TU Clausthal

Clausthal University of Technology, Germany
Faculty of Science
Institute for Informatics

Cumulative Habilitation Thesis
for Granting of the *Venia Legendi* in the Field of Informatics

Multimodal Human Machine Interaction Technologies at Controller Working Positions

DOI: 10.21268/20260420-0

from: Oliver Ohneiser
ORCID: 0000-0002-5411-691X

Submission of Habilitation Thesis: 19th December 2024
Scientific Presentation and Colloquium: 15th April 2026
Public Trial Lecture: 29th April 2026

Internal and External Reviewers of the Habilitation Thesis:

- Prof. Dr. Sven Hartmann (Big Data and Technical Information Systems, Institute for Informatics, Clausthal University of Technology, Germany)
- Prof. Dr. Christian Siemers (Automation Technology, Institute for Electrical Information Technology, Clausthal University of Technology, Germany)
- Prof. Dr. Jonas Lundberg (Human Centered Design, Media and Information Technology, Department of Science and Technology, Linköping University, Sweden)
- Prof. Dr. Max Mulder (Aerospace Human-Machine Systems, Control and Simulation Section, Delft University of Technology, The Netherlands)

Habilitation Commission:

- Prof. Dr. Jörg P. Müller (Mobility and Enterprise Computing, Institute for Informatics, Clausthal University of Technology, Germany)
- Prof. Dr. Sven Hartmann (Big Data and Technical Information Systems, Institute for Informatics, Clausthal University of Technology, Germany)
- Prof. Dr. Rüdiger Ehlers (Automating CPS Design, Institute for Software and Systems Engineering, Clausthal University of Technology, Germany)
- Prof. Dr. Benjamin Säfken (Applied Statistics, Institute of Mathematics, Clausthal University of Technology, Germany)
- Prof. Dr. Gunther Brenner (Fluid Mechanics, Institute of Applied Mechanics, Clausthal University of Technology, Germany)
- Prof. Dr. Thomas Niemand (Business Administration and Management of Digital Transformation, Institute of Management, Economics, and Law, Clausthal University of Technology, Germany)
- Dean at Time of Thesis Submission at Faculty of Mathematics/Computer Science and Mechanical Engineering: Prof. Dr. Armin Lohrengel (Machine Elements and Engineering Design, Institute of Mechanical Engineering, Clausthal University of Technology, Germany)
- Dean at Time of Oral Examinations at Faculty of Science: Prof. Dr. René Wilhelm (Institute of Organic Chemistry, Clausthal University of Technology, Germany)

Abstract

Multimodal Human Machine Interaction Technologies at Controller Working Positions

Human machine interaction technologies such as automatic speech recognition and understanding (ASRU), text-to-speech, multitouch, and eye tracking have found their way into daily life. However, they can hardly be found in safety-critical environments such as controller working positions for air traffic management (ATM) tasks. There is a lack of investigation how such interaction technologies and their combination could support human aviation operators. Therefore, a row of human machine interaction concepts has been prototypically implemented in software and hardware to gather subjective and objective data from human study subjects using those prototypes. More than 200 international subjects participated as controllers and pilots in human-in-the-loop simulation trials of 16 validation campaigns to validate the prototypes on multimodal interaction technologies at aeronautical working positions. The various prototypic solutions with advanced interaction methods have shown potential benefits over the baseline conditions with interaction methods close to today's operations such as (i) an improvement of input speed by up to 15% as well as a reduction of errors by a factor of two in entering digitized ATM data, (ii) a statistically significant reduction of human aviation operators' mental workload, and (iii) an increase in their situation awareness with succeeding effects like (iv) improved safety and efficiency of ATM. The high performance especially of the ASRU prototypes and the human-centered design of all interaction technologies led to high acceptance and usability ratings as well. The individual interaction technologies and also their multimodal combination have shown feasibility in high-fidelity laboratory environments and were even already partly evaluated on operational data. Thus, they are ready to be put into operational practice.

Keywords
Human Machine Interaction
Air Traffic Management
Air Traffic Control
Controller Working Position
Multimodal Interaction
Automatic Speech Recognition and Understanding
Text-To-Speech
Gesture Recognition
Tactile Cue
Eye Tracking
Visual Cue
Attention Guidance

Kurzfassung

Multimodale Mensch-Maschine-Interaktionstechnologien an Lotsenarbeitsplätzen

Mensch-Maschine-Interaktionstechnologien wie automatische Spracherkennung und Sprachverstehen (ASRU), Sprachsynthese, Gestenerkennung (Multi-Touch) und Blickerkennung haben ihren Weg in den Alltag gefunden. In sicherheitskritischen Umgebungen wie an Lotsenarbeitsplätzen für Aufgaben des Luftverkehrsmanagements (ATM) sind sie jedoch kaum zu finden. Es mangelt an Untersuchungen, wie solche Interaktionstechnologien und ihre Kombinationen menschliche Operateure in der Luftfahrt unterstützen könnten. Daher wurde eine Reihe von Mensch-Maschine-Interaktionskonzepten prototypisch in Software und Hardware implementiert, um subjektive und objektive Daten von menschlichen Versuchspersonen zu sammeln, die diese Prototypen verwenden. Mehr als 200 internationale Probanden nahmen als Lotsen oder Piloten an Human-in-the-Loop-Simulationsversuchen von 16 Validierungskampagnen teil, um die Prototypen mit multimodalen Interaktionstechnologien an Luftfahrt-Arbeitspositionen zu validieren. Die verschiedenen prototypischen Lösungen mit fortschrittlichen Interaktionsmethoden haben potentielle Vorteile gegenüber den Vergleichsbedingungen mit heutigen operationellen Interaktionsmethoden aufgezeigt, wie z.B. (i) eine Erhöhung der Eingabegeschwindigkeit um bis zu 15% sowie eine Halbierung der Fehlerquote bei der Eingabe digitalisierter ATM-Daten, (ii) eine statistisch signifikante Verringerung der mentalen Arbeitsbelastung von Luftfahrt-Operateuren sowie (iii) eine Steigerung ihres Situationsbewusstseins mit Folgeeffekten wie (iv) verbesserte Sicherheit und Effizienz des Luftverkehrsmanagements. Die hohe Performanz insbesondere der ASRU-Prototypen und die mensch-zentrierte Gestaltung aller Interaktionstechnologien führten ebenso zu hohen Akzeptanz- und Nutzbarkeitsbewertungen. Die einzelnen Interaktionstechnologien sowie deren multimodale Kombinationen haben sich in realistischen Laborumgebungen als praktikabel erwiesen und wurden teilweise sogar bereits mit operationellen Daten evaluiert. Daher sind sie bereit, in die operationelle Praxis überführt zu werden.

Schlagwörter

Mensch-Maschine-Interaktion
Luftverkehrsmanagement
Flugsicherung
Lotsenarbeitsplatz
Multimodale Interaktion
Automatische Spracherkennung und Sprachverstehen
Sprachsynthese
Gestenerkennung
Taktile Hinweise
Blickerkennung
Visuelle Hinweise
Aufmerksamkeitslenkung

Acknowledgments

Many thanks to:

- Prof. Dr. Sven Hartmann, Chair of Big Data and Technical Information Systems in Department of Informatics at Clausthal University of Technology, for welcoming me with open arms and offering great self-determined opportunities,
- Apl. Prof. Dr.-Ing. Umut Durak, for initializing the topic of Aeronautical Informatics at a Northern German university and for connecting me with the Department of Informatics at Clausthal University of Technology,
- Hon. Prof. Dr.-Ing. Hartmut Helmke, the unofficial leader of the Automatic Speech Recognition and Understanding (ASRU) group within DLR's Institute of Flight Guidance and deputy head of the Controller Assistance department for your enthusiasm, hard work, and meticulousness from which I learned and benefitted a lot during the last decade,
- Dipl.-Ing. Jürgen Rataj, the former head of DLR's department Controller Assistance at the Institute of Flight Guidance, as each talk with you was like a professional coaching,
- The complete ASRU group with amongst others M.Sc. Matthias Kleinert, Heiko Ehr, and M.Sc. Shruthi Shetty for the great mutual support in all the past ASRU projects,
- The Controller Interaction group with amongst others M.Sc. Malte-Levin Jauer, M.Sc. Hejar Gürlük, and Dipl.-Inf. Maria Uebbing-Rumke for great support regarding TriControl and Attention Guidance, as well as Dr. Marco-Michael Temme and Ph.D. Ulf Ahlstrom (FAA) for the weather situation awareness research,
- The colleagues of DLR's Institute of Flight Guidance such as the current head of DLR's department Controller Assistance Dr.-Ing. Sebastian Schier-Morgenthal, e.g., for supporting our fascinating projects and human-in-the-loop simulation studies with controllers,
- Many national and international DLR-external project partners for very good common achievements and co-authoring of scientific papers,
- A lot of further colleagues and students who I came across with during the last years, who worked with me and supported some of the relevant aspects of this cumulative habilitation thesis,
- Prof. Dr. Maarten Boersma, Vice Director of the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, for being a great mentor starting in the Helmholtz Mentoring programme "Helmholtz Advance",
- My "multiple"-great uncle Attorney at Law Bob Ohneiser from Virginia for performing an English native speaker review of this cumulative habilitation thesis,
- My parents for shaping me to always critically, but positively reflect progress in life and my brother Dr. Kevin Ohneiser also for pushing scientific achievements,
- My wife for giving me the space to pursue my habilitation and my two little kids for motivation when I see how you explore the world.

Table of Contents

List of Figures

List of Abbreviations

A-SMGCS .. **A**dvanced - **S**urface **M**ovement **G**uidance and **C**ontrol **S**ystem

ABSR ..... **A**ssistant **B**ased **S**peech **R**ecognition

AcListant ... **Ac**tive **List**ening **A**ssista**nt**

AG ..... **A**ttention **G**uidance

ANSP ..... **A**ir **N**avigation **S**ervice **P**rovider

ASR ..... **A**utomatic **S**peech **R**ecognition

ASRU ..... **A**utomatic **S**peech **R**ecognition and **U**nderstanding

ATC ..... **A**ir **T**raffic **C**ontrol

ATCo ..... **A**ir **T**raffic **Co**ntroller

ATM ..... **A**ir **T**raffic **M**anagement

ATS ..... **A**ir **T**raffic **S**ervice

CARS ..... **C**ontroller **A**cceptance **R**ating **S**cale

CPDLC .... **C**ontroller-**P**ilot **D**ata **L**ink **C**ommunications

CWP ..... **C**ontroller **W**orking **P**osition

DLR ..... **D**eutsches Zentrum für **L**uft- und **R**aumfahrt e.V.; German Aerospace Center

DTT ..... **D**igital **T**ower **T**echnologies

EEG ..... **E**lectro**e**ncephalo**g**raphy

EU ..... **E**uropean **U**nion

EUROCAE . **Eur**opean **O**rganization for **C**ivil **A**viation **E**quipment

FAA ..... **F**ederal **A**viation **A**dministration

FL ..... **F**light **L**evel

FN ..... **F**alse **N**egative

FP ..... **F**alse **P**ositive

ft ..... **f**ee**t**

HAAWAII .. **H**ighly **A**utomated **A**ir Traffic Controller **W**orkstations with **A**rtificial **I**ntelligence **I**ntegration

HMI ..... **H**uman **M**achine **I**nterface

ICAO ..... **I**nternational **C**ivil **A**viation **O**rganization

ISA ..... **I**nstantaneous **S**elf-**A**ssessment (of Workload)

MALORCA **Ma**chine **L**earning **o**f Speech **R**ecognition Models for **C**ontroller **A**ssistance

MET4ATM . **Met**eorology **for A**ir **T**raffic **M**anagement

MINIMA ... **Mi**tigating **N**egative **I**mpacts of **M**onitoring high levels of **A**utomation

NASA ..... **N**ational **A**eronautics and **S**pace **A**dministration

NASA-TLX NASA-**T**ask **L**oad Inde**x**

PROSA .... Controller Tools and Team Organisation for the **Pro**vision of **S**eparation in **A**ir Traffic Management

PTT ..... **P**ush-**T**o-**T**alk

SAGAT ..... **S**ituation **A**wareness **G**lobal **A**ssessment **T**echnique

SASHA ..... **S**ituation **A**wareness for **S**olutions for **H**uman **A**utomation Partnerships in European ATM

SATI ..... **S**hape **A**utomation **T**rust **I**ndex

SESAR ..... **S**ingle **E**uropean **S**ky **A**TM **R**esearch Programme

SINOPTICA **S**atellite-borne and **In**-situ **O**bservations to **P**redict **T**he **I**nitiation of **C**onvection for **A**TM

SPAM ..... **S**ituation **P**resence **A**ssessment **M**ethod

STARFiSH . **S**afe**t**y and **A**r**t**i**f**icial **I**ntelligence **S**peec**h** Recognition

SUS ..... **S**ystem **U**sability **S**cale

TA ..... **T**ot**a**l Number

TLX ..... **T**ask **L**oad Inde**x**

TMA ..... **T**erminal **M**anoeuvring **A**rea

TN ..... **T**rue **N**egative

TP ..... **T**rue **P**ositive

TRL ..... **T**echnology **R**eadiness **L**evel

TTS ..... **T**eext-**T**o-**S**peech

USA ..... **U**nited **S**tates of **A**merica

VAD ..... **V**oice **A**ctivity **D**etection

VHF ..... **V**ery **H**igh **F**requency

WER ..... **W**ord **E**rror **R**ate

1 Introduction of Human Machine Interaction in Aviation

Human operators as supervisors of automated systems are using their auditory, tactile, and visual sensing modality to different extents during their work. In the aviation domain, the visual and auditory modality are the main means for interaction of air traffic controllers (ATCos) and cockpit crews (pilots). While the visual modality is mostly used by human operators to acquire information from situation data displays for air traffic services (ATS) or flight guidance, the auditory modality is predominantly utilized for verbal communication, e.g., via radio telephony.

The main task of ATCos is to ensure a "safe, orderly and expeditious flow of air traffic" [ICAO, 2016]. Therefore, ATCos at airport towers and control centers coordinate and communicate intensely with cockpit crews and other controllers. Even though, digital controller-pilot data link communications (CPDLC) are in use especially for non-time-critical air traffic control (ATC) instructions, verbal communication remains critical in ATC.

Mental workload and situation awareness are known as some of the most critical human factors for human operators in aviation in order to enable safe and efficient flights. However, during peak hours at hub airports and within central airspace sectors it can be hard to always be mentally "ahead of traffic" and to avoid overload. These aspects are even more important facing increasing additional tasks such as continuous descent operations or abatement of noise and contrails.

Further challenges come with regulatory requirements and the way of human usage of software and hardware. If, for example, regulations require an ATCo to manually input all given ATC instructions into the electronic ATC system, which does not exist to this extent today, the ATCo workload will increase. The number and size of displays as well as unnatural or unintuitive human machine interaction can also complicate the work of human operators in aviation.

A better support of ATCos and pilots with advanced human machine interaction technologies can help to positively affect a row of human factors such as mental workload, situation awareness, trust, acceptance, usability and at the same time increase efficiency and safety in the air traffic management (ATM) domain – also facing the current shortage of ATCos [Wallace, 2024], [Eccles, 2023] and pilots [Placek, 2024].

Controllers should have proper awareness regarding air traffic situation, weather data, and upcoming air traffic control events in order to issue timely, mostly verbal instructions to pilots and enter the instruction content manually into digital air traffic management systems. All those tasks are supported with the described multimodal interaction technology prototypes at controller working positions (CWP) in this thesis – as visualized with the images in Figure 1.1. The research question is formulated as:

"How can human machine interaction technologies and their multimodal combination support the work of human aviation operators?"

This thesis presents concepts, implementations, and validation results for software prototypes integrated into realistic CWPs covering the auditory, tactile, and visual modality as well as combinations of those.

Figure 1.1: Futuristic sketches of multimodal interaction technologies at controller working positions created by artificial intelligence.

Chapter 2 outlines related work on three modalities for human machine interaction, i.e., (1) auditory modality on speech recognition and understanding as well as text-to-speech, (2) tactile modality comprising gesture recognition and tactile cues, (3) visual modality with eye tracking, display aids, and even attention guidance mechanisms for human aviation operators as well as (4) combinations of the aforementioned modalities with a focus on the integration into applications. This chapter also describes validation measures for research prototypes regarding human and system performance as well as the technology readiness level metric.

Chapter 3 explains a bundle of concepts and prototypic implementations during the habilitation phase considering and integrating human machine interaction technologies regarding the three modalities and their multimodal combination. The prototypes cover the active and passive side of each of the three interaction modalities – speaking and listening, touching and feeling, looking and highlighting for auditory, tactile, and visual interaction. The developed prototypes for working positions in aviation are validated through human-in-the-loop simulation studies as described with study setups, experiment conduction, results, and their discussion mainly considering human performance of air traffic control operators – as well outlined in Chapter 3.

Chapter 4 summarizes, concludes, and gives an outlook on future work. The annex contains a list of references and provides copies of eleven research publications with more details on the aforementioned concepts, prototypic implementations, and validation campaigns as well as explains the contribution of the habilitand to them in the course of this cumulative habilitation thesis.

2 Related Work on Human Machine Interaction Technologies and Applications

This chapter outlines related work[1] on interaction between human and machine/automation [Sheridan and Parasuraman, 2005] regarding three interaction modalities and their integration. Relevant literature will be presented[2] about (1) the acoustic modality on speech recognition and understanding as well as text-to-speech, (2) the tactile modality mainly comprising gesture recognition, (3) the visual modality with eye tracking, visual cues, and attention guidance mechanisms for human operators, as well as (4) combinations of the aforementioned modalities.

2.1 Human Control of Automated Systems in Aviation

There are two central human operators in aviation: controllers to guide air traffic and pilots to steer aircraft. The majority of ATCos works in control centers. They deal with en-route, approach, and departure traffic. Tower ATCos handle the traffic close to airports and on the airport runways. Apron controllers are responsible for aircraft movements on an airport's ramp area. Hereinafter, the term *controller* includes air traffic controllers from the en-route, approach, and tower environment as well as apron controllers.

The CWPs differ due to the different controller tasks and due to the ATM system suppliers that equip the national air navigation service providers (ANSP). The central CWP elements are voice communication systems for radio telephony communication with aircraft pilots and situation data displays, i.e., mostly radar displays which are especially relevant for center ATCos. Nowadays, the ATC systems are mainly digitized. Relevant information about flights under responsibility of a controller such as callsigns of the flights as well as current altitudes and speeds can be acquired from electronic aircraft radar labels or electronic flight strips. The controllers manage the relevant air traffic mainly through communication and monitoring to assure separation minima between aircraft [Brooker, 2011a]. Communication includes instructing verbal commands to pilots in English language for flights in controlled airspace above a certain flight level following a phraseology as defined by the International Civil Aviation Organization (ICAO) [ICAO, 2016] and checking the pilots' readback – the hearback. Roughly every hundredth readback contains an unintended error quite independent of the ATC environment such as en-route [Cardosi, 1993], approach [Cardosi et al., 1996], or tower [Cardosi, 1994]. Even if this readback error evaluation based on data from three decades ago, the basic mechanisms of ATC radio telephony did not change so that the frequency of readback errors is expected to roughly remain the same until today. The verbal utterances can also contain requests or pilot reportings next to commands and their readbacks.

[1] None of the literature references in this chapter was authored or co-authored by the habilitand.

[2] The copies of the eleven published research articles in the annex contain 38, 41, 47, 58, 40, 37, 20, 102, 31, 35, and 110 references with some degree of overlap, respectively. In order to not completely repeat the literature research of those papers within this chapter, only the most relevant paper citations will be re-used. Furthermore, some additional and more recent papers are added.

Especially in non-time-critical environments such as en-route, data-link messages with ATC instructions can be sent by the controllers to an aircraft instead of verbal communication. After readback and hearback, controllers continuously monitor the potentially changed aircraft motion regarding the prior issued instructions in order to enable a safe and swift air traffic. The described ATCo tasks can be supported by electronic assistance systems such as decision support systems or conflict probing tools at some CWPs [Brooker, 2011b].

Primarily flights in controlled airspace or at a controlled airport, independent of aircraft size or flight type, have the obligation to be in contact with the controllers on the ground. Thus, pilots communicate with controllers and actively steer their aircraft. Hereinafter, the term *pilot* includes flying and monitoring pilots of flights under both instrument and visual flight rules. Pilots of commercial flights usually have electronic cockpit systems such as a primary flight display, a flight management system, and an electronic flight bag.

In case of analogue radio telephony communication, controllers and pilots are required to manually enter the relevant communication content into their electronic systems. The interaction between human aviation operators and automated systems often focuses on the support of the electronic system. However, there are various possibilities to support the human operators with multimodal interaction technologies as well.

2.2 Auditory Modality: Automatic Speech Recognition and Understanding, Text-To-Speech

Given the aviation radio communication as sketched in the previous section, there are three aspects of interest: (1) automatic speech recognition (ASR), i.e., speech-to-text (transcription) of ATCo and pilot utterances, (2) automatic speech understanding, i.e., text-to-concepts (annotation) as relevant for ATC, and (3) automatic text-to-speech, i.e., verbal machine readings that sound like human utterances.

The history of ASR in ATC can be traced back to the 1970's [Connolly, 1977]. Initially, ASR was used to support ATC training and to reduce the number of required simulation-pilots [Schäfer, 2000], [Ciupka, 2012]. Furthermore, the workload of human aviation operators was assessed using the transcriptions of their utterances [Cordero et al., 2013], [Cordero et al., 2012]. A row of ATC speech corpora have been recorded [Zuluaga-Gómez et al., 2023a], [Smídl et al., 2019], [Hofbauer et al., 2008] and frameworks [Lin et al., 2021b], [Povey et al., 2011] have been proposed in order to provide a data base for research and development [Pellegrini et al., 2019].

The content of aviation radio communication mainly comprises of three elements. First, an aircraft callsign to indicate who or whose pilot should adhere to the following commands and should react on potential instructions. Second, a command type to indicate what aspect of the current flight requires an action, e.g., changing the altitude/speed/heading/frequency/etc. or reporting something. Third, a command value to indicate what exact change of the current flight is required, e.g., a flight level or a waypoint. If there are multiple commands in an utterance, there will usually also be multiple command type-value pairs. The three basic radio communication elements already cover many ATC instruction elements. Further elements and their interrelation will be analyzed in Section 3.1.2.

The ASR performance on ATC speech data is usually measured with the word error rate (WER). The WER bases on the Levenshtein distance [Levenshtein, 1966], i.e., the sum of substitutions, deletions, and insertions, which is divided by the number of words in the reference sentence.

Reference (e.g., gold transcription): air france one four eight zero you are cleared to luton perit five delta departure initially climb five thousand feet squawk three one zero three

Hypothesis (e.g., speech-to-text output): airfrans one air four eight zero you are cleared to ten united five delta departure initially climb five thousand feet squawk three one zero three

Substitutions: 3, Deletions: 1, Insertions: 1

Figure 2.1: Calculation of word error rate for a given reference and hypothesis sentence.

In the example presented in Figure 2.1, three substitutions, one deletion, and one insertion result in a Levenshtein distance of five that needs to be divided by the number of 25 reference words, i.e., the WER is 20%.

With evolving technology such as deep neural networks and end-to-end speech recognition, the ASR performance improved compared to earlier approaches [Zuluaga-Gómez et al., 2020]. Challenges for speech recognition in ATC are voice activity detection and speaker role detection [Khalil et al., 2023]. With more advanced applications, the complexity of not only recognizing, but also understanding the semantics of the recognized word sequences from spoken instructions [Lin, 2021], [Pardo et al., 2011] increases.

Hence, it is not sufficient to only spot keywords such as “reduce”/“descend” and simple values. ATC concept extraction from recognized utterances often focuses on the important ATC concepts only such as callsigns [Kasttet et al., 2024], [García et al., 2023] or runway information [Badrinath and Balakrishnan, 2022]. Variations in the used phraseology and deviations from the ICAO phraseology require deeper analysis and use of background information.

Additional data to improve ASR can be used through lip reading in audio-visual speech recognition [Rudregowda et al., 2024]. But also knowledge about the domain and the current situation help to improve speech recognition accuracy. Contextual air traffic situation knowledge for controllers was considered to improve the accuracy of speech recognition [Guo et al., 2021], [Nguyen and Holone, 2016], [Shore et al., 2012] and speech understanding [Kleinert et al., 2018], e.g., to improve the quality of electronic decision support systems [Helmke et al., 2013].

The speech recognition and understanding technology was enrolled for various ATC research purposes such as runway incursion detection [Chen et al., 2015], readback error detection [Rajendram Bashyam et al., 2023], [Chen et al., 2017], CPDLC support at ATC workstations [Lechner et al., 2002], digitizing ground taxi instructions [Steinmetz et al., 2024], or pilot report analysis [Chen et al., 2022].The work in this thesis makes use of the relevant state-of-the-art techniques and audio corpora in speech recognition, but goes far beyond existing speech understanding methodologies and results in the ATM context. First ASRU systems at least for data recording have recently been deployed in ATC centers [Lin et al., 2023].

Initial experiments with text-to-speech (TTS) technology in aviation found that intelligibility [Manaker, 1982] and response times to synthetic speech [Simpson and Williams, 1980] are at least comparable to human voice for cockpit audio. To ease understanding and reduce miscommunication in ATC communication especially based on accents, the use of TTS has been suggested [Dhavala, 2014].

Commercial products utilize TTS for automatic terminal information service in tower and en-route environment, for ATC modules in flight simulation, or for cockpit voice warning systems [Thorburn, 1971]. When combining speech-to-text, text-to-concepts, and TTS, this also enables virtual simulation pilots [Zuluaga-Gómez et al., 2023b]. Pilot repetitions to recognized and interpreted ATCo instructions can be automatically generated as text and synthesized to speech [Zhang et al., 2022]. A prototypic ATC communication training environment lacked realistic artificial voices [Auinger, 2019]. A more professional TTS system has been used by the German ANSP to train multiple hundred ATCos [Slotty and Rühl, 2012]. The lack of TTS models fine-tuned with audio data from ATC environments is often a limitation of prototypes [Lin et al., 2021a] that can lead to reduced realism.

A soundscape for air traffic control data to be used through the auditory channel of a human operator, can be a good complement to visual cues as shown with a combination of visual attention guidance cues and sonified aircraft CPDLC login at a CWP [Hunger et al., 2024]. The sonification of remote tower control data, especially about aircraft sector entries with information on the aircraft's cardinal direction was as well appreciated by ATCos [Elmquist et al., 2023]. The above-reported works indicate the potential of utilizing different aspects of the auditory channel.

2.3 Tactile Modality: Two- and Three-Dimensional Gesture Recognition, Tactile Cues

Recognizing two- or three-dimensional hand gestures [Mitra and Acharya, 2007] is an important aspect for tactile interaction of human aviation operators. Different aspects of touch screens were investigated for aircraft flight decks [Rouwhorst et al., 2017], [Dodd et al., 2014], e.g., manipulating radio frequencies [Avsar et al., 2016]. Visual gesture recognition has been implemented to detect signs for aircraft from marshallers at an airport's apron [Prakash et al., 2016], [Singh et al., 2005].

A design standard exists for ATC touch interfaces [Dodd et al., 2022]. The DigiStrips experiment is an early examination of touch screens for ATC CWPs [Mertz et al., 2000]. A multi-user tabletop surface has been used to foster collaboration of ATCos [Conversy et al., 2011]. A touch-based prototype [Wald, 2011] and indirect touch interaction [Causse et al., 2014] have been explored in the ATC context. A study on multitouch gestures for a stripless ATC system confirmed the usability of most touch gestures [Hagemann and Udovic, 2019]. Another initial multitouch display mock-up for a terminal manoeuvring area (TMA) CWP was evaluated with 14 controllers showing better usability and less workload than with a mouse mock-up display [Uebbing-Rumke et al., 2014]. In this thesis multitouch gestures will be used together with other interaction means based on the gained usability experience described above.

Vibrotactile and especially audiotactile cues were found to potentially improve the detection time of visual events in ATC displays [Ngo et al., 2012]. In a simulator with an automated cockpit, tactile cues led to faster response times and improved event detection rates compared to visual cues [Sklar and Sarter, 1999]. Haptic feedback reduced the workload of tele-operators for uncrewed aerial vehicles [Smisek et al., 2017]. If haptic feedback, e.g., on a side-stick in a cockpit, is enhanced with visual indications on a primary flight display, this at least improved the subjective user interface assessment of study subjects [de Rooij et al., 2023]. These findings suggest to examine and combine cues of different interaction modes if the guidance of attention is desired.

2.4 Visual Modality: Eye Tracking Technology, Visual Cues, and Attention Guidance

Eye tracking technology can be used for various tasks of human operators in aviation as the complementary literature reviews for cockpit use cases reveal [Peißl et al., 2018], [Ziv, 2016]. Eye tracking data – including fixations and saccades – can serve as a proxy for human operator attention [Wang et al., 2021]. With this, it is used to study monitoring tasks of ATCos [Hasse et al., 2012] and pilots [Peysakhovich et al., 2018], [Cox et al., 2005] as well as situation awareness [Dehais et al., 2017] and workload of pilots [Faulhaber and Friedrich, 2019]. Furthermore, training of ATCos' monitoring tasks can be supported with the use of eye tracking [Barzantny, 2018], [Kang et al., 2017].

Visual scan patterns have been investigated in different ATC environments such as radar control [Wee et al., 2017], tower control [Westin et al., 2019], [Manske and Schier, 2015], and multiple remote tower control [Li et al., 2018]. Eye tracking data has also been used to study the issue of forgetting in ATC [Jin et al., 2023] as well as fatigue [Nealley and Gawron, 2015], the latter human factor even with a combination of gaze and speech data [Xu et al., 2024]. Apart from pure analysis of eye tracking data, it has also been explored for interactive ATC systems [Traoré and Hurter, 2016], [Alonso et al., 2013].

Visual attention guidance to relevant spots or ATC events has been investigated for tower controller working positions with augmented reality [Teutsch et al., 2022] and en-route controller working positions for human-machine teaming [Jameel et al., 2023]. Visual cues for attention guidance usually contain text information such as advisories for ATC commands, e.g., to fly around severe weather at an en-route controller working position [Ahlstrom, 2015]. A basic prototype utilizing soft visual cues to smoothly guide the ATCo's gaze has been evaluated showing a preference for changes of symbol properties compared to color changes [Nylin et al., 2020]. In another study, the display contrast has been found to be an important factor to gain visual attention [Palmer et al., 2008]. Acoustic attention guidance usually focuses on very important information such as altitude during landing or warnings and alerts for potential separation infringements. Within this thesis, eye tracking is used as a mean to select displayed items and to guide attention in an ATC context. Furthermore, advanced visual cues will be developed and examined based on the knowledge of existing prototypes.

2.5 Multimodality: Prototypes with more than One Interaction Technology

Multimodal interaction can be understood as combining "natural input modes such as speech, pen, touch, hand gestures, eye gaze, and head and body movements" [Oviatt, 1999, p. 74]. Multimodal interaction in general has already been analyzed at the beginning of this century [Dumas et al., 2009], [Oviatt, 2002]. In addition, mental load [Oviatt et al., 2004] and evaluation of multimodal systems [Wechsung, 2014] have been explained. Furthermore, combining gaze based control and ASR for pilot interaction in the cockpit [Merchant and Schnell, 2000], [Gauci et al., 2018] and to support adaptivity of multimodal human machine interfaces in aviation [Lim et al., 2018] were analyzed. Another multimodal prototype offers pilots to interact with an autopilot through voice commands or via multitouch [Gauci et al., 2017]. Combining the analysis of three interaction modes, i.e., gestures, speech, and gaze has already been investigated for general discourse [Quek et al., 2000] and in the automotive domain [Neßelrath et al., 2016]. The combination of ASR and multitouch inputs to enter ATC commands [Jauer, 2014] as well as a combination of eye tracking and multitouch [Seelmann, 2015] have been prototypically developed and evaluated on a low technological maturity level.

The multimodal combination including multitouch on an ATC flight strip board has been investigated with ATCos [Savery et al., 2013]. However, there was no prototype and no systematic evaluation yet to combine speech, gestures, and gazes for digitized ATC system input as described in this thesis.

2.6 Validating System and Human Performance in Human-in-the-Loop-Simulations

Mental workload, situation awareness, and trust in automation [Parasuraman et al., 2008] are three cornerstones of successfully usable prototypes for human aviation operators [Parasuraman and Riley, 1997]. Those human factors [Wickens et al., 1997] and their measurement will be outlined in the following. Afterwards, some measures for system performance and a taxonomy for technological maturity will be presented.

2.6.1 Workload Measurement

Mental workload is an important human factor influencing the performance of human operators. The manual workload is usually of much less importance for aviation operators. There are various subjective and objective assessment methods for mental workload. The *Instantaneous Self-Assessment of Workload (ISA)* [Jordan and Brennen, 1992] is a retrospective self-rating of human operators. They need to rate their mental workload of the last five minutes during a simulation on a five-point Likert scale [Likert, 1932]. Such Likert scales are frequently used in various domains [Joshi et al., 2015].

The *Bedford workload scale* [Roscoe, 1984] – originally designed for pilots – requires a retrospective rating on a ten-point Likert scale. One of the most prominent subjective methods is the *National Aeronautics and Space Administration – Task Load Index (NASA-TLX)* [Hart, 2006]. The human operators need to retrospectively rate their experienced workload in terms of mental, physical, and temporal demand as well as performance, effort, and frustration. These six factors can also be weighted regarding their importance. As the above-described ratings are done retrospectively, this can lead to a bias towards the more recent or more intense experiences. Furthermore, physiological measures such as via electroencephalography (EEG) can be used to derive mental workload in ATC [Bernhardt et al., 2019] and to even trigger adaptive automation functionalities [Aricò et al., 2016]. Combining EEG and eye-tracking for workload measurement can also help to recognize situation awareness of ATCos [Li et al., 2023].

Another more objective measure of mental workload and any assumed free mental capacity can be obtained with a secondary task [Kaber and Riley, 1999] that human study subjects need to fulfill. The primary task, e.g., controlling air traffic, remains most important. However, if the human study subjects have time, they need to solve short repetitive tasks like math exercises or color-word assignment tasks such as the Stroop test [Stroop, 1935]. All of these workload measurement techniques have been used for the work in this thesis.

2.6.2 Determining Situation Awareness

Situation awareness of human operators is defined by Endsley as “the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future” [Endsley, 1988, p. 792]. It is essential for human aviation operators in order to safely and efficiently perform their tasks [Landry, 2011].

Several techniques were proposed and widely used to assess aviation operators' situation awareness [Kaber et al., 2006]. The *situation awareness global assessment technique (SAGAT)* stops and blacks the displays of the running simulation trial and asks questions about the situation directly before the stop, e.g., about aircraft characteristics [Endsley, 1988]. The percentage of correct answers serves as an indicator for the situation awareness. The *situation presence assessment method (SPAM)* method [Endsley, 2021] takes the timely delay of the answer without blacking the simulation trial displays as an indicator for situation awareness.

The *SASHA* questionnaire *Situation Awareness for Solutions for Human Automation Partnerships in European ATM (SHAPE)* [Dehn, 2008] needs to be filled by the study subjects after the simulation run, i.e., they rate situation awareness related statements on a Likert scale. The different techniques have different intrusiveness levels, different points in time when they are executed, require different amount of effort and comprise different queries to answer or statements to rate, respectively.

Thus, it needs to be decided individually for each study, which technique or combinations of them fits best to the given purpose and research questions. *SASHA* and a sophisticated version of *SPAM* have been used for the work in this thesis.

2.6.3 Usability, Acceptance, and Trust Questionnaires

A row of further factors next to workload and situation awareness influences the human performance when working with automated systems. System usability is often assessed with the *System Usability Scale (SUS)* [Brooke, 1996]. The underlying questionnaire requires ratings about ten usability statements on a five-point Likert scale. The acceptance of automated systems can be assessed with the *Controller Acceptance Rating Scale (CARS)* [Lee et al., 2001].

The trust of human aviation operators in a system may be measured with the questionnaire *SHAPE Automation Trust Index (SATI)* [Dehn, 2008] that needs to be filled by the study participants. All those questionnaire items need to be rated after finishing a simulation run, which can lead to results that focus more on the last part of the simulation or any intense phases instead of an average. *SUS*, *CARS*, and *SATI* have been used for the work in this thesis.

2.6.4 Confusion Matrix, Precision, Recall, and Accuracy Metric

The system performance can usually be measured more objectively than human performance based on data recordings during simulation runs. The portion of correctness and the number of errors regarding a prototype's functionality often play a central role. A confusion matrix divides true positives (correct classification as target), true negatives (correct classification as non-target), false positives (incorrect classification as target), and false negatives (incorrect classification as non-target).

The precision is then defined as the number of true positives divided by the sum of true and false positives [Powers, 2011]. The recall is defined as the number of true positives divided by the sum of true positives and false negatives. The accuracy can be calculated as the number of correct classifications (true positives plus true negatives) divided by the number of all classifications. These classifications are relevant for sophisticated metrics in this thesis.

2.6.5 Technology Readiness Levels

NASA defined nine different technology readiness levels (TRL) ranging from basic principle reports in TRL1 to successful operations in TRL9 [Mankins, 1995]. The prototypes in this thesis range from technology concepts at TRL2 to prototype demonstrations in a relevant environment at TRL6. The process of rapid prototyping for operational concept development has been explained for the ATM domain [Edinger and Schmitt, 2012].

On lower TRLs, the prototypes rather undergo "proof-of-concept" trials and evaluations of the technical system while more complex validation trials on higher TRLs answer research hypotheses and assess human performance as well as system performance based on objective and subjective data. For simplicity, the term "validation trials" will be used hereinafter if human aviation operators participated in a real-time human-in-the-loop simulation study with research prototypes independent of the TRL.

In view of the literature research above, it can be stated that the technological development of individual interaction technologies in general and in non-aviation domains already reached industrial TRLs. In the aviation domain there are some research prototypes on low TRLs using the individual interaction technologies. However, it is rare to find multimodal combinations for human machine interaction integrating speech recognition and understanding, gesture recognition, eye tracking, or acoustic, tactile, and visual cues, and even more rare to find any conclusive research focused on supporting human aviation operators at controller working positions. Furthermore, a systematic quantification of potential benefits and drawbacks validated with human operators in realistic simulation trials is missing in the ATC domain. This thesis closes the described gap and is supported by 52 publications on different aspects of human machine interaction in the aviation domain. The relevant concepts, prototypic implementations, and validation results will be outlined in Chapter 3.

3 Concepts, Implementations, and Validation Campaigns for Interaction Prototypes

This chapter explains the cumulative habilitation thesis relevant concepts and prototypic implementations for multimodal interaction technologies in ATM. This encompasses outlining technologies on the auditory, tactile, and visual modality as well as a combination of these three modalities[1].

Usually, the concept development and implementation activities use a human-centered, iterative approach. This includes an early involvement of system matter experts for conceptualization and a row of pre-trials with potential end users to gather continuous qualitative and quantitative feedback for adjusting the further rapid prototyping development. Finally, human-in-the-loop validation trials on different TRLs were conducted in order to validate the research hypotheses or to answer research questions. The conduction and results of validation trials on each prototype are presented directly after each prototype description.

3.1 Auditory Modality: Automatic Speech Recognition and Understanding as well as Text-To-Speech for Air Traffic Management Applications

The auditory modality mainly focuses on the support for automatic understanding of aviation radio communication in this thesis. There are three basic steps to make use of aviation voice communication. First, the process of ASR also known as speech-to-text. Second, the natural language processing step, i.e., automatic speech understanding to convert text to relevant semantic concepts focusing on ATC as the core part of ATS. Third, the output of the automatic speech recognition and understanding (ASRU) chain is used to feed any ATM application like supporting controllers with automatic aircraft label updates. Some details of these three steps and prototypic implementations are explained in the following sections.

3.1.1 Transcription Rules for Speech-To-Text

The conversion of speech to written text should be unambiguous independent of the natural (human) or artificial (machine) transcriber in order to enable comparability and interoperability. Thus, a set of transcription rules was needed to unify and simplify the style and – more important – enable complex word transcriptions such as abbreviations, numbers, thinking sounds, partly and non-intelligible sounds, non-English words, etc. Let a controller utter an airline name, a flight number, a command type, and a command value with some of the above-mentioned complexities, then the transcription could look like one of the four examples shown in Figure 3.1.

[1]The habilitand is one of the authors of all cited 52 different references in this chapter, i.e., it summarizes the habilitation relevant work. The concept and implementation description remains at a high level to avoid extensive duplications and refers to the relevant published papers in the annex for more details.

1. Hello ~k~l~m umm three nine alpha [UNINTELLIGIBLE] descend flight level two hundred
2. Hallo klm 39A – Descent FL 200
3. hello KLM [hes] tree niner alfa descent flight lev* 2-00
4. hello KLM [hes] three nine alfa descend flight lev* two hundred

Figure 3.1: Four alternatives to transcribe the same ATC utterance.

These transcriptions have different benefits and drawbacks. While "FL 200" can be quickly written, this format does not save how the words have been pronounced, e.g., "level two zero zero" or "flight level two hundred". Furthermore, the smart usage of uppercase letters and the notation of non-intelligible sounds can help to better keep information on the actual phonetics of an utterance. The use of American or British English or the concrete spelling of words that sound the same like "alfa" and "alpha" is of minor importance, but can save a format transformation step if there are unique rules.

The correct transcription for the given example with lower case style for spoken words and upper case style for spelled letters as agreed in all recent projects focusing on automatic speech recognition and understanding [Shetty et al., 2020] is number four of Figure 3.1. The agreed rules shall make sure that all sounds of an utterance have an expression, as much information as possible on the word level is saved, and that they can easily be applied for transcriptions by humans and machines. Any hesitation sound is transcribed as *[hes]* while incomplete word pronunciations at the beginning or at the end of a word are marked with *.

3.1.2 Annotation Rules for Text-To-Concepts

The automatic transformation of text to concepts is known as automatic speech understanding. The conversion of text to ATC concepts should as well be unambiguous independent of the natural (human) or artificial (machine) annotator in order to enable comparability and interoperability. Thus, a set of annotation rules – an ontology – was needed to unify the style and to enable saving as many semantic information as reasonable from the word transcriptions [Chen et al., 2023]. In the transcription example of Figure 3.1, a royal Dutch airline flight with two digits and a letter in its flight number are forming the callsign. Furthermore, a command type and a command value are included. Possible annotation alternatives are shown in Figure 3.2.

1. klm39a descent FL 200
2. KLM39A DESCEND 200
3. KLM39A ALTITUDE 200
4. KLM39A DESCEND 200 FL

Figure 3.2: Four alternatives to annotate the same ATC utterance.

These annotations have different benefits and drawbacks. Depending on the application basing on the annotations, it can be important to know if the unit "FL" has been uttered and in which vertical direction the altitude should be modified. The correct annotation (gold annotation) for the given example as agreed with the major European ATM service providers and ANSPs under DLR leadership [Helmke et al., 2018] and used in recent ASRU projects is number four of Figure 3.2. The complete scheme for the annotation of ATS utterances with clearly defined elements and their contents is shown in Figure 3.3.

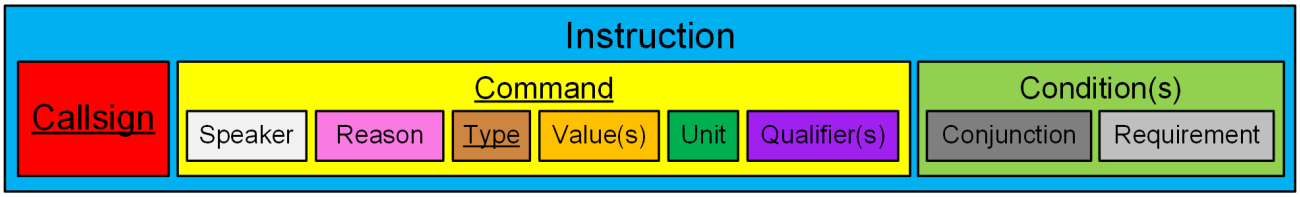


Figure 3.3: The basic scheme elements for the annotation of utterances in air traffic service communication with underlined mandatory elements.

Each instruction (blue) in an utterance must contain a callsign (red) such as *DLH7HT*, *BAW931*, *OENKF*, or *NO_CALLSIGN*. Each command (yellow) must at least contain one of the more than 120 currently defined types (brown) such as *REDUCE*, *DIRECT_TO*, *TAXI VIA*, *CLEARED TAKEOFF*, or *NO_CONCEPT*. Commands often have a single value (orange) like *200* for headings, speeds, or flight levels, *RW25R* for airport surface entities or even multiple values such as *GUNPA DEXON* for waypoints of a route. Commands can have units (dark green) such as *kt* or *ft_min* and qualifiers (purple) such as *LEFT* or *OR_BELOW*. Furthermore, the speaker (white) and reason (pink) need to be set if they are not the default values. The speaker can be *PILOT* or *ATCO* with the latter as default if nothing is explicitly set. The reason can be *REQUEST*, *REPORTING*, or empty as default for all other utterance reasons including readbacks. All instructions may have conditions (light green) that consist of a conjunction (dark grey) such as *UNTIL* or *WHEN* plus the requirement (light grey) such as a waypoint name *DL455* or *SPEED_CONVERSION*. More details on ASRU ontologies used at both sides of the Atlantic can be found in [Chen et al., 2023] and [Helmke et al., 2023b].

This ontology is one of the inputs for the European Organization for Civil Aviation Equipment (EUROCAE) Working Group 126[2] to standardize European voice communication system and ATC system integration for exchanging ATM information and Working Group 41 on Advanced - Surface Movement Guidance and Control System (A-SMGCS). The ontology further helps to build a common base for interoperability of ASRU applications as well as comparability of annotations when using defined metrics (Section 3.1.6).

3.1.3 Voice Activity and Speaker Detection, Audio Splitting, Speech Recognition

The audio data containing ATC voice utterances has different characteristics depending on the source. Usually, audio files or streams from the laboratory environment consist of a few seconds long elements that contains either an ATCo or a pilot utterance only. The begin and end of those elements can be triggered by the push-to-talk (PTT) mechanism of a headset or with a foot switch. This audio data can directly feed an ASR engine that automatically picks the best models for ATCo or pilot [Kleinert et al., 2021b]. If the audio comes from an operational ATC environment, it is sometimes provided as a continuous stream without PTT information and therefore includes utterances of an ATCo, utterances of multiple pilots, and phases of silence on the radio frequency. Then, the audio information needs to be split into short segments of distinct speakers. This is done via analyzing the frequency amplitudes and silence thresholds with a voice activity detection (VAD) to cut the relevant audio sections [Motlíček et al., 2023].

[2]This Working Group was founded with the support of DLR in August 2023.

In order to achieve high speech recognition accuracy in this case as well, again the speaker type should be known, i.e., ATCo or pilot [Helmke et al., 2021b]. Therefore, a speaker detection[3] algorithm analyzes the speaker type of the small segments [Zuluaga-Gómez et al., 2023d]. This can be based on certain keywords that correlate with either ATCo or pilot or the position of the words in an utterance, respectively [Prasad et al., 2022]. It can also evaluate the voice quality which is usually better for ATCo voice recorded on the ground than for pilot voice in noisy cockpit environment received via very high frequency (VHF) antenna on the ground as well [Zuluaga-Gómez et al., 2023c].

Finally, the speech recognition process can run on the short audio segments. The ASR's acoustic and language models are trained on ATC data[4]. As a starting point, some ATC data can be retrieved from a few publicly available ATC speech corpora. Additionally, transcribed and untranscribed speech data from the ATM target environment should be integrated into the training data to reach sufficient performance, i.e., voice recordings of controllers in a laboratory environment and voice recordings from controllers and pilots in ATC operations rooms depending on the project scope. The WER varies from 1% to 20%, i.e., the WER is generally lower if the amount of training data is bigger and the recording conditions are cleaner.

The WER is an important indicator in order to evaluate the quality of the succeeding step – the semantic understanding of the speech content. ASR is especially challenging in case of words that are rare in training data such as artificial waypoint names [Bhattacharjee et al., 2023], [Bhattacharjee et al., 2024]. Depending on the use case and environment, the speech recognition process continuously outputs the words of a received stream (online) or starts only after an utterance has been finished (offline). The same two options for online and offline processing are as well applicable to the succeeding speech understanding step, i.e., concept extraction (Section 3.1.4).

3.1.4 Air Traffic Control Concept Extraction

The basic components of an ASRU system are shown in Figure 3.4. The *audio signal* containing voice utterances from ATCos or pilots (speaker symbol in Figure 3.4) is the mandatory input for the *Speech-To-Text* block as described in Section 3.1.3. Thus, this audio input is usually already split based on PTT or VAD information. The *Speech-To-Text* block benefits from three elements: (1) a *Lexicon* with English words plus common words in ATC speech, (2) an *Acoustic Model* trained on general English and ATC audio and transcription corpora, and (3) a *Language Model* trained on ATC utterance transcriptions. The word sequence resulting from the speech-to-text process is the mandatory input for the *Text-To-ATC-Concepts* block. The *ATC Concept Extraction Model* is used by the below described ATC concept extraction algorithm to automatically extract semantic ATC concepts from the recognized word sequence. Each element (colored block in Figure 3.3) is called an ATC concept, i.e., an ATC concept can also comprise of other ATC concepts. The *Plausibility Checker* verifies if the set of commands is reasonable, e.g., does not contain two opposing *HEADING* commands and then forwards the ATC concepts to an application, e.g., display at a *controller working position* (human and monitor symbol). The *Text-To-ATC-Concepts* block is supported by the *ATC Concept Hypotheses Generator* to improve the ASRU output.

[3]The speech analysis steps before the speech understanding part were done by European ASR partners with the support of DLR, e.g., Idiap, BUT.

[4]This process is predominantly done by project partners with expertise in ASR.

With the use of *surveillance data* (aircraft symbol), *meteorological data* (thundercloud symbol), and further *Contextual Knowledge*, e.g., about the airspace and procedures, the *ATC Concept Prediction Model* helps to forecast the most probable ATC ontology concepts that will be uttered in one of the next utterances. These predicted concepts improve the performance of the *Speech-To-Text* block, the *Text-To-ATC-Concepts* block, and the *Plausibility Checker*.

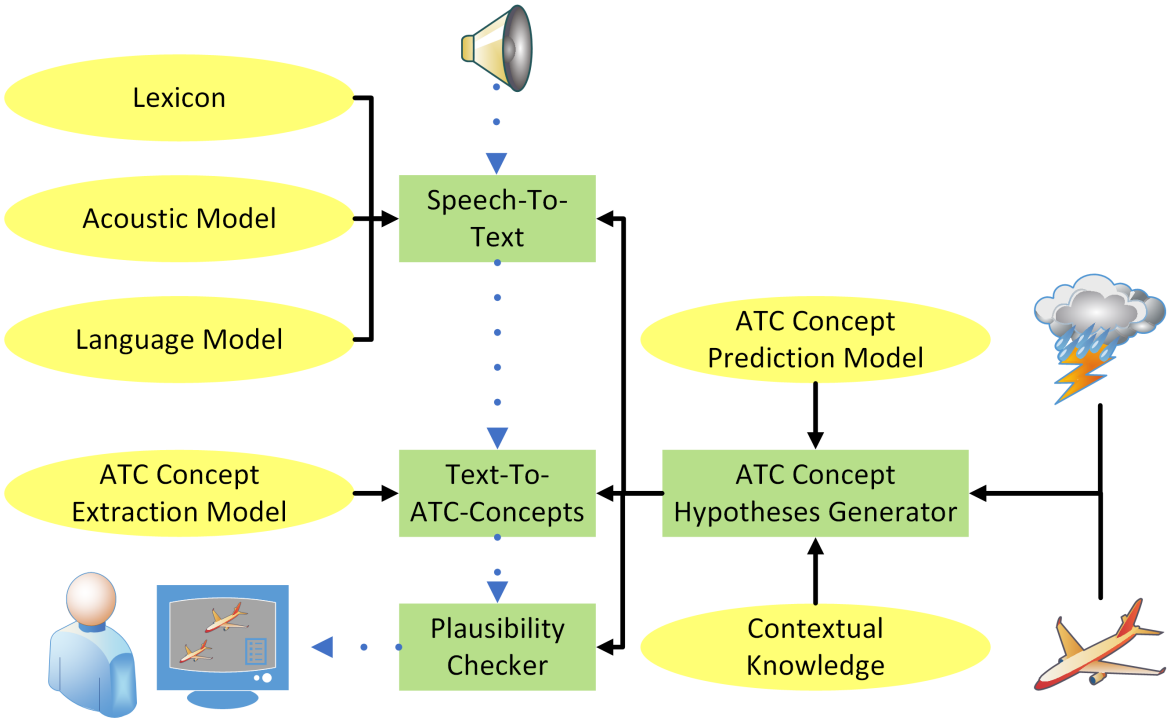


Figure 3.4: Components (green rectangle modules and yellow oval models) and data flow (black and blue arrows) of an ASRU system for ATC after [Helmke et al., 2020].

Over the last years, a sophisticated prototypic implementation of the above-mentioned ontology has been developed for the *Text-To-ATC-Concepts* block. The used rule-based *ATC Concept Extraction* algorithm has been described in [Helmke et al., 2020]. First, the algorithm tries to extract an aircraft callsign from the utterance transcription that is part of the set of predicted callsigns. Callsigns usually consist of an airline designator and two to four digits or letters in case of commercial flights. Thus, the concept extraction algorithm seeks for transcribed words such as “lufthansa” to be mapped to *DLH* or “speed bird”/“british airways” to be mapped to *BAW*. Afterwards, letters pronounced according to the ICAO alphabet such as *alfa*, *romeo*, *x-ray*, or digits from *zero* to *nine* are extracted. The callsign is one of the most important ATC concepts as it denotes the communication partner, i.e., the aircraft that is concerned with the following extracted commands.

Afterwards, the commands are extracted based on matching command keyword sequences. This step includes the extraction of types, values, units, qualifiers, conditions, etc. If all matching keyword sequences have been handled, the extraction of commands is repeated based on matching ATC concepts and on further command hints. Finally, further callsigns might get extracted from the remaining unmatched words of the transcription. All the extracted callsigns and commands together form the annotation of the transcribed utterance. Furthermore, the extracted commands with identified speaker and reason can be grouped as conversations. This means to form a thread of pilot and controller utterances regarding the same aircraft callsign that belong together content-wise. An example could be a pilot request, followed by an ATCo instruction, a pilot report, another ATCo instruction, and the required pilot readback.

3.1.5 Air Traffic Control Concept Prediction

As shown in Figure 3.4, the ATC concept extraction benefits from contextual knowledge. This knowledge based on flight plan data, radar data, meteorological data, and rather static airspace or airport dependent data. The data contains aircraft callsigns, current three-dimensional aircraft positions, speeds, headings, vertical rates, mean sea level pressure (QNH), waypoint names, arrival routes, runways, taxiways, and many more. If, for example, all aircraft in an airspace at a certain point in time are known, the callsign transcriptions that will most likely appear in the next utterances of the human ATC operators can be predicted. This tremendously helps to increase the callsign recognition rate and to reduce the callsign error rate (Section 3.1.6). The following example illustrates how contextual knowledge helps. The callsign *AUA452R* is pronounced as *austrian four five two romeo* in its long form. If there is no chance for mixing up the addressed flight, ATCos also just utter *four five two romeo*. Thus, only with the contextual knowledge, the latter utterance transcription can be mapped to the callsign annotation *AUA452R*. The contextual knowledge of ASRU can even help to correct errors from ASR. If *austrian four five three romeo* has been recognized, but *AUA453R* does not exist, the closest Levenshtein deviation match *AUA452R* can be the result of the callsign extraction process.

Furthermore, commands can be predicted in a similar way than callsigns [Ohneiser et al., 2019b]. More precisely, the knowledge about an aircraft to land at an airport soon, helps to assume that *DESCEND* commands are more likely than *CLIMB* commands. It also supports the prediction of reasonable altitude and flight level values when considering the current altitude of an aircraft. A trade-off needs to be found to make correct predictions, but to predict as few ATC concepts as possible [Ohneiser et al., 2019b]. More details about the prediction of ATC concepts in tower environments can be found in [Ohneiser et al., 2019b].

The ATC concept prediction, i.e., callsigns and commands, can be modeled with costly expert knowledge to define relevant geographical areas for command types and value limits [Rataj et al., 2019]. However, it is far more efficient to use machine learning on a set of historical data and automatically learn where which command types and values are usually given. This method allows to quickly adapt the concept prediction to new ATM environments.

3.1.6 Metrics for Speech Understanding in Air Traffic Management

The speech recognition community generally uses the metric WER (Section 2.2) to show and compare the accuracy of automatic speech-to-text transformation. However, this metric hardly provides information about the speech understanding performance that follows up on the speech-to-text output. Thus, the accuracy of extracted concepts from the text output needs to be measured [Kleinert et al., 2021a]. To make it even more complex for the ATM environment, the ATC concepts are of different value and with the exception of callsigns, it is important that all of the concepts in the command part are correct. It is for example of utmost importance – after extracting a value of *200* – to also correctly extract the command type like *REDUCE*, *DESCEND*, or *HEADING*. Commands with one or more sub-elements can be called concepts as well.

The complete performance of ASRU in ATM can be judged when calculating a concept recognition rate, a concept recognition error rate, and a concept recognition rejection rate considering each concept as an entity for the Levenshtein distance calculation.

An example in Figure 3.5 demonstrates the metrics for ATC concepts in general [Kleinert et al., 2021a]. The colored cells of the automatic annotations are always divided by the total number of cells from the gold annotation.

The recognition rate is calculated as two correct (green automatic annotations) divided by four gold annotations. The error rate is computed as two wrong (red automatic annotations) – missing or erroneous – divided by four gold annotations. The rejection rate is calculated in the same way than the other two metrics. The reason for the yellow colored automatic annotation is that the type *NO_CONCEPT* contributes to a substitution and is, therefore, counted as a rejection and not as an error.

Metric		**Calculation**	
Concept Recognition Rate		#matches / #gold	
Concept Error Rate		(#substitutions + #insertions) / #gold	
Concept Rejection Rate		#deletions / #gold	
Example			
Gold Annotation		**Automatic Annotation**	
		AFR27V DIRECT_TO DEXON none	
AFR27V INIT_RESPONSE		AFR27V INIT_RESPONSE	
AFR27V TURN LEFT		AFR27V TURN RIGHT	
AUA2EB SPEED 140 kt		AUA2EB NO_CONCEPT	
DLH7HT NO_CONCEPT		DLH7HT NO_CONCEPT	
Result			
Recognition Rate: 2 / 4 = 50%	Error Rate: 2 / 4 = 50%	Rejection Rate: 1 / 4 = 25%	on command level

Figure 3.5: Calculation of concept recognition, error, and rejection rate based on an example for ATC commands.

A true positive (TP) is defined as a concept that is correctly and fully annotated. A true negative (TN) is counted if a concept is correctly not annotated. A false negative (FN) is a concept that should have been annotated, but actually has not been. A false positive (FP) depicts a concept that is incorrectly annotated, i.e., either an actual concept is completely missing or one or more of its sub-concepts are incorrect. Furthermore, there is the total number (TA) of gold annotations. The concept error rate is then defined as FP divided by TA, the concept rejection rate is calculated as FN divided by TA, and the concept recognition rate can be computed with TP plus TN divided by TA [Chen et al., 2023]. These tailored metrics can be helpful if, for example, contextual knowledge is used to finally correct or add ASRU output, which can be relevant for the analyses in this thesis. Otherwise, precision, recall, and accuracy are also often used metrics.

The most relevant ATC concepts are callsigns and commands. A command is only considered correctly recognized if all sub-concepts such as type, unit, qualifier, or condition are correct [Helmke et al., 2021c]. The WER of the speech-to-text output is usually a good hint for the expected text-to-concepts output, but a strong correlation cannot be assumed in general [Prasad et al., 2021]. The sum of recognition, error, and rejection rate can exceed 100 % due to its definition with comparing extracted concepts with actual (gold) concepts, i.e., there is not mandatorily a one-on-one match of the number and content of both concept lists to compare.

The accuracy of predicted concepts is measured with the concept prediction error rate that is calculated as the number of extracted concepts that are not part of the predicted concepts divided by the total number of extracted concepts.

3.1.7 Application Areas of Speech Understanding in Air Traffic Management

The visualization of an ASRU system's output with three examples at a controller working position is highlighted in Figure 3.6. The transcriptions and annotations of ATC utterances are shown in different ways depending on the application's use case.

Figure 3.6: ASRU output in (1) an electronic flight strip with colored icons (lower right red box), (2) an aircraft radar label with colored values (center red box), and (3) in an outside view of a multiple remote tower setup with transcribed utterance text and annotated ATC concepts (top red box).

Over the last years there were a row of succeeding European and national funded projects on different aspects of ASRU in ATM [Rataj et al., 2021a] as outlined in the following two sections. The complete chain of airport apron and air traffic control environments for tower, approach/departure, and en-route/oceanic has been tackled by the different prototypes as shown in Figure 3.7[5]. The early projects focused on the approach environment. Later, tower ATCos and apron controllers were supported as well. Finally, the en-route/oceanic environment was covered. Furthermore, the pilot utterances were considered in addition to the ATCo utterances.

[5]The noted project acronyms are explained in Section 3.1.8 and 3.1.9

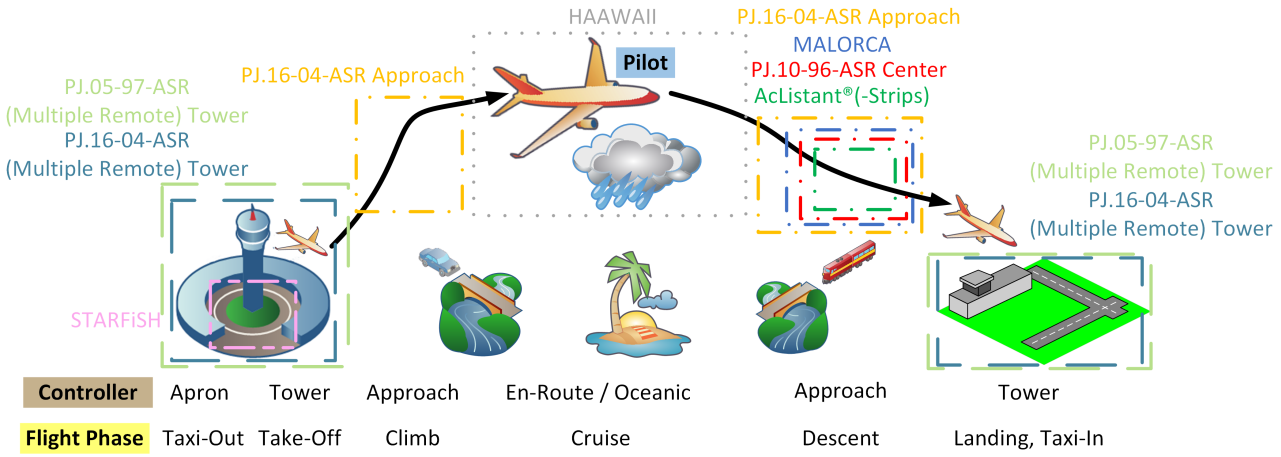


Figure 3.7: Schematic view of typical flight phases with controller type responsibility showing recent ASRU projects and their environment.

3.1.8 Prototypes in Air Traffic Control Approach and En-Route Environment

The sequence of DLR projects on ASRU in ATM began with "Active Listening Assistant" (*AcListant®* and *AcListant®-Strips*[6]). A first ASRU prototype for radio telephony communication utterances of approach ATCos was implemented. The extracted ATC concepts, i.e., callsigns, command types, and command values were displayed to support the conventional, manual maintenance of aircraft radar labels. The text-to-concepts output was also used to dynamically update the suggestions of the arrival management system, an electronic decision support tool for approach ATCos. The use case for the prototypic implementation was Düsseldorf approach.

Roughly a dozen ATCos – from Germany, Austria, Czechia, Croatia, Sweden, Denmark, Ireland – participated in the *AcListant®* and *AcListant®-Strips* validation trials, respectively. In the baseline simulation runs, ATCos needed to enter all verbally issued ATC command content manually into the aircraft radar labels. In the *AcListant®-Strips* solution simulation runs, ATCos were automatically supported in the radar label maintenance task by the ASRU system, i.e., some manual corrections were needed in seldom cases. The ASRU component achieved ATC command recognition error rates below 2% and command recognition rates above 95% [Helmke et al., 2016a]. With this, the number of inconsistent aircraft radar label content was reduced by a factor of up to two in the solution condition [Helmke et al., 2016a]. Furthermore, the human workload for aircraft radar label maintenance, i.e., the mouse clicking time, was reduced by a factor of three. The workload reduction was also measured and confirmed with NASA-TLX and ISA. The workload reduction induced a better human use of mental capacity leading to more efficient flight trajectories. The flight time per aircraft in the Düsseldorf approach area was reduced by 77 seconds in the solution condition [Helmke et al., 2017]. This can be converted to a reduced fuel consumption of more than 50 liters of kerosene and far more than savings of 100 kg carbon dioxide per flight. However, ATC experts provided a lot of input regarding the ATC concept prediction for models that are defined manually making *AcListant®* expensive to automatically roll out to other ATM environments.

[6]DLR acted as project lead; Helmholtz Validation Fund supported the project AcListant® (until late 2016).

Therefore, the *Single European Sky ATM Research Programme (SESAR)* exploratory research project *MALORCA*[7] (Machine Learning of Speech Recognition Models for Controller Assistance) proposes a general, cheap, and effective solution to automate this re-learning, adaptation, and customization process by automatically learning local speech recognition and controller models from radar and speech data recordings. *MALORCA* validated the use cases Vienna and Prague approach.

Audio data and surveillance data from five ATCos in proof-of-concept simulation trials and from twelve ATCos within the operational environment was recorded in *MALORCA*. These more than 7000 recordings contain ATC communication. The speech-to-text engine achieved a WER of 2.3%. The command recognition error rate was 0.6% for Prague approach and 3.2% for Vienna approach. The WER enables a callsign recognition rate of 98.7% and a command recognition rate of 96.5% for the recorded almost 13,000 commands [Helmke et al., 2020]. If a perfect speech-to-text engine with a WER of 0% would output the word sequences, the extraction of the text-to-concepts module would be even better, i.e., a callsign recognition rate of 99.7% and a command recognition rate of 98.8% were achieved [Helmke et al., 2020].

The SESAR2020 industrial research Solution *PJ.16-04 CWP HMI*[8] investigated different interaction technologies for controller working positions such as ASR[9]. The ASR partners of the project agreed on the ontology for annotation of ATC utterances as described in Section 3.1.2 [Helmke et al., 2018]. This was the first time that a relevant group of stakeholders defined rules for ASRU formats in ATC. An ASRU prototype for approach ATCos based on commercial-off-the-shelf software comprised the use case of Prague approach [Kleinert et al., 2019]. This prototype implemented a first version of the mentioned ontology.

Four Czech and two Austrian approach ATCos participated in the *PJ.16-04-ASR-220* TRL4 validation trials. It could be shown that commercial off-the-shelf speech recognition engines are not feasible for ATC, i.e., resulting command recognition rates were in a range of 31% to 82% for different ATCos. However, the plausibility checker module of the ASRU system significantly reduced the number of erroneous commands shown to the ATCos from around 20% to around 5% [Kleinert et al., 2019].

The SESAR2020 industrial research project *PJ.10-W2 PROSA* included an ASR activity as well in its *Solution 96*[10]. An ASRU prototype for aircraft radar label maintenance support for Vienna approach was validated on TRL6. The recognized ATC concepts from the approach ATCo utterances were shown in the radar display, e.g., the command values were highlighted in purple color in the relevant command type fields as visualized in the center red box of Figure 3.6. Those values needed to be confirmed by the ATCo through clicking on a check mark or rejected through clicking on a cross. In the most recent ASRU projects, the aircraft callsigns were as well highlighted directly after they have been uttered and recognized [Shetty et al., 2022]. Twelve approach ATCos from Austria participated in the *PJ.10-W2-96-ASR* TRL6 validation trials. The command recognition rate relevant for the aircraft radar label maintenance task was 92.5% with an error rate of 2.4% [Ahrenhold et al., 2023b].

[7]The Horizon 2020 SESAR project *MALORCA* (2016-2018) was led by DLR and was partly funded by SESAR Joint Undertaking (Grant Number 698824). Homepage: `https://www.malorca-project.de`

[8]This solution (2016-2019) with 22 European partners was led by the habilitand. It was partly funded by SESAR Joint Undertaking (Grant Number 734141). Homepage: `https://www.sesarju.eu/projects/cwphmi`

[9]This solution's activity was led by DLR.

[10]An ASRU validation exercise was led by DLR (2019-2022). It was partly funded by SESAR Joint Undertaking (Grant Number 874464). Homepage: `https://www.sesarju.eu/sesar-solutions/automatic-speech-recognition/`

The reported recognition rates seem to decrease over time for succeeding ASRU projects. However, it has to be noted that the utterance analysis became more and more holistic over time. Earlier ASRU prototypes did not consider command units and conditional clearances. The number of different command types was less. The handled command complexity was lower, e.g., corrections or commands for multiple aircraft in one utterance were not foreseen in earlier analyses. Furthermore, the recognition and error rates up to TRL4 often refer to ASRU results applied offline. Thus, this shows the rates that could potentially be achieved with enough computing time on completed utterances – especially influencing the WER of the ASR process. The online recognition rates that the human operators really "experience" in their tested prototypic applications are usually a low one-digit percent number below the offline recognition rates. Nevertheless, it is of value to analyze offline and online recognition results in order to improve their quality overall.

Within the TRL6 validation trials ATC safety could be improved due to the reduction of missing or wrong label inputs through ASRU support from 11% to 4% [Helmke et al., 2023a]. When considering additional time for mental workload of ATCos verifying ASRU output, the savings still sum up to more than one third of the time for radar label updates, i.e., the ASRU support condition requires less time for radar label maintenance than without ASRU support [Helmke et al., 2024a]. A secondary task, i.e., the Stroop test, that ATCos should only perform if their primary ATC task allows, confirmed with statistical significance that ATCos had more mental spare capacity in the solution condition with ASRU support compared to the baseline condition with solely manual label input [Helmke et al., 2023a], [Ahrenhold et al., 2023a]. Furthermore, satisfaction and trust of ATCos in the system were significantly better in the solution condition [Ahrenhold et al., 2023b]. Four ASRU use cases have been evaluated considering eight functional hazards and their defined mitigations. Supported by the results of two real-time simulations, it was shown that the use of ASRU does not increase current safety risks [Pinska-Chauvin et al., 2023]. When artificial intelligence-driven technologies like ASRU should be brought to operations rooms, such a safety assessment with analyzing normal, abnormal, and degraded modes of ATC operations is a prerequisite.

The *HAAWAII*[11] project (Highly Automated Air Traffic Controller Workstations with Artificial Intelligence Integration) developed a reliable, error resilient, and adaptable solution to automatically transcribe and annotate voice commands from ATCos and pilots in different ATC environments. Therefore, the ontology for annotation of ATC utterances was enhanced for pilot utterance concepts and en-route concepts. The output was used by the British ANSP NATS to evaluate ATCo workload of London Heathrow approach. It also built a first readback error detection prototype facing the challenge that readback errors occur very seldom in ATC communication. This prototype based on seven defined use cases including different types of readback errors such as reading back wrong values or missing readbacks [Helmke et al., 2021a]. The *HAAWAII* prototype was running in the operational ATC room of Isavia ANSP in Iceland for en-route radio telephony in June 2022. Six Icelandic en-route ATCos participated in the *HAAWAII* proof-of-concept trials. They generated verbal readback error cases to be automatically detected. The WER for ATCos was 5% and for pilots 10%. This enabled a readback error detection rate of 80% with a false alarm rate of 11%. The false alarm rate is very crucial in order to achieve a high acceptance rate of human operators. The command recognition error rate must be below 0.2% in order to achieve a false alarm rate less than 10% [Helmke et al., 2022].

[11]The Horizon 2020 SESAR exploratory research project *HAAWAII* (2020-2022) was led by DLR and was partly funded by SESAR Joint Undertaking (Grant Number 884287). Homepage: `https://www.haawaii.de`

3.1.9 Prototypes in Tower and Apron Environment

The above-mentioned Solution *PJ.16-04 CWP HMI* comprised another prototype that explored the callsign and command prediction [Ohneiser et al., 2019b] as well as recorded voice data in a multiple remote tower environment for training of automatic callsign and command extraction [Ohneiser et al., 2021b]. Each ATCo was responsible for controlling ground and air traffic on three remote airports at the same time as well as to maintain electronic flight strips of all aircraft. In order to this, the uttered tower ATCo utterances were recognized and relevant concepts were automatically extracted [Ohneiser et al., 2021c]. The ontology has been expanded with regards to tower concepts. More details on the ASRU performance can be found in:

Ohneiser, O., Sarfjoo, S., Helmke, H., Shetty, S., Motlíček, P., Kleinert, M., Ehr, H., and Murauskas, S. Robust Command Recognition for Lithuanian Air Traffic Control Tower Utterances. In *22nd Annual Conference of the International Speech Communication Association, Proc. of Interspeech 2021*, pages 3291–3295, Brno, Czechia (hybrid), 30 Aug - 03 Sep, 2021. ISCA.
`https://doi.org/10.21437/Interspeech.2021-935` reprinted as article →**1** in the annex

More than a dozen of tower ATCos from Hungary, Lithuania, and Finland participated in human-in-the-loop trials for *PJ16-04-ASR-240*. The command prediction error rate for the multiple remote tower environment was around 7.5% [Ohneiser et al., 2019b]. More details on the technical ASRU background and results can be found in:

Ohneiser, O., Helmke, H., Shetty, S., Kleinert, M., Ehr, H., Murauskas, S., and Pagirys, T. Prediction and extraction of tower controller commands for speech recognition applications. ***Journal of Air Transport Management***, 95:102089, 2021.
`https://doi.org/10.1016/j.jairtraman.2021.102089` reprinted as article →**2** in the annex

The SESAR2020 industrial research project *PJ.05-W2 DTT* also included an ASR activity in its *Solution 97*[12] and validated an ASRU prototype for electronic flight strip maintenance support in a multiple remote tower environment on TRL4 [Ohneiser et al., 2022]. More details on different ASRU validation exercises of European project partners that also used the defined ontology[13] for ATC utterance annotations can be found in [Ohneiser et al., 2022]. The extracted concepts were used to highlight recognized callsigns and to input commands in a highlighted color at the flight strip display. These commands automatically entered the ATC system if they were not corrected by the ATCo within ten seconds. The automatic acceptance goes back to earlier ATCo feedback in an approach environment [Helmke et al., 2016a]. Five Austrian and five Lithuanian tower ATCos participated in the *PJ.05-W2-97* TRL4 validation trials [Ohneiser et al., 2023]. A reduction of ATCo workload and an improvement of usability have been measured based on the prototype's callsign recognition rate of 94.2% and a command recognition rate of 82.9%. More details about the validation setup and results can be found in:

[12]The project (2019-2022) was led by DLR; an ASRU validation exercise was led by the habilitand as explained by him in a video: `https://s.dlr.de/bzxME`. The project was partly funded by SESAR Joint Undertaking (Grant Number 874470). Homepage: `https://www.remote-tower.eu/wp/project-pj05-w2/`

[13]Mainly in the course of SESAR projects there is a row of ATM system suppliers, ANSPs, and research institutions that is as well involved in applying the ontology to different ASRU prototypes such as CRIDA, DFS, HungaroControl, Indra, Leonardo, and Thales. MITRE from the US compared the European ontology with theirs.

Ohneiser, O., Helmke, H., Shetty, S., Kleinert, M., Ehr, H., Schier-Morgenthal, S., Sarfjoo, S., Motlíček, P., Murauskas, S., Pagirys, T., Usanovic, H., Meštrović, M., and Černá, A. Assistant based speech recognition support for air traffic controllers in a multiple remote tower environment. ***Aerospace***, Special Issue *Automatic Speech Recognition and Understanding in Air Traffic Management*, 10(6), 2023. `https://doi.org/10.3390/aerospace10060560` reprinted as article →**3** in the annex

The integration of artificial intelligence based ASRU into decision support systems of flexible endorsed ATCos is explored for remote tower centers [Meier et al., 2024]. The *STARFiSH*[14] project (Safety and Artificial Intelligence Speech Recognition) developed an ASRU prototype to support apron controllers at Frankfurt airport. Again, the ontology has been enhanced considering concepts of apron control. The ASRU system was integrated with an A-SMGCS [Kleinert et al., 2023]. Thus, apron controllers are supported through visualizing their verbally instructed taxi routes on a situation display. The simulation pilots are as well supported through automatically recognized callsigns and commands, i.e., commands are automatically executed in case simulation pilots do not veto [Kleinert et al., 2022]. 14 German apron controllers participated in the *STARFiSH* human-in-the loop trials in the Fraport simulator at Frankfurt airport [Kleinert et al., 2023]. The integration of A-SMGCS data into the ASRU component improved the command recognition rate by more than 15% absolute, i.e., a command recognition rate of 91.8% and a callsign recognition rate of 97.4% were achieved based on a WER of 3.1% [Kleinert et al., 2023]. This performance led to a significant reduction of controller workload and an increase in safety. Currently, ASRU is developed further to automate aircraft pushbacks[15].

3.1.10 Text-To-Speech Application for Air Traffic Control Utterances

The training of aviation communication operators often comes with the use of intense human resources even if their tasks could be automated. Based on open-source ATC communication datasets, 20 ATCo voice models and 8 pilot voice models have been fine-tuned for a TTS application using an end-to-end method and a voice cloning method, respectively [Ohneiser and Ahmed, 2025]. This TTS application was able to simulate aviation radio telephony communication operators – ATCo instructions and pilot responses – based on textual ATC utterances. The application's synthesized speech has been validated online by 20 international aviation experts including 14 ATCos and 4 pilots from Germany, Austria, France, the Netherlands, Egypt, and Pakistan. The study subjects provided more than 4100 ratings on overall experience, clarity, pronunciation, intonation, naturalness, and speed of generated speech. The voice cloning models – especially the female voices – received significantly better ratings than the end-to-end models. The speech has been synthesized faster than real-time. While some issues with pronunciation of waypoint names non-existing in training data and with pauses existed, the overall feedback demonstrated feasibility and realism of the artificial voices as discussed in:

Ohneiser, O., Ahmed, U. Text-To-Speech Application for Training of Aviation Radio Telephony Communication Operators. ***IEEE Transactions on Aerospace and Electronic Systems***, 61(2), pages 4542–4560, 2025.
`https://doi.org/10.1109/TAES.2024.3504493` reprinted as article →**4** in the annex

[14]The project STARFiSH (2020-2022) was funded by the German Federal Ministry of Education and Research (BMBF).
[15]The project AeM-Speedport (2024-2026) is funded by the German Federal Ministry of Digital and Transport (BMDV): `https://s.dlr.de/XPZb2`

3.2 Tactile Modality: Gesture Recognition and Tactile Cues for Aviation Purposes

There have been experiments on the usability of multitouch devices for interaction of ATCos with the ATC system. First ideas explored an overlay display on a radar screen background to (1) select aircraft icons with a one-finger tap, (2) select command types with a two-dimensional gesture of one or more fingers, and (3) select command values with interactive sliders and menus. These initial experiments revealed that the human operator's arm and hand as well as the overlay interaction menus might hide too much relevant information from the radar screen below. Furthermore, a "gorilla arm" was suspected in order to often use the arms for data input without any hold. It turned out that the two-dimensional gestures to depict the ATC command types could be beneficial especially if they are executed intuitively and without looking on the multitouch device.

During the above reported *AcListant®-Strips* validation trials a second solution condition was tested. In the above-described solution condition, ATCos needed to use the mouse in order to correct and confirm ASRU output. In the second solution condition, a multitouch device was used to correct and confirm ASRU output through swipe and single tap gestures of an approach ATCo [Helmke et al., 2016a]. The perceived workload as measured with NASA-TLX was highest in the baseline condition with a value of 87.0 where ATCos used the mouse for all inputs. The workload score for ASRU plus manual correction only was much lower, i.e., corrections via a mouse were slightly better with a value of 74.2 than via multitouch with a value of 78.5 [Helmke et al., 2016b]. There was a row of measurement categories in which the use of multitouch outperformed the use of a mouse. The time needed to solve a secondary task was 15% less in average with multitouch than with a mouse for ASRU corrections [Helmke et al., 2016b]. Also, the deviations in aircraft radar labels between the correct and the actual input was much lower with multitouch compared to a mouse for ASRU correction, more precisely, the number of inconsistent label states was 7.5% less and the duration of inconsistencies was 10 seconds less [Helmke et al., 2016b].

Seventy US American general aviation pilots participated in a human-in-the-loop validation study at FAA[16] on weather state-change notifications [Ahlstrom et al., 2015]. The study subjects flew a simulated aircraft under visual meteorological conditions and should avoid significant precipitation cells based on weather symbology on different maps and their dynamic changes. The experimental group wore a vibrating bracelet notifying subjects about visualized weather state-changes. This tactile cue – compared to visual cues only – improved weather situation awareness of pilots and reduced their mental workload even if experimental group and control group kept similar distances to hazardous weather.

3.3 Visual Modality: Eye Tracking, Display Cues, and Attention Guidance for Air Traffic Control Centers and Cockpit

All visual elements and information presented on controller screens such as air situation displays shall support decision making of human operators. Some visual cues indicate recommended times and content of aviation actions such as uttering ATC commands. However, there is no assurance that the relevant displayed information is also noticed at the right time.

[16]This work has been supported by the habilitand in the course of a research semester at FAA William J. Hughes Technical Center in Atlantic City, NJ, USA.

The eye tracking technology is used to determine where human operators in aviation are looking at. For example, a low-budget infrared eye tracker that is mounted at the bottom of a display analyzes at which aircraft radar label or aircraft icon a controller is looking at. An accuracy of around one centimeter on the screen is sufficient for this use case as the aircraft radar labels seldom overlap each other. If it is assumed that the current gaze point equals the point of the human's current visual attention, eye tracking serves as a proxy for analysis of visual attention. Similarly, mouse tracking can complement the determination of visual attention. Knowing the current visual attention is valuable because it can help to predict the next actions of human operators. If a controller is visually scanning an aircraft radar label, the likelihood of this aircraft receiving an ATC command in the near future is high. If a controller does not visually scan a displayed airspace area or an aircraft radar label, the current importance of this element for the controller can be assumed to be low. However, if an assistant system recommends a visual check of such an area, measures to guide the attention can be initiated.

ATCos not only need to assure safe but also efficient air traffic, which can be achieved by adhering to optimized target times of aircraft at waypoints. The following sections outline visual display aids to meet target times, explain visualized re-routing advisories and weather data in ATCo displays, show how eye tracking can support to guide ATCo attention, and connect eye tracking data with ASRU.

3.3.1 Visual Display Aids for Approach Controller Support

Approach ATCos at highly congested airports need to take care of issuing instructions to pilots on time to avoid any unnecessary delay in the air traffic flow. One of the most critical phases is merging aircraft streams from different cardinal directions into one final arrival stream.

In a simplified prototype with three aircraft streams, five different visual display techniques as shown in Figure 3.8 have been implemented to support merging into one final arrival stream in a time-based manner [Ohneiser et al., 2018a]: (1) a *Baseline* display with current time and target times of aircraft; (2) a *Time-To-Gain/Lose* display that shows alphanumeric values on accuracy in seconds at each aircraft label; (3) a *Slot Marker*, i.e., a circle indicating the optimum position of an aircraft in order to meet its time constraints; (4) a *TargetWindow* on a centerline to present the final positions of all aircraft in the merged arrival stream; and (5) a *Timeline* display with aircraft strips assigned to a scheduled time to be compared with the aircraft strips on projected times. The simplified aircraft labels consist of a callsign with abbreviated cardinal direction for North (N), Center (C), or South (S not shown) and a number in the first line as well as the current ground speed value in knots in the second line.

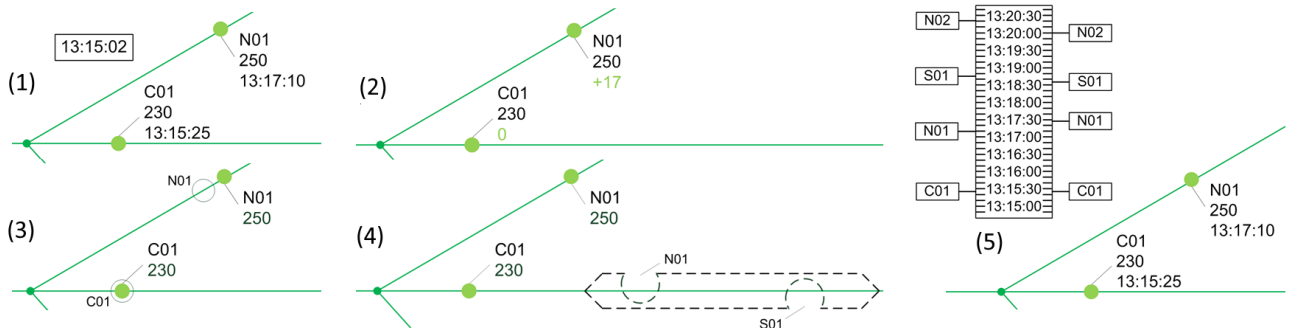


Figure 3.8: Time-based visual display aids: *Baseline* (top left), *Slot Marker* (bottom left), *Time-To-Gain/Lose* (middle top), *TargetWindow* (middle bottom), *Timeline* (right).

More details about the visual display aids and evaluation results can be found in:

Ohneiser, O., Ahlstrom, V., Tracy, K., and Williams, B. Comparison of Air Traffic Controller Display Techniques for Reaching Target Times at Significant Waypoints. In *IEEE/AIAA 37th Digital Avionics Systems Conference, DASC 2018*, pages 1092-1101. London, UK, 23-27 Sep, 2018. `https://doi.org/10.1109/DASC.2018.8569365` reprinted as article →**5** in the annex

Fifteen US American ATC experts and one ATCo participated in a human-in-the-loop simulation at Federal Aviation Administration (FAA) regarding five different display-techniques for time-based merging of arrival streams [Ohneiser et al., 2018a]. The participants had to adjust aircraft speeds in order to meet the aircraft target times. Most of the participants performed best with the *Slot Marker* aid, i.e., the aircraft reached their target times accurately. However, the number of speed commands was also higher as for other visual aids. The higher the number of speed changes, the more inefficient the flight profile. When analyzing the weighted performance scores, the *Timeline* and *Time-To-Gain/Lose* aid also led to good performances.

3.3.2 Situation Awareness Aspects for Aviation Weather

The projects *MET4ATM*[17] (Meteorology for Air Traffic Management) and *SINOPTICA*[18] (Satellite-borne and In-situ Observations to Predict The Initiation of Convection for ATM) integrated advanced meteorological information into the visual modality elements of controller working positions as shown in Figure 3.9.

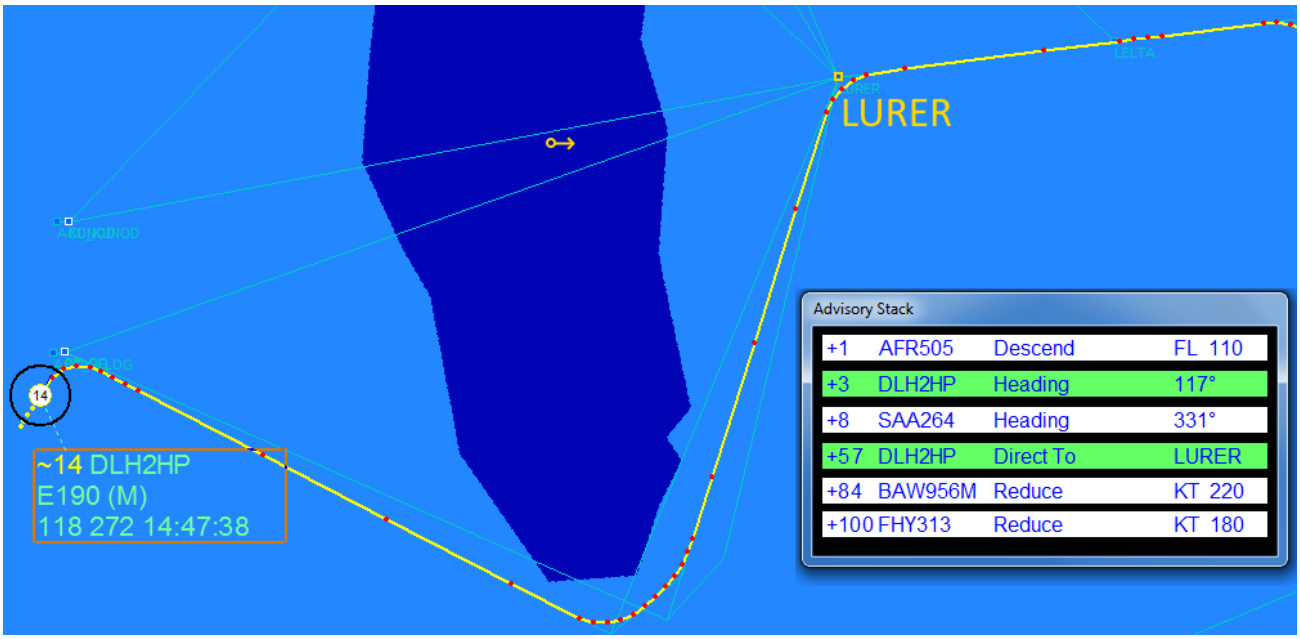


Figure 3.9: Re-routing trajectory (yellow) of aircraft *DLH2HP* around severe weather cell (dark blue) with ATC advisories (green highlighted commands in *Advisory Stack*) after [Ohneiser et al., 2019c].

[17]The German national project MET4ATM (2016-2018) was partly funded by the Federal Ministry for Economic Affairs and Energy/LuFo.

[18]The Horizon 2020 SESAR research project SINOPTICA (2020-2022) was partly funded by SESAR Joint Undertaking (Grant Number 892362).

The aircraft *DLH2HP* with landing sequence number *14* on the left side of Figure 3.9 is affected by severe weather as shown with the yellow tilde symbol. The yellow line with red dots represents the two-dimensional view of a calculated four-dimensional re-routing trajectory around the severe weather cell. The balance point of the weather polygon is shown with a yellow circle, the predicted direction and speed of movement are visualized with the orientation and length of the yellow arrow. The *Advisory Stack* on the lower right side gives time-based recommendations for ATC commands to ATCos to implement the calculated plan. More precisely, for *DLH2HP* it recommends a heading change to *117* degrees in a few seconds and roughly one minute later flying directly towards the waypoint *LURER* to rejoin the original route.

The weather visualization implementation also includes dynamic updates of convective weather cells morphing from the current to the predicted shape integrated into an air traffic radar display [Ohneiser et al., 2019c]. Furthermore, automatic detour planning with reroute command advisories resulting in updated four-dimensional aircraft trajectories to avoid convective weather are calculated and displayed in real-time [Ohneiser et al., 2019c]. More details about concept and visualization of weather and re-routes can be found in:

Ohneiser, O., Kleinert, M., Muth, K., Gluchshenko, O., Ehr, H., Groß, N., and Temme, M.-M. Bad Weather Highlighting: Advanced Visualization of Severe Weather and Support in Air Traffic Control Displays. In *IEEE/AIAA 38th Digital Avionics Systems Conference, DASC 2019.* San Diego, CA, USA, 08-12 Sep, 2019.
`https://doi.org/10.1109/DASC43569.2019.9081773` reprinted as article →**6** in the annex

It has been evaluated in the simulated Milano-Malpensa airspace with two parallel runways if the situation awareness of controllers can be increased by these advanced weather visualization techniques [Temme et al., 2023]. Five controllers from Germany, Austria, and Poland participated in the *SINOPTICA* human-in-the-loop simulation [Temme et al., 2023]. The arrival management system used in the simulation provided fly-around routes to avoid severe weather areas. The human operators appreciated the support in approach planning and re-routing from the visual and from the planning point of view [Temme et al., 2023].

Improved weather visualization has also been analyzed in cockpits [Ahlstrom et al., 2016]. More precisely, the effect on general aviation pilot behavior when using portable weather applications was investigated in a simulator study at FAA[19].

The portable device visualized the planned flight route as well as any hazardous weather to improve weather situation awareness as well as decision making, cognitive engagement as measured with functional near-infrared spectroscopy methods, and the resulting distance to convective weather.

Seventy US American pilots participated in a human-in-the-loop validation study at FAA for mobile weather applications in the cockpit [Ahlstrom et al., 2016]. The solution group used a mobile weather app, the baseline group did not. The use of portable applications did not degrade pilot performance in safety related tasks. It increased weather situation awareness and led to larger distances from hazardous weather in the solution group compared to the baseline group [Ahlstrom et al., 2016].

[19]This work has been supported by the habilitand in the course of a research semester at FAA William J. Hughes Technical Center in Atlantic City, NJ, USA.

3.3.3 Attention Guidance for Human Operators in Aviation

If eye tracking measurements can be used to deliver the actual visual attention of a human operator, this data can serve as an input to also guide the operator's attention. The attention guidance functionality defined herein contains of three steps [Rataj et al., 2021b] – with actual attention measurement being step 1. For step 2, it must be determined where the current target attention should be, i.e., a display area that the operator should visually check. This information can be derived with the support of electronic assistant systems such as arrival manager, departure manager, or surface manager in case of an ATCo. For example, the current attention should focus on two aircraft involved in a short-term conflict or on an aircraft about to be handed over into the responsibility of another airspace sector.

If steps 1 and 2 are executed – independently of their order – and there is a mismatch between the current attention area and the target attention area, step 3 consists of taking measures to smoothly guide the human operator attention. The attention guidance mechanisms can be auditory, tactile, or visual. While auditory and tactile measures are independent of the human's field of view, the visual attention guidance mechanism is more precise if it comes to certain areas of interest in a display. Before actually executing the attention guidance cues, it should be evaluated by considering the current progress of the operator's scan pattern if the target area of attention will be reached soon. If so, it might be less disturbing for the human operator to wait until escalating attention guidance cues.

3.3.4 Attention Guidance Prototype for Flight-Centric Air Traffic Control

An attention guidance (AG) prototype has been developed in the project *PJ.16-04 CWP HMI AG*[20] for the use case of flight-centric ATC in Hungary's upper airspace. Several upcoming ATC events such as handovers as well as medium- and short-term conflict alerts are weighted given their urgency and importance [Ohneiser et al., 2018c]. If the AG mechanism is started, visual cues are escalated as shown in Figure 3.10 [Rataj et al., 2021b]. Level 0 includes the default view. In escalation level 1, a rectangle around an aircraft radar label is highlighted.

In escalation level 2, an additional semi-transparent circle highlights the aircraft icon. In escalation level 3, a glowing effect strengthens the visual cuing effect of the semi-transparent circle to draw the human operator's attention. The highlighting immediately disappears if the controller visually checks the relevant display spot.

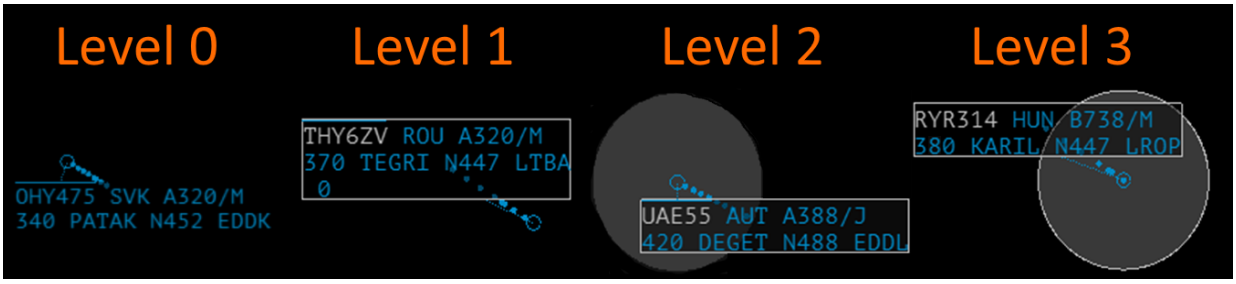


Figure 3.10: Escalation levels of attention guidance prototype for air traffic control events with their visual cues.

[20]This solution's activity with 11 European partners was led by the habilitand.

More details about the attention guidance concept and validation results can be found in:

Ohneiser, O., Gürlük, H., Jauer, M.-L., Szöllősi, Á., and Balló, D. Please have a Look here: Successful Guidance of Air Traffic Controller's Attention. In *9th SESAR Innovation Days, SID 2019.* Athens, Greece, 02-05 Dec, 2019.
`https://s.dlr.de/qyGZM` reprinted as article →**7** in the annex

Five Hungarian ATCos participated in the *PJ.16-04-AG* trials at HungaroControl [Ohneiser et al., 2019a]. The ATCos perceived less workload and improved situation awareness when they got support from the attention guidance mechanisms compared to no support at all [Ohneiser et al., 2019a]. They also rated acceptance and confidence with the attention guidance condition higher as without. They appreciated the smooth nature of attention guidance via visual cues very much.

3.3.5 Vigilance and Attention Controller

The *MINIMA*[21] project (Mitigating Negative Impacts of Monitoring high levels of Automation) developed adaptive automation functionalities for an approach ATCo air situation display. The adaptive automation was implemented in an assumed future environment where the human operator primarily does monitor the traffic instead of interactively guiding air traffic and speaking to flight crews [Berberian et al., 2017]. The use case comprised arriving and departing aircraft for two independent parallel runways. The adaptive ATCo tasks and support activation were triggered by a vigilance and attention controller consisting of two modules. The first module was an eye tracking based attention controller that guided the ATCo attention to display spots of upcoming ATC events and to seldom visually scanned aircraft via different visual cues [Ohneiser et al., 2018b]. The second module was an EEG based vigilance controller. The brain activity data from the head-mounted EEG activated or deactivated ATCo support functionality [Di Flumeri et al., 2019]. If the human vigilance decreased, more tasks were allocated from the machine back to the human operator in order to keep the human better in the mental loop. If the vigilance increased, some tasks were allocated back from the human operator to the machine.

Examples for those dynamically allocated tasks or offered support tools are the sector size, i.e., the area of ATCo responsibility, manual/automatic assuming of aircraft, displaying/hiding ATC command advisories, displaying/hiding inter-aircraft distances on the runway centerlines, situation awareness questions to be answered, and the visual attention guidance cues [Ohneiser et al., 2018b]. Fourteen Italian ATCos participated in the *MINIMA* proof-of-concept trials [Di Flumeri et al., 2019]. While the human operators' vigilance decreased over time in the monitoring task as intended, the decrease was much less in the solution condition compared to the baseline condition [Di Flumeri et al., 2019]. More precisely, the EEG-based vigilance score after 40 minutes within a simulation run decreased by almost 50% compared to the start of the simulation run in the baseline condition whereas it decreased by only 21% in the solution condition. The operator involvement in the baseline condition was much less compared to the solution condition in which the vigilance and attention controller component actively managed the assignment of tasks to the human and the machine as well as adapted the availability of automation support [Di Flumeri et al., 2019].

[21] The Horizon 2020 SESAR exploratory research project MINIMA (2016-2018) was led by DLR – project responsibility was assumed by the habilitand in the project's fade-out phase – and was partly funded by SESAR Joint Undertaking (Grant Number 699282).

3.3.6 Improve Air Traffic Control Concept Prediction

As outlined in Section 3.1.5, the accuracy of ATC concept prediction improves the accuracy of ATC concept extraction. The visual attention of a controller on specific areas on the radar screen, e.g., aircraft labels, indicates that the characteristics of this very aircraft is currently analyzed to derive any future instruction. An eye tracker can evaluate how often, how long, and how recent a controller looked at a certain aircraft radar label. The most recent watched labels with their aircraft callsigns as well as commands for this aircraft are likely candidates to be verbalized in the next utterances.

Thus, callsigns and commands can be assigned with a higher probability to influence the automatic recognition of speech and extraction of concepts [Ohneiser et al., 2021a]. Hence, eye tracking, i.e., fixations, can serve as a proxy to predict the next, most likely ATC concepts that will be uttered.

Two German controllers participated in a human-in-the-loop simulation regarding the prediction of ATC concepts [Ohneiser et al., 2021a]. The probabilities for predicted commands improved by a factor of four when considering eye tracking data of ATCos [Ohneiser et al., 2021a]. The error rate of the predictions was reduced by a factor of 25 regarding the three most probable aircraft to be uttered in the next commands [Ohneiser et al., 2021a]. More details on improving ATC concept prediction and visual verification of ASRU output can be found in:

Ohneiser, O., Adamala, J., and Salomea, I.-T. Integrating Eye- and Mouse-Tracking with Assistant Based Speech Recognition for Interaction at Controller Working Positions. ***Aerospace***, Special Issue *Aeronautical Informatics*, 8(9), 2021.
`https://doi.org/10.3390/aerospace8090245` reprinted as article →**8** in the annex

3.3.7 Support Verification of Displayed Speech Recognition and Understanding Output

As mentioned in Section 3.1.7, the ASRU output for aircraft radar label maintenance is visualized with highlighted values in the air situation display. Earlier feedback of ATCos indicated that they do not want to confirm the output by clicking on an aircraft radar label check mark each time as the majority of outputs is correct [Helmke et al., 2016a].

In a row of further validation trials with other ASRU prototypes, ATCos stated that they would prefer having a time to correct or reject the ASRU output with succeeding automatic entry of the electronic ATC system.

However, it would be beneficial to check if the ATCo at least visually scanned the output value to replace the active mouse click confirmation. Again, the fixations on aircraft radar labels as determined by an eye tracker can support such a functionality. If the ATCo did not look at the radar label with ASRU output, the visual saliency can be escalated until the output may be discarded if visually unchecked in the final escalation step [Ohneiser et al., 2021a].

The eye tracking-based verification support and the prediction support are already a combination of two modalities to use the mutual benefits. Two German controllers tested the visual verification of ASRU output [Ohneiser et al., 2021a]. This worked technically well and might replace manual click-based confirmation of ASRU output [Ohneiser et al., 2021a].

3.4 Multimodality: Combination and Integration of Three Interaction Technologies for Controller Working Positions

After exploring the benefits and drawbacks of the auditory, tactile, and visual interaction, a combination of them promised to reveal further potential. The feasibility of the three modalities has been assessed for the three basic ATC instruction parts to span a matrix as shown in Figure 3.11.

		Command Element		
		Callsign	Type	Value
Interaction Technology	Speech Recognition & Understanding	medium	medium	good
	Gesture Recognition	medium	good	medium
	Eye Tracking	good	poor	poor

Figure 3.11: Assessment of interaction technology feasibility for human input of ATC command elements after [Ohneiser et al., 2016].

Eye tracking can poorly be used to select values or types from a long list. However, it worked well for visually selecting aircraft radar labels that are usually separated from each other on the display and have a maximum number of around a dozen. Gestures to select aircraft via icons as well as to select values via sliders or from a list were feasible. However, they were ergonomically unfavorable and hid important information in the background. Thus, intuitive no-look gestures – restricted in their number – to be performed on an extra multitouch device were well usable to identify command types such as swiping down/up for *DESCEND/CLIMB* or swiping left/right for *REDUCE/INCREASE*. ASRU basically works well on callsigns with contextual knowledge as well as on types or numbers.

However, usually the values in a reasonable unit are uttered in natural communication. For example, when asking somebody for the way to the airport, the response typically includes looking in the direction of the airport, performing a hand/arm gesture towards the airport walkway, and verbalizing the distance value, e.g., in kilometers. Thus, ASRU promises to be most intuitive for the recognition of command values when combining three different modalities. The green diagonal from the matrix (Figure 3.11) was chosen for the concept of a multimodal controller working position [Ohneiser et al., 2016]. The three interaction technologies and three ATC concepts were integrated for the *TriControl*[22] prototype as sketched in Figure 3.12 and detailed in:

Ohneiser, O., Jauer, M.-L., Gürlük, H., and Uebbing-Rumke, M. TriControl – A Multimodal Air Traffic Controller Working Position. In *6th SESAR Innovation Days, SID 2016*. Delft, The Netherlands, 08-10 Nov, 2016.
`https://s.dlr.de/bM1th` reprinted as article →**9** in the annex

To form an instruction, the ATCo looks at an aircraft label on the radar screen. The eye tracker detects the gaze on the label and uses the aircraft callsign, e.g., *DLH4TN* to send the next instruction to. The ATCo also performs a gesture on the multitouch device to instruct the command type, e.g., *swipe down*. The value needs to be spoken by the ATCo, e.g., *six zero*. The gaze, the gesture, and the speech can be done in sequence or in parallel.

[22]The number *three* is pronounced as *tri* in ATC radio telephony and serves for the prototype name.

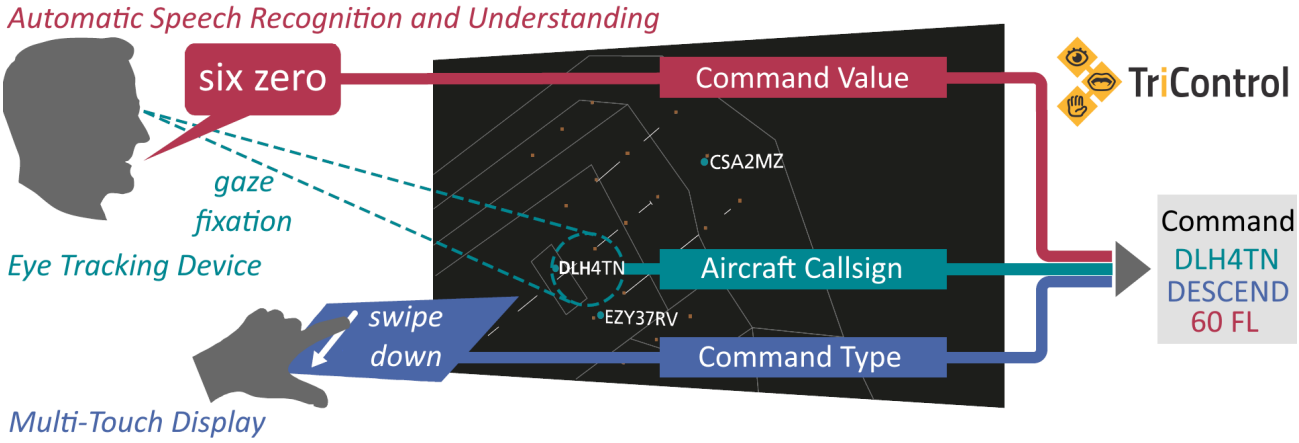


Figure 3.12: Scheme of the multimodal controller working position *TriControl* to combine command elements as entered via different interaction technologies.

Finally, the completed instruction is checked for plausibility of the three single concepts and the combined command[23] in order to increase robustness. Thus, the only reasonable unit FL for the given command is added automatically. The final command, e.g., *DLH4TN DESCEND 60 FL* must be confirmed with a one-finger tap gesture.

The instruction can then be up-linked to the aircraft or read via text-to-speech functionality on the radio frequency. This multimodal working position can be seen as the overarching prototype for the investigated interaction technologies in this thesis as shown in Figure 3.13.

Figure 3.13: Validation setup of the multimodal controller working position prototype *TriControl* integrating eye tracking, gesture recognition, as well as speech recognition and understanding.

[23]The habilitand co-holds the European patent No. 17158692.8 on an *Assistance Based Controller Command Creator*

A human-in-the-loop simulation study has been conducted with the *TriControl*[24] prototype in order to assess the current technical feasibility [Ohneiser et al., 2020] and to evaluate a potential speed gain in entering instructions into the electronic ATC system [Ohneiser et al., 2018d].

The *TriControl* prototype has been evaluated in proof-of-concept trials with fourteen ATCos from the German ANSP at the DFS site. The participants' performance showed potential for a 15% speed gain to enter ATC commands multimodally into the ATC system compared to conventional manual input mode [Ohneiser et al., 2018d].

However, the variability in results was high for the solution condition, i.e., some ATCos performed command input with *TriControl* much faster than with the baseline system while others were slower compared to the quite homogeneous performance with the baseline input. This could indicate that some more training is needed with the *TriControl* system in order to achieve a more homogeneous performance of human operators. Some more details on those results can be found in:

Ohneiser, O., Jauer, M.-L., Rein, J. R., and Wallace, M. Faster Command Input Using the Multimodal Controller Working Position "TriControl". ***Aerospace***, 5(2), 2018.
`https://doi.org/10.3390/aerospace5020054` reprinted as article →**10** in the annex

The active approach ATCos rated system usability, ease of use, user-friendliness, learnability, and acceptance of the early-stage prototype very positive [Ohneiser et al., 2020]. They even stated to be willing to use such a system in their daily ATC work. A row of further aspects of feasibility of this prototype can be found in:

Ohneiser, O., Biella, M., Schmugler, A., and Wallace, M. Operational Feasibility Analysis of the Multimodal Controller Working Position "TriControl". ***Aerospace***, 7(2), 2020.
`https://doi.org/10.3390/aerospace7020015` reprinted as article →**11** in the annex

3.5 Limitations of Human-in-the-loop Simulation Setups

All validated research prototypes were not intended to be as robust as they would need to be in operations, i.e., a re-start of simulation runs have happened once in a while. The working positions of controllers and pilots were designed to provide a look-and-feel as close to operations as reasonable.

However, as the operational working positions differ, e.g., between ANSPs across Europe and even inside a country, the only goal for the simulations could be to find commonalities and to provide layout and tools that come close to the majority of working positions. Hence, the main part of adjusting the simulation environment usually contains of adapting the air situation data displays, e.g., showing radar data of aircraft.

With this, also the tasks and the assistance systems that human operators perform in real operations differ from those during the simulations. Furthermore, there was no coordination between ATCos working at the same working position, between ATCos from different air traffic service units, and from adjacent airspace sectors or between pilots and co-pilots.

[24]The simulation study setup and basic functionality of *TriControl* can be seen in this video: `https://s.dlr.de/sLmxi`

Further restrictions mainly apply due to the limited resources to run simulations, i.e., the budget and time required for expensive professional study subjects like ATCos or pilots, for validation staff, and for the maintenance and operation of high-fidelity validation infrastructure. Telephone surveys in other domains usually have more than 1000 participants in order to generate representative results. This number would mean that 50% of the German ATCos would need to travel to and take part in a study that typically requires the presence on one or two simulation days. As this is not feasible with the given resources, the number of ATCos participating in a simulation study is usually around ten.

All study subjects undergo a training at the prototypic working position at the beginning of each study. However, the amount of time was limited and not comparable to operational trainings regarding newly introduced features and technologies at working positions. In addition, the number of different air traffic scenarios and their content that can be part of simulations is small. Thus, simulations often focus on one or two baseline and solution runs to be compared. Non-nominal circumstances such as emergency flights, runway closures, separation infringements, etc. can only be integrated in a very small number and must be well-planned in order to guarantee comparability between different study subjects' results and to keep a high degree of realism in the simulation. The simulation realism is as well affected by further aspects such as individual characteristics of human simulation pilots that communicate with ATCos, the familiarity of study subjects with an airport or airspace setup, the study subjects' experience with tested technologies, and the general attitude towards simulations. Hence, it is often difficult to achieve statistically significant results, i.e., where the performance and judgments of human aviation operators and the overall system performance is far away from any ambiguity.

Still the early and iterative involvement of system matter experts with recording of objective and subjective data tremendously helps to build prototypes that will be accepted by human aviation operators in later operations after they have been tested in field trials on higher TRLs. The used simulation setups are repeatedly confirmed to be close to real operations. Each validation campaign and each tested software iteration reveals aspects for improving the studied prototypes in order to minimize the gap between pre-industrial prototypes and potential future applications.

4 Summary and Conclusions of Validations with Outlook on Future Work

This cumulative habilitation thesis described multimodal interaction technologies at aviation working positions to support human operator awareness regarding the air traffic situation, different types of weather data, and pending air traffic events in order to instruct timely verbal commands to pilots and to manually feed digital air traffic management systems with the command content. The following three sections summarize the developed and validated prototypes, conclude the main results, and give an outlook on future directions.

4.1 Summary of Activities on Aviation Interaction Technology Prototypes

The prototypes of the reported validation studies usually make use of hardware and basic software that already exists for automatic recognition of speech, gestures, and gazes. The applied research prototypes built advanced software functionalities on top using the air traffic domain knowledge and covering the active and passive side of three interaction modalities.

Speech understanding for air traffic control (ATC) utterances analyzes the *active auditory* side when human aviation operators are *speaking*, while text-to-speech for ATC utterances supports the *passive* side of a *listening* operator. Sophisticated ATC command gestures are used on the *active tactile* side when operators are *touching* interactive elements, while tactile cues for notification purposes in aviation support the *passive* side of *feeling* touch. Eye tracking input is utilized on the *active visual* side when operators are *looking*, while visual cues for advanced air traffic management applications serve for *highlighting* on the *passive* side. Reasonable combinations of those enhanced interaction technologies have been integrated into a common conceptualized multimodal aviation working position.

As mentioned in Chapter 3, each implemented prototype has undergone human-in-the-loop real-time simulation studies, i.e., controllers or pilots actively using the systems. These validation trials were used to record objective and subjective data based on measurements, questionnaires, and semi-structured interviews.

The measures of interest heavily depended on the tested use case. The measurement of human operator workload and situation awareness was quite common in all trials. But also, usability, acceptance, trust, satisfaction, and human errors have been evaluated. Those aforementioned human factors have been evaluated with subjective measures, and wherever possible with more objective means. If applicable, flight times, flight distances, and number of movements in a specific area have been recorded in order to objectively calculate potential capacity benefits. Furthermore, the accuracy of the prototype functionality itself has been analyzed in the form of correct and erroneous ATC concept predictions and extractions.

In the course of the outlined research activities on human machine interaction in aviation, more than a dozen of prototypes using some or all of the three modalities *auditory/tactile/visual* have been developed and tested.

More than 130 controllers from more than a dozen different, mainly European countries, more than 70 pilots, mainly US American general aviation pilots and a few European commercial pilots, as well as other ATC experts participated in altogether 16 human-in-the-loop validation campaigns. Those simulation studies on different technology readiness levels often included iterative pre-trials to ensure suitability of final trials to validate the research hypotheses. The majority of those studies took place in DLR Braunschweig's Air Traffic Validation Center while a minority was held at the research facilities of European project partners, at the FAA in the US, or partly online.

In the field of automatic speech recognition and understanding (ASRU), ATC radio telephony commands have been predicted and actual voice utterances were automatically recognized on word-level as well as on semantic level with the extraction of ATC concepts based on an ontology. Those functionalities were driven by machine learning techniques in order to reduce the manual effort for building rules. The ASRU command extraction, still relying on a well-performing rule-based implementation, has shown to deliver robust results on operational and simulation data from all phases of flights as covered by controllers and pilots in en-route, approach, tower, and apron phase. Text-to-speech technology has successfully been explored for synthesizing ATC utterances with realistic audio output.

Eye tracking was used to derive the visual attention of human operators in different air traffic management environments. It enabled to guide the operator attention to relevant displayed ATC events, to more accurately predict upcoming ATC commands, and to verify if the ASRU output has been visually scanned before it enters the electronic ATC system. The combination of ASRU, touch gestures, and eye tracking has been integrated in a multimodal controller working position to enter ATC commands. Within the latter prototype, the aircraft callsign was selected via gaze, the command type via two-dimensional multitouch gestures, and the command value via ASRU.

4.2 Conclusions of Human Machine Interaction Prototypes Validation Results

The comparison of results with prototypic implementations against baseline systems showed several benefits for human aviation operators in the tested environments such as a statistically significant reduction of human aviation operators' mental workload and an increase in situation awareness. The interaction technologies have shown to offer intuitive human machine interaction and achieved good acceptance and usability ratings. The multimodal interaction prototype revealed a potential speed gain for ATC command input of 15%.

Different prototypes enabled automated digitized air traffic control system input while reducing the safety-relevant number of errors by a factor of two at the same time. Many visual cues including weather visualization or time-based information and especially the non-intrusive attention guidance have shown that lightweight implementations can already be enough to effectively support controllers or pilots. However, the validation studies on a broad spectrum of interaction technology prototypes mainly considered TRL2 to TRL4 and are based on laboratory environments. Only the ASRU technology already investigated operational data and has proven to fulfill preindustrial TRL6.

Regarding ASRU prototypes, the often asked question remains: Are the achieved command recognition rates beyond 90% enough or are 100% needed? ASRU in ATC is not replacing any existing method, e.g., the manual data input of human aviation operators; it is only complementing it.

Thus, the human operators can be relieved of workload in 9 of 10 cases. Only one case remains to require very similar manual work than without ASRU. Of course, a low command error rate as achieved with rates below 5% in the majority of ATC environments and ASRU prototypes is important in order to avoid intense manual error correction. Hence, a command recognition rate of 90% already enables benefits for human operators even if there is nothing against pushing the rates towards 100%. Most of the ASRU prototypes have shown a reduction of mental workload of human aviation operators based on objective secondary tasks. This mainly results from reduced time for mouse clicking due to automatic radar label and flight strip maintenance, e.g., for an ATC approach scenario the clicking time was reduced by a factor of three. The electronic ASRU support even came along with less wrong or missing inputs from controllers into aircraft radar labels, i.e., a reduction by a factor of roughly two compared to only manual input.

For ASRU prototypes there is no need for the controllers and pilots to adapt their behavior, i.e., the active communication speech can remain the same as usual. This is even more the case as not only ICAO standard phraseology is supported by ASRU prototypes, but a huge number of common phraseology deviations. However, if controllers see and feel the support through ASRU and the following electronic support systems, they automatically adapted their speech, i.e., they spoke clearer and with less phraseology deviations to get even better automatic support. Thus, the pure presence of an ASRU system in ATC could already increase safety if human operators in aviation stick closer to the safety relevant ICAO communication rules.

Coming back to the research question in Chapter 1, human machine interaction technologies such as ASRU, gesture recognition, and eye tracking on the active side, text-to-speech, tactile and visual cues on the passive side as well as their multimodal combination have shown to support digitization of ATC data, to improve usability, to reduce mental workload of human aviation operators, to enhance situation awareness, and to even improve ATM efficiency and safety.

4.3 Outlook on Further Research Directions and Industrialization of Interaction Concepts and Applications

All investigated interaction technologies should be brought to higher technology readiness levels (TRL) to enable field trials or at least human-in-the-loop simulations in environments close to operations. This means to mature the multimodal controller working position prototype *TriControl*, multitouch inputs, and eye tracking based applications such as attention guidance and visual cues beyond TRL4. The ASRU technology will be brought beyond TRL6 with a technology transfer as a commercial company recently took over the know-how via license. However, the safety certification process will also need to be tackled. So, the target is to have ASRU operational for ATC before the end of this decade. Further interaction technologies can become operational at a later stage. The work and results in this thesis build the basis for further technology maturation, which would otherwise not have been possible in a streamlined way.

Additionally, more applications and an expansion to further aviation environments building on the generated results should be fostered. For the acoustic modality, this may include ASRU for digitizing ATS communication inside aircraft cockpits or between human aviation operators using contextual knowledge as well as analysis of aviation incidents and accidents.

An integration of advanced text-to-speech technologies can help to build a realistic simulated ATC environment for training of controllers and pilots. Given the semantic extraction of ATC concepts from radio telephony utterances, there can be an analysis of the human operator speech conformity with the prescribed rules. This can be done in an anonymized way for a group of human operators' speech or if it is ethically justifiable, in training phases to improve trainees' speech conformity. The sonification of air traffic control data[1] as another auditory interaction prototype has as well the potential to improve controllers' situation awareness.

The initial exploration of three-dimensional gestures, e.g., via leap control, could be intensified in order to evaluate potential benefits and drawbacks when using them in air traffic management contexts. Eye tracking data could be used to derive complex visual scan patterns of human aviation operators in their individual environments to enable an even more operator-centered guidance of attention to relevant events and display spots. With this, the visual cues and the information conveyed with them can be optimized and adapted for the relevant use case.

Multimodal human machine interaction could benefit from individualizing the choice of interaction modes for the user due to personal preferences and to overcome any disabilities. Further information could be integrated into even broader multimodal interaction such as lip reading to improve ASRU performance or stress detection of human operators to offer customized interaction functionalities. There is also further potential to suggest ATC commands or to automatically complement command parts as well as to perform plausibility checks of generated commands.

Different levels of automation as materialized with the interaction prototypes describe the path from manual control via decision support to full automation [Endsley and Kaber, 1999]. Using multimodal interaction technologies in a smart way can help to further walk along the path towards higher levels of automation in different aviation domains such as air traffic management [SESAR3JU, 2020]. The presented technologies also support the operational need considering future regulations that require more alerting functions to be implemented, which comes from the likely need for more entered data into digital ATC systems than today, e.g., due to Commission Implementing Regulation (EU) 2021/116 [European Commission, 2021]. Thus, multimodal interaction technologies have shown their potential to support human aviation operators especially at controller working positions and will be subject to further research and development considering this Aeronautical Informatics thesis as relevant input.

[1]The habilitand co-holds the German patent No. 10 2019 113 680 on a *Melody Message*

Bibliography (List of 204 References, 69 including Habilitand Contribution)

[Ahlstrom, 2015] Ahlstrom, U. (2015). Experimental evaluation of the AIRWOLF weather advisory tool for en route air traffic controllers. *Aviation Psychology and Applied Human Factors*, 5(1):18–35. https://doi.org/10.1027/2192-0923/a000070.

[Ahlstrom et al., 2015] Ahlstrom, U., Caddigan, E., Schulz, K., Ohneiser, O., Bastholm, R., and Dworsky, M. (2015). The Effect of Weather State-change Notifications on General Aviation Pilots' Behavior, Cognitive Engagement, and Weather Situation Awareness. DOT/FAA/TC-15/64. Technical report, Federal Aviation Administration, Atlantic City, NJ, USA. https://s.dlr.de/KoZ7y.

[Ahlstrom et al., 2016] Ahlstrom, U., Ohneiser, O., and Caddigan, E. (2016). Portable Weather Applications for General Aviation Pilots. *Human Factors*, 58(6):864–885. https://doi.org/10.1177/0018720816641783.

[Ahrenhold et al., 2023a] Ahrenhold, N., Helmke, H., Mühlhausen, T., Kleinert, M., Ohneiser, O., and Ehr, H. (2023a). Influence of Automatic Speech Recognition and Understanding on Flight Efficiency and Throughput - A Human-in-the-Loop Study. In *IEEE/AIAA 42nd Digital Avionics Systems Conference*, DASC 2023, Barcelona, Spain, 01-05 Oct, 2023. https://doi.org/10.1109/DASC58513.2023.10311293.

[Ahrenhold et al., 2023b] Ahrenhold, N., Helmke, H., Mühlhausen, T., Ohneiser, O., Kleinert, M., Ehr, H., Klamert, L., and Zuluaga-Gómez, J. P. (2023b). Validating Automatic Speech Recognition and Understanding for Pre-Filling Radar Labels - Increasing Safety While Reducing Air Traffic Controllers' Workload. *Aerospace*, 10(6). https://doi.org/10.3390/aerospace10060538.

[Alonso et al., 2013] Alonso, R., Causse, M., Vachon, F., Parise, R., Dehais, F., and Terrier, P. (2013). Evaluation of Head-free Eye Tracking as an Input Device for Air Traffic Control. *Ergonomics*, 56(2):246–255. https://doi.org/10.1080/00140139.2012.744473.

[Aricò et al., 2016] Aricò, P., Borghini, G., Di Flumeri, G., Colosimo, A., Bonelli, S., Golfetti, A., Pozzi, S., Imbert, J.-P., Granger, G., Benhacene, R., and Babiloni, F. (2016). Adaptive Automation Triggered by EEG-Based Mental Workload Index: A Passive Brain-Computer Interface Application in Realistic Air Traffic Control Environment. *Frontiers in Human Neuroscience*, 10. https://doi.org/10.3389/fnhum.2016.00539.

[Auinger, 2019] Auinger, T. (2019). Design and implementation of a simulated aircraft for air traffic control communication training. Master's thesis, Paris Lodron UniversitÃ¤t Salzburg, Salzburg, Austria. https://s.dlr.de/b1Etd.

[Avsar et al., 2016] Avsar, H., Fischer, J. E., and Rodden, T. (2016). Designing Touch Screen User Interfaces for Future Flight Deck Operations. In *IEEE/AIAA 35th Digital Avionics Systems Confer-*

ence, DASC 2016, Sacramento, CA, USA, 25-29 Sep, 2016.
https://doi.org/10.1109/DASC.2016.7777976.

[Badrinath and Balakrishnan, 2022] Badrinath, S. and Balakrishnan, H. (2022). Automatic Speech Recognition for Air Traffic Control Communications. *Transportation Research Record*, 2676(1):798–810.
https://doi.org/10.1177/03611981211036359.

[Barzantny, 2018] Barzantny, C. (2018). Training operational monitoring in future ATCOs using eye tracking: extended abstract. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ETRA 2018, Warsaw, Poland, 14-17 Jun, 2018, New York, NY, USA. Association for Computing Machinery.
https://doi.org/10.1145/3204493.3207412.

[Berberian et al., 2017] Berberian, B., Ohneiser, O., De Crescenzio, F., Babiloni, F., Di Flumeri, G., and Hasselberg, A. (2017). MINIMA Project: Detecting and Mitigating the Negative Impact of Automation. In Harris, D., editor, *Engineering Psychology and Cognitive Ergonomics: Performance, Emotion and Situation Awareness, 14th International Conference, EPCE 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part 1*, pages 87–105. Springer International Publishing, Cham, Switzerland.
https://doi.org/10.1007/978-3-319-58472-0_8.

[Bernhardt et al., 2019] Bernhardt, K. A., Poltavski, D., Petros, T., Ferraro, F. R., Jorgenson, T., Carlson, C., Drechsel, P., and Iseminger, C. (2019). The effects of dynamic workload and experience on commercially available EEG cognitive state metrics in a high-fidelity air traffic control environment. *Applied Ergonomics*, 77:83–91.
https://doi.org/10.1016/j.apergo.2019.01.008.

[Bhattacharjee et al., 2024] Bhattacharjee, M., Motlíček, P., Madikeri, S., Helmke, H., Ohneiser, O., Kleinert, M., and Ehr, H. (2024). Minimum effort adaptation of automatic speech recognition system in air traffic management. *European Journal of Transport and Infrastructure Research (EJTIR)*, 24(4):133–153.
https://doi.org/10.59490/ejtir.2024.24.4.7531.

[Bhattacharjee et al., 2023] Bhattacharjee, M., Motlíček, P., Nigmatulina, I., Helmke, H., Ohneiser, O., Kleinert, M., and Ehr, H. (2023). Customization of Automatic Speech Recognition Engines for Rare Word Detection Without Costly Model Re-Training. In *13th SESAR Innovation Days*, SID 2023, Seville, Spain, 27-30 Nov, 2023.
https://s.dlr.de/kSSsy.

[Brooke, 1996] Brooke, J. (1996). SUS - A 'Quick and Dirty' Usability Scale. In Jordan, P. W., Thomas, B., McClelland, I. L., and Weerdmeester, B. A., editors, *Usability Evaluation in Industry (1st ed.)*, pages 189–194. Taylor and Francis, London, United Kingdom.
https://doi.org/10.1201/9781498710411.

[Brooker, 2011a] Brooker, P. (2011a). Air Traffic Control Separation Minima: Part 1 - The Current Stasis. *Journal of Navigation*, 64(3):449–465.
https://doi.org/10.1017/S0373463311000129.

[Brooker, 2011b] Brooker, P. (2011b). Air Traffic Control Separation Minima: Part 2 - Transition to a Trajectory-based System. *Journal of Navigation*, 64(4):673–693.
https://doi.org/10.1017/S0373463311000221.

[Cardosi, 1993] Cardosi, K. M. (1993). An Analysis of En Route Controller-Pilot Voice Communications, DOT/FAA/RD-93-11. Technical report, Federal Aviation Administration, Washington D.C., USA.
https://s.dlr.de/scJB0.

[Cardosi, 1994] Cardosi, K. M. (1994). An Analysis of Tower (Local) Controller-Pilot Voice Communications, DOT/FAA/RD-94/15. Technical report, Federal Aviation Administration, Washington D.C., USA.
https://s.dlr.de/VHwbf.

[Cardosi et al., 1996] Cardosi, K. M., Brett, B., and Han, S. (1996). An Analysis of TRACON (Terminal Radar Approach Control) Controller-Pilot Voice Communications, DOT/FAA/AR-96/66. Technical report, Federal Aviation Administration, Washington D.C., USA.
https://s.dlr.de/RHlKR.

[Causse et al., 2014] Causse, M., Alonso, R., Vachon, F., Parise, R., Orliaguet, J.-P., Tremblay, S., and Terrier, P. (2014). Testing Usability and Trainability of Indirect Touch Interaction: Perspective for the Next Generation of Air Traffic Control Systems. *Ergonomics*, 57(11):1616–1627.
https://doi.org/10.1080/00140139.2014.940400.

[Chen et al., 2023] Chen, S., Helmke, H., Tarakan, R. M., Ohneiser, O., Kopald, H., and Kleinert, M. (2023). Effects of Language Ontology on Transatlantic Automatic Speech Understanding Research Collaboration in the Air Traffic Management Domain. *Aerospace*, 10(6).
https://doi.org/10.3390/aerospace10060526.

[Chen et al., 2022] Chen, S., Kopald, H. D., Avjian, B., and Fronzak, M. (2022). Automatic Pilot Report Extraction from Radio Communications. In *IEEE/AIAA 41st Digital Avionics Systems Conference*, DASC 2022, Portsmouth, VA, USA, 18-22 Sep, 2022.
https://doi.org/10.1109/DASC55683.2022.9925803.

[Chen et al., 2017] Chen, S., Kopald, H. D., Chong, R. S., Wei, Y.-J., and Levonian, Z. (2017). Read back error detection using automatic speech recognition. In *12th USA/Europe Air Traffic Management Research and Development Seminar*, ATM 2017, Seattle, WA, USA, 27-30 Jun, 2017.
https://s.dlr.de/V3St8.

[Chen et al., 2015] Chen, S., Kopald, H. D., Elessawy, A., Levonian, Z., and Tarakan, R. M. (2015). Speech Inputs to Surface Safety Logic Systems. In *IEEE/AIAA 34th Digital Avionics Systems Conference*, DASC 2015, pages 3B2–1–3B2–11, Prague, Czechia, 13-17 Sep, 2015.
https://doi.org/10.1109/DASC.2015.7311392.

[Ciupka, 2012] Ciupka, S. (2012). Siris big sister captures DFS, original German title: Siris große Schwester erobert die DFS. *transmission*, 1:14–15.
https://s.dlr.de/XBrvI.

[Connolly, 1977] Connolly, D. W. (1977). Voice Data Entry in Air Traffic Control, N93-72621. Technical report, National Aviation Facilities Experimental Center, Atlantic City, NJ, USA.
https://s.dlr.de/FA5kx.

[Conversy et al., 2011] Conversy, S., Gaspard-Boulinc, H., Chatty, S., Valès, S., Dupré, C., and Ollagnon, C. (2011). Supporting air traffic control collaboration with a TableTop system. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW 2011, pages 425–434, Hangzhou, China, 19-23 Mar, 2011, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/1958824.1958891.

[Cordero et al., 2012] Cordero, J. M., Dorado, M., and de Pablo, J. M. (2012). Automated Speech Recognition in ATC Environment. In *Proceedings of the 2nd International Conference on Application and Theory of Automation in Command and Control Systems*, ATACCS 2012, pages 46–53, London, United Kingdom, 29-31 May, 2012. IRIT Press, Toulouse, France. https://s.dlr.de/vWjMz.

[Cordero et al., 2013] Cordero, J. M., Rodríguez, N., de Pablo, J. M., and Dorado, M. (2013). Automated speech recognition in controller communications applied to workload measurement. In *3rd SESAR Innovation Days*, SID 2013, Stockholm, Sweden, 26-28 Nov, 2013. https://s.dlr.de/noHk6.

[Cox et al., 2005] Cox, G., Sharples, S. C., Patel, H., and Stedmon, A. W. (2005). Human Factors Issues in ATC: The "Use" of an Eye Tracking Methodology. In Bust, P. and McCabe, P., editors, *Contemporary Ergonomics 2005: Proceedings of the International Conference on Contemporary Ergonomics*, CE 2005, Hatfield, United Kingdom, 05-07 Apr, 2005. 1st ed., Taylor & Francis. https://doi.org/10.1201/9781003419969.

[de Rooij et al., 2023] de Rooij, G., Van Baelen, D., Borst, C., van Paassen, M. M., and Mulder, M. (2023). Supplementing Haptic Feedback in Flight Envelope Protection Through Visual Display Indications. *Journal of Aerospace Information Systems*, 20(6):351–367. https://doi.org/10.2514/1.I011191.

[Dehais et al., 2017] Dehais, F., Behrend, J., Peysakhovich, V., Causse, M., and Wickens, C. D. (2017). Pilot Flying and Pilot Monitoring's Aircraft State Awareness During Go-Around Execution in Aviation: A Behavioral and Eye Tracking Study. *The International Journal of Aerospace Psychology*, 27(1-2):15–28. https://doi.org/10.1080/10508414.2017.1366269.

[Dehn, 2008] Dehn, D. M. (2008). Assessing the Impact of Automation on the Air Traffic Controller: The SHAPE Questionnaires. *Air Traffic Control Quarterly*, 16:127–146. https://doi.org/10.2514/atcq.16.2.127.

[Dhavala, 2014] Dhavala, L. (2014). Use of Synthetic Voice to Improve Communication between Air Traffic Controllers and Pilots. Technical report, Emirates Aviation University, Dubai, UAE. https://s.dlr.de/4MZuC.

[Di Flumeri et al., 2019] Di Flumeri, G., De Crescenzio, F., Berberian, B., Ohneiser, O., Kramer, J., Aricò, P., Borghini, G., Babiloni, F., Bagassi, S., and Piastra, S. (2019). Brain-Computer Interface-Based Adaptive Automation to Prevent Out-Of-The-Loop Phenomenon in Air Traffic Controllers Dealing With Highly Automated Systems. *Frontiers in Human Neuroscience*, 13. https://doi.org/10.3389/fnhum.2019.00296.

[Dodd et al., 2014] Dodd, S., Lancaster, J., Miranda, A., Grothe, S., DeMers, B., and Rogers, B. (2014). Touch Screens on the Flight Deck: The Impact of Touch Target Size, Spacing, Touch Technology

and Turbulence on Pilot Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1):6–10.
https://doi.org/10.1177/1541931214581002.

[Dodd et al., 2022] Dodd, S., Lubold, N., Bui, B., and Finseth, T. (2022). Touch User Interfaces in Air Traffic Control: Final Guidelines Report with Recommended Updates for the FAA Human Factors Design Standard (HFDS); FAA HF-STD-001B. Technical report, Federal Aviation Administration, Washington D.C., USA.
https://s.dlr.de/TSyAC.

[Dumas et al., 2009] Dumas, B., Lalanne, D., and Oviatt, S. (2009). Multimodal Interfaces: A Survey of Principles, Models and Frameworks. In Lalanne, D. and Kohlas, J., editors, *Human Machine Interaction: Research Results of the MMI Program*, pages 3–26. Springer Berlin Heidelberg.
https://doi.org/10.1007/978-3-642-00437-7_1.

[Eccles, 2023] Eccles, M. (2023). Europe's air traffic controllers are falling off the radar. politico.eu.
https://s.dlr.de/s6aC9 [Accessed: 16 Dec, 2024].

[Edinger and Schmitt, 2012] Edinger, C. and Schmitt, A. R. (2012). Rapid Prototyping for ATM Operational Concept Development. In *Deutscher Luft- und Raumfahrtkongress*, DLRK 2012, Berlin, Germany, 10-12 Sep, 2012.
https://s.dlr.de/BzC5D.

[Elmquist et al., 2023] Elmquist, E., Bock, A., Lundberg, J., Ynnerman, A., and Rönnberg, N. (2023). SonAir: the design of a sonification of radar data for air traffic control. *Journal on Multimodal User Interfaces*, 17(3):137–149.
https://doi.org/10.1007/s12193-023-00404-x.

[Endsley, 1988] Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, volume 3, pages 789–795.
https://doi.org/10.1109/NAECON.1988.195097.

[Endsley, 2021] Endsley, M. R. (2021). A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM. *Human Factors*, 63(1):124–150.
https://doi.org/10.1177/0018720819875376.

[Endsley and Kaber, 1999] Endsley, M. R. and Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, 42(3):462–492.
https://doi.org/10.1080/001401399185595.

[European Commission, 2021] European Commission (2021). L 36/10; Commission Implementing Regulation (EU) 2021/116 of 1 February 2021 on the Establishment of the Common Project One Supporting the Implementation of the European Air Traffic Management Master Plan Provided for in Regulation (EC) No 550/2004 of the European Parliament and of the Council, Amending Commission Implementing Regulation (EU) No 409/2013 and Repealing Commission Implementing Regulation (EU) No 716/2014. *Official Journal of the European Union*, 64:10–38.
https://s.dlr.de/CQ6Y5.

[Faulhaber and Friedrich, 2019] Faulhaber, A. K. and Friedrich, M. (2019). Eye-Tracking Metrics as an Indicator of Workload in Commercial Single-Pilot Operations. In Longo, L. and Leva, M. C., editors, *Human Mental Workload: Models and Applications*, pages 213–225. Springer International Publishing, Cham, Switzerland.
https://doi.org/10.1007/978-3-030-32423-0_14.

[Förster et al., 2011] Förster, F., Ohneiser, O., and Temme, M.-M. (2011). Manual Approach Path Adaptation with Controller Assistance System Support for Implementation of a time-based Approach Guidance, original German title: Manuelle Anflugroutenanpassung mit Unterstützung eines Lotsenassistenzsystems zur Implementierung einer zeitbasierten Anflugführung. In *Deutscher Luft- und Raumfahrtkongress*, DLRK 2011, pages 261–268, Bremen, Germany, 27-29 Sep, 2011.
https://s.dlr.de/Cftde.

[García et al., 2023] García, R., Albarrán, J., Fabio, A., Celorrio, F., Pinto de Oliveira, C., and Bárcena, C. (2023). Automatic Flight Callsign Identification on a Controller Working Position: Real-Time Simulation and Analysis of Operational Recordings. *Aerospace*, 10(5).
https://doi.org/10.3390/aerospace10050433.

[Gauci et al., 2018] Gauci, J., Theuma, K., Muscat, A., and Zammit-Mangion, D. (2018). Evaluation of a Multimodal Interface for Pilot Interaction with Avionic Systems. In *IEEE/AIAA 37th Digital Avionics Systems Conference*, DASC 2018, London, United Kingdom, 23-27 Sep, 2018.
https://doi.org/10.1109/DASC.2018.8569607.

[Gauci et al., 2017] Gauci, J., Xuereb, M., Muscat, A., and Zammit-Mangion, D. (2017). Multi-modal Interaction Between Pilots and Avionic Systems On-Board Large Commercial Aircraft. In Harris, D., editor, *Engineering Psychology and Cognitive Ergonomics: Cognition and Design, 14th International Conference, EPCE 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part 1*, pages 200–210. Springer International Publishing, Cham, Switzerland.
https://doi.org/10.1007/978-3-319-58475-1_15.

[Guo et al., 2021] Guo, D., Zhang, Z., Fan, P., Zhang, J., and Yang, B. (2021). A Context-Aware Language Model to Improve the Speech Recognition in Air Traffic Control. *Aerospace*, 8(11).
https://doi.org/10.3390/aerospace8110348.

[Gürlük et al., 2015] Gürlük, H., Helmke, H., Wies, M., Ehr, H., Kleinert, M., Mühlhausen, T., Muth, K., and Ohneiser, O. (2015). Assistant Based Speech Recognition - Another Pair of Eyes for the Arrival Manager. In *IEEE/AIAA 34th Digital Avionics Systems Conference*, DASC 2015, pages 3B6–1–3B6–14, Prague, Czechia, 13-17 Sep, 2015.
https://doi.org/10.1109/DASC.2015.7311396.

[Hagemann and Udovic, 2019] Hagemann, K. and Udovic, A. (2019). Tactical inputs for a stripless ATC system via Multi-Touch Gestures. *Innovation im Fokus*, 1:21–27.
https://s.dlr.de/No2vp.

[Hart, 2006] Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908.
https://doi.org/10.1177/154193120605000909.

[Hasse et al., 2012] Hasse, C., Grasshoff, D., and Bruder, C. (2012). How to measure monitoring performance of pilots and air traffic controllers. In *Proceedings of the Symposium on Eye Tracking*

Research and Applications, ETRA 2012, pages 409–412, Santa Barbara, CA, USA, 28-30 Mar, 2012, New York, NY, USA. Association for Computing Machinery.
https://doi.org/10.1145/2168556.2168649.

[Helmke et al., 2013] Helmke, H., Ehr, H., Kleinert, M., Faubel, F., and Klakow, D. (2013). Increased Acceptance of Controller Assistance by Automatic Speech Recognition. In *10th USA/Europe Air Traffic Management Research and Development Seminar*, ATM 2013, Chicago, IL, USA, 10-13 Jun, 2013.
https://s.dlr.de/CzcsV.

[Helmke et al., 2023a] Helmke, H., Kleinert, M., Ahrenhold, N., Ehr, H., Mühlhausen, T., Ohneiser, O., Klamert, L., Motlíček, P., Prasad, A., Zuluaga-Gómez, J. P., Dokic, J., and Pinska-Chauvin, E. (2023a). Automatic Speech Recognition and Understanding for Radar Label Maintenance Support Increases Safety and Reduces Air Traffic Controllers' Workload. In *15th USA/Europe Air Traffic Management Research and Development Seminar*, ATM 2023, Savannah, GA, USA, 05-09 Jun, 2023.
https://s.dlr.de/10OZw.

[Helmke et al., 2024a] Helmke, H., Kleinert, M., Ohneiser, O., Ahrenhold, N., Klamert, L., and Motlíček, P. (2024a). Safety and Workload Benefits of Automatic Speech Understanding for Radar Label Updates. *Journal of Air Transportation*, 32(4):155–168.
https://doi.org/10.2514/1.D0419.

[Helmke et al., 2020] Helmke, H., Kleinert, M., Ohneiser, O., Ehr, H., and Shetty, S. (2020). Machine Learning of Air Traffic Controller Command Extraction Models for Speech Recognition Applications. In *IEEE/AIAA 39th Digital Avionics Systems Conference*, DASC 2020, Virtual, 11-16 Oct, 2020.
https://doi.org/10.1109/DASC50938.2020.9256484.

[Helmke et al., 2021a] Helmke, H., Kleinert, M., Shetty, S., Ohneiser, O., Ehr, H., Arilíusson, H., Simiganoschi, T. S., Prasad, A., Motlíček, P., Veselý, K., Ondřej, K., Smrz, P., Harfmann, J., and Windisch, C. (2021a). Readback Error Detection by Automatic Speech Recognition to Increase ATM Safety. In *14th USA/Europe Air Traffic Management Research and Development Seminar*, ATM 2021, Virtual, 20-24 Sep, 2021.
https://s.dlr.de/LtMbB.

[Helmke et al., 2017] Helmke, H., Ohneiser, O., Buxbaum, J., and Kern, C. (2017). Increasing ATM Efficiency with Assistant Based Speech Recognition. In *12th USA/Europe Air Traffic Management Research and Development Seminar*, ATM 2017, Seattle, WA, USA, 27-30 Jun, 2017.
https://s.dlr.de/Vm1Eg.

[Helmke et al., 2023b] Helmke, H., Ohneiser, O., Kleinert, M., Chen, S., Kopald, H., and Tarakan, R. M. (2023b). Transatlantic Approaches for Automatic Speech Understanding in Air Traffic Management. In *15th USA/Europe Air Traffic Management Research and Development Seminar*, ATM 2023, Savannah, GA, USA, 05-09 Jun, 2023.
https://s.dlr.de/AOUnj.

[Helmke et al., 2024b] Helmke, H., Ohneiser, O., and many other Special Issue contributors (2024b). *Automatic Speech Recognition and Understanding in Air Traffic Management*. Helmke, H. and Ohneiser, O., editors, MDPI, Basel, Switzerland, ISBN 978-3-7258-0316-3.
https://doi.org/10.3390/books978-3-7258-0315-6.

[Helmke et al., 2016a] Helmke, H., Ohneiser, O., Mühlhausen, T., and Wies, M. (2016a). Reducing Controller Workload with Automatic Speech Recognition. In *IEEE/AIAA 35th Digital Avionics Systems Conference*, DASC 2016, Sacramento, CA, USA, 25-29 Sep, 2016.
https://doi.org/10.1109/DASC.2016.7778024.

[Helmke et al., 2016b] Helmke, H., Ohneiser, O., Wies, M., and Kleinert, M. (2016b). AcListant®-Strips - Validation Results of Main Trials. Technical Report DLR-IB-FL-BS-2016-19, German Aerospace Center (DLR), Institute of Flight Guidance, Braunschweig, Germany.
https://s.dlr.de/GcL0r.

[Helmke et al., 2022] Helmke, H., Ondřej, K., Shetty, S., Arilíusson, H., Simiganoschi, T. S., Kleinert, M., Ohneiser, O., Ehr, H., and Zuluaga-Gómez, J. P. (2022). Readback Error Detection by Automatic Speech Recognition and Understanding - Results of HAAWAII project for Isavia's Enroute Airspace. In *12th SESAR Innovation Days*, SID 2022, Budapest, Hungary, 05-08 Dec, 2022.
https://s.dlr.de/9NrZG.

[Helmke et al., 2015] Helmke, H., Rataj, J., Mühlhausen, T., Ohneiser, O., Ehr, H., Kleinert, M., Oualil, Y., and Schulder, M. (2015). Assistant-Based Speech Recognition for ATM Applications. In *11th USA/Europe Air Traffic Management Research and Development Seminar*, ATM 2015, Lisbon, Portugal, 23-26 Jun, 2015.
https://s.dlr.de/LS7sS.

[Helmke et al., 2021b] Helmke, H., Shetty, S., Kleinert, M., Ohneiser, O., Prasad, A., Motlíček, P., Černá, A., and Windisch, C. (2021b). How to Measure Speech Recognition Performance in the Air Traffic Control Domain? The Word Error Rate is only half of the truth. In *Interspeech 2021 Satellite Workshop 'Automatic Speech Recognition in Air Traffic Management'*, ASR-ATM 2021, Brno, Czechia (hybrid), 30 Aug, 2021.
https://s.dlr.de/3Ydz6.

[Helmke et al., 2021c] Helmke, H., Shetty, S., Kleinert, M., Ohneiser, O., Prasad, A., Motlíček, P., Černá, A., and Windisch, C. (2021c). Measuring Speech Recognition And Understanding Performance in Air Traffic Control Domain Beyond Word Error Rates. In *11th SESAR Innovation Days*, SID 2021, Virtual, 07-09 Dec, 2021.
https://s.dlr.de/Rt8dv.

[Helmke et al., 2018] Helmke, H., Slotty, M., Poiger, M., Herrer, D. F., Ohneiser, O., Vink, N., Černá, A., Hartikainen, P., Josefsson, B., Langr, D., García Lasheras, R., Marin, G., Mevatne, O. G., Moos, S., Nilsson, M. N., and Pérez, M. B. (2018). Ontology for Transcription of ATC Speech Commands of SESAR 2020 Solution PJ.16-04. In *IEEE/AIAA 37th Digital Avionics Systems Conference*, DASC 2018, London, United Kingdom, 23-27 Sep, 2018.
https://doi.org/10.1109/DASC.2018.8569238.

[Hofbauer et al., 2008] Hofbauer, K., Petrik, S., and Hering, H. (2008). The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC 2008, Marrakech, Morocco, 28-30 May, 2008. European Language Resources Association (ELRA).
https://s.dlr.de/mKsCN.

[Hunger et al., 2024] Hunger, R., Christoffels, L., Friedrich, M., Jameel, M., Pick, A., Gerdes, I., von der Nahmer, P. M., and Sobotzki, F. (2024). Lesson Learned: Design and Perception of Single Controller Operations Support Tools. In Harris, D. and Li, W.-C., editors, *Engineering Psychology and Cognitive Ergonomics: Cognition and Design, 21st International Conference, EPCE 2024, Held as Part of HCI International 2024, Washington D.C., USA, 29 Jun-04 Jul, 2024*, volume 14693, pages 15–33. Springer Nature, Cham, Switzerland.
https://doi.org/10.1007/978-3-031-60731-8_2.

[ICAO, 2016] ICAO (2016). Procedures for Air Navigation Services - Air Traffic Management, Doc 4444, Sixteenth Edition. Technical report, International Civil Aviation Organization, Montréal, QC, Canada.
https://s.dlr.de/daW7j.

[Jameel et al., 2023] Jameel, M., Tyburzy, L., Gerdes, I., Pick, A., Hunger, R., and Christoffels, L. (2023). Enabling Digital Air Traffic Controller Assistant through Human-Autonomy Teaming Design. In *42nd IEEE/AIAA Digital Avionics Systems Conference*, DASC 2023, Barcelona, Spain, 01-05 Oct, 2023. IEEE.
https://doi.org/10.1109/DASC58513.2023.10311220.

[Jauer, 2014] Jauer, M.-L. (2014). Multimodal Controller Working Position, Integration of Automatic Speech Recognition and Multi-Touch Technology, original German title: Multimodaler Fluglotsenarbeitsplatz, Integration von Automatischer Spracherkennung und Multi-Touch Technologie; DLR-IB 112-2014/39. Technical report, Duale Hochschule Baden-Württemberg Mannheim, Germany, Bachelor Thesis.
https://s.dlr.de/PsFsX.

[Jin et al., 2023] Jin, H., Gao, W., Li, K., and Chu, M. (2023). Air Traffic Control Forgetting Prediction based on Eye Movement Information and Hybrid Neural Network. *Scientific Reports*, 13(13084).
https://doi.org/10.1038/s41598-023-40406-z.

[Jordan and Brennen, 1992] Jordan, C. S. and Brennen, S. D. (1992). Instantaneous self-assessment of workload technique (ISA). Technical report, Defence Research Agency, Portsmouth, United Kingdom.
https://s.dlr.de/kx2W3.

[Joshi et al., 2015] Joshi, A., Kale, S., Chandel, S., and Pal, D. K. (2015). Likert Scale: Explored and Explained. *Current Journal of Applied Science and Technology*, 7(4):396–403.
https://doi.org/10.9734/BJAST/2015/14975.

[Kaber et al., 2006] Kaber, D. B., Perry, C. M., Segall, N., McClernon, C. K., and Prinzel III, L. J. (2006). Situation awareness implications of adaptive automation for information processing in an air traffic control-related task. *International Journal of Industrial Ergonomics*, 36(5):447–462.
https://doi.org/10.1016/j.ergon.2006.01.008.

[Kaber and Riley, 1999] Kaber, D. B. and Riley, J. M. (1999). Adaptive Automation of a Dynamic Control Task Based on Secondary Task Workload Measurement. *International Journal of Cognitive Ergonomics*, 3(3):169–187.
https://doi.org/10.1207/s15327566ijce0303_1.

[Kang et al., 2017] Kang, Z., Mandal, S., and Dyer, J. (2017). Data Visualization Approaches in Eye Tracking to Support the Learning of Air Traffic Control Operations. In *Proceedings of the National Training Aircraft Symposium*, NTAS 2017, Daytona Beach, FL, USA, 14 Aug, 2017.
https://s.dlr.de/eh9yH.

[Kasttet et al., 2024] Kasttet, M. S., Lyhyaoui, A., Zbakh, D., Aramja, A., and Kachkari, A. (2024). Toward Effective Aircraft Call Sign Detection Using Fuzzy String-Matching between ASR and ADS-B Data. *Aerospace*, 11(1).
https://doi.org/10.3390/aerospace11010032.

[Khalil et al., 2023] Khalil, D., Prasad, A., Motlíček, P., Zuluaga-Gómez, J. P., Nigmatulina, I., Madikeri, S., and Schuepbach, C. (2023). An Automatic Speaker Clustering Pipeline for the Air Traffic Communication Domain. *Aerospace*, 10(10).
https://doi.org/10.3390/aerospace10100876.

[Kleinert et al., 2018] Kleinert, M., Helmke, H., Ehr, H., Kern, C., Klakow, D., Motlíček, P., Singh, M., and Siol, G. (2018). Building Blocks of Assistant Based Speech Recognition for Air Traffic Management Applications. In *8th SESAR Innovation Days*, SID 2018, Salzburg, Austria, 03-07 Dec, 2018.
https://s.dlr.de/VCbsx.

[Kleinert et al., 2019] Kleinert, M., Helmke, H., Moos, S., Hlousek, P., Windisch, C., Ohneiser, O., Ehr, H., and Labreuil, A. (2019). Reducing Controller Workload by Automatic Speech Recognition Assisted Radar Label Maintenance. In *9th SESAR Innovation Days*, SID 2019, Athens, Greece, 02-05 Dec, 2019.
https://s.dlr.de/P3TMA.

[Kleinert et al., 2021a] Kleinert, M., Helmke, H., Shetty, S., Ohneiser, O., Ehr, H., Prasad, A., Motlíček, P., and Harfmann, J. (2021a). Automated Interpretation of Air Traffic Control Communication: The Journey from Spoken Words to a Deeper Understanding of the Meaning. In *IEEE/AIAA 40th Digital Avionics Systems Conference*, DASC 2021, San Antonio, TX, USA, 03-07 Oct, 2021.
https://doi.org/10.1109/DASC52595.2021.9594387.

[Kleinert et al., 2023] Kleinert, M., Ohneiser, O., Helmke, H., Shetty, S., Ehr, H., Maier, M., Schacht, S., and Wiese, H. (2023). Safety Aspects of Supporting Apron Controllers with Automatic Speech Recognition and Understanding Integrated into an Advanced Surface Movement Guidance and Control System. *Aerospace*, 10(7).
https://doi.org/10.3390/aerospace10070596.

[Kleinert et al., 2022] Kleinert, M., Shetty, S., Helmke, H., Ohneiser, O., Wiese, H., Maier, M., Schacht, S., Nigmatulina, I., Sarfjoo, S. S., and Motlíček, P. (2022). Apron Controller Support by Integration of Automatic Speech Recognition with an Advanced Surface Movement Guidance and Control System. In *12th SESAR Innovation Days*, SID 2022, Budapest, Hungary, 05-08 Dec, 2022.
https://s.dlr.de/EPpjr.

[Kleinert et al., 2021b] Kleinert, M., Venkatarathinam, N., Helmke, H., Ohneiser, O., Strake, M., and Fingscheidt, T. (2021b). Easy Adaptation of Speech Recognition to Different Air Traffic Control Environments using the DeepSpeech Engine. In *11th SESAR Innovation Days*, SID 2021, Virtual, 07-09 Dec, 2021.

https://s.dlr.de/hAG7U.

[Landry, 2011] Landry, S. J. (2011). Human centered design in the air traffic control system. *J. Intell. Manuf.*, 22(1):65–72.
https://doi.org/10.1007/s10845-009-0278-6.

[Lechner et al., 2002] Lechner, A., Mattson, P., and Ecker, K. (2002). Voice recognition: software solutions in real-time ATC workstations. *IEEE Aerospace and Electronic Systems Magazine*, 17(11):11–16.
https://doi.org/10.1109/MAES.2002.1047373.

[Lee et al., 2001] Lee, K. K., Kerns, K., Bone, R., and Nickelson, M. (2001). Development and Validation of the Controller Acceptance Rating Scale (CARS): Results of Empirical Research. In *4th USA/Europe Air Traffic Management Research and Development Seminar*, ATM 2001, Santa Fe, NM, USA, 03-07 Dec, 2001.
https://s.dlr.de/fq5xH.

[Levenshtein, 1966] Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10(8):707–710.
https://s.dlr.de/cXp8H.

[Li et al., 2023] Li, Q., Ng, K. K., Yu, S. C., Yiu, C. Y., and Lyu, M. (2023). Recognising situation awareness associated with different workloads using EEG and eye-tracking features in air traffic control tasks. *Knowledge-Based Systems*, 260:110179.
https://doi.org/10.1016/j.knosys.2022.110179.

[Li et al., 2018] Li, W.-C., Kearney, P., Braithwaite, G., and Lin, J. J. (2018). How much is too much on monitoring tasks? Visual scan patterns of single air traffic controller performing multiple remote tower operations. *International Journal of Industrial Ergonomics*, 67:135–144.
https://doi.org/10.1016/j.ergon.2018.05.005.

[Likert, 1932] Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, 22(140):5–55.
https://s.dlr.de/pP40x.

[Lim et al., 2018] Lim, Y., Gardi, A., Ezer, N., Kistan, T., and Sabatini, R. (2018). Eye-Tracking Sensors for Adaptive Aerospace Human-Machine Interfaces and Interactions. In *5th IEEE International Workshop on Metrology for AeroSpace*, MetroAeroSpace 2018, pages 311–316.
https://doi.org/10.1109/MetroAeroSpace.2018.8453509.

[Lin, 2021] Lin, Y. (2021). Spoken Instruction Understanding in Air Traffic Control: Challenge, Technique, and Application. *Aerospace*, 8(3).
https://doi.org/10.3390/aerospace8030065.

[Lin et al., 2023] Lin, Y., Ruan, M., Cai, K., Li, D., Zeng, Z., Li, F., and Yang, B. (2023). Identifying and managing risks of AI-driven operations: A case study of automatic speech recognition for improving air traffic safety. *Chinese Journal of Aeronautics*, 36(4):366–386.
https://doi.org/10.1016/j.cja.2022.08.020.

[Lin et al., 2021a] Lin, Y., Wu, Y., Guo, D., Zhang, P., Yin, C., Yang, B., and Zhang, J. (2021a). A Deep Learning Framework of Autonomous Pilot Agent for Air Traffic Controller Training. *IEEE*

Transactions on Human-Machine Systems, 51(5):442–450.
https://doi.org/10.1109/THMS.2021.3102827.

[Lin et al., 2021b] Lin, Y., Yang, B., Li, L., Guo, D., Zhang, J., Chen, H., and Zhang, Y. (2021b). ATCSpeechNet: A multilingual end-to-end speech recognition framework for air traffic control systems. *Applied Soft Computing*, 112:107847.
https://doi.org/10.1016/j.asoc.2021.107847.

[Manaker, 1982] Manaker, E. (1982). Pilot ability to understand synthetic voice and radio voice when received simultaneously, ACT-R-82-O1. Technical report, Grumman Aerospace Corporation, New York, NY, USA.
https://s.dlr.de/LABw3.

[Mankins, 1995] Mankins, J. C. (1995). Technology Readiness Levels - A White Paper. Technical report, NASA, Office of Space Access and Technology, Advanced Concepts Office.
https://s.dlr.de/4OZnA.

[Manske and Schier, 2015] Manske, P. G. and Schier, S. L. (2015). Visual Scanning in an Air Traffic Control Tower - A Simulation Study. *Procedia Manufacturing*, 3:3274–3279.
https://doi.org/10.1016/j.promfg.2015.07.397.

[Meier et al., 2024] Meier, J., Finke, M., Ohneiser, O., and Jameel, M. (2024). Flexible Air Traffic Controller Deployment with Artificial Intelligence based Decision Support: Literature Survey and Evaluation Framework. In *Deutscher Luft- und Raumfahrtkongress*, DLRK 2024, Hamburg, Germany, 30 Sep-02 Oct, 2024.
https://doi.org/10.25967/630258.

[Merchant and Schnell, 2000] Merchant, S. and Schnell, T. (2000). Applying Eye Tracking as an Alternative Approach for Activation of Controls and Functions in Aircraft. In *IEEE/AIAA 19th Digital Avionics Systems Conference*, volume 2 of *DASC 2000*, pages 5A5/1–5A5/9, Philadelphia, PA, USA, 07-13 Oct, 2000.
https://doi.org/10.1109/DASC.2000.884872.

[Mertz et al., 2000] Mertz, C. P., Chatty, S., and Vinot, J.-L. (2000). The influence of design techniques on user interfaces: the DigiStrips experiment for air traffic control. In *Proceedings of the International Conference on Human-Computer Interaction in Aerospace*, HCI-Aero 2000, Toulouse, France, 27-29 Sep, 2000.
https://s.dlr.de/HPQMr.

[Mitra and Acharya, 2007] Mitra, S. K. and Acharya, T. (2007). Gesture Recognition: A Survey. *Trans. Sys. Man Cyber Part C*, 37(3):311–324.
https://doi.org/10.1109/TSMCC.2007.893280.

[Motlíček et al., 2023] Motlíček, P., Prasad, A., Nigmatulina, I., Helmke, H., Ohneiser, O., and Kleinert, M. (2023). Automatic Speech Analysis Framework for ATC Communication in HAAWAII. In *13th SESAR Innovation Days*, SID 2023, Seville, Spain, 27-30 Nov, 2023.
https://s.dlr.de/ost2A.

[Nealley and Gawron, 2015] Nealley, M. A. and Gawron, V. J. (2015). The Effect of Fatigue on Air Traffic Controllers. *The International Journal of Aviation Psychology*, 25(1):14–47.
https://doi.org/10.1080/10508414.2015.981488.

[Neßelrath et al., 2016] Neßelrath, R., Moniri, M. M., and Feld, M. (2016). Combining Speech, Gaze, and Micro-gestures for the Multimodal Control of In-Car Functions. In *12th International Conference on Intelligent Environments*, IE 2016, pages 190–193, London, United Kingdom, 14-16 Sep, 2016. https://doi.org/10.1109/IE.2016.42.

[Ngo et al., 2012] Ngo, M. K., Pierce, R. S., and Spence, C. (2012). Using Multisensory Cues to Facilitate Air Traffic Management. *Human Factors*, 54(6):1093–1103. https://doi.org/10.1177/0018720812446623.

[Nguyen and Holone, 2016] Nguyen, V. N. and Holone, H. (2016). N-best list re-ranking using Syntactic Score: A Solution for Improving Speech Recognition Accuracy in Air Traffic Control. In *16th International Conference on Control, Automation and Systems*, ICCAS 2016, pages 1309–1314, Gyeongju, Korea, 16-19 Oct, 2016. https://doi.org/10.1109/ICCAS.2016.7832482.

[Nylin et al., 2020] Nylin, M., Lundberg, J., and Johansson, J. (2020). Attention support with soft visual cues in control room environments. In *24th International Conference Information Visualisation*, IV 2020, pages 160–165. https://doi.org/10.1109/IV51561.2020.00035.

[Ohneiser, 2012] Ohneiser, O. (2012). Command Support for the Integration of Conventionally Equipped Aircraft into a Time-based Approach Stream, original German title: Führungsunterstützung zur Integration konventionell ausgerüsteter Luftfahrzeuge in einen zeitbasierten Anflugstrom. In *54. Fachausschusssitzung Anthropotechnik der Deutschen Gesellschaft für Luft- und Raumfahrt*, DGLR-Antrhopotechnik 2012, DGLR-B (01), pages 175–192, Koblenz, Germany, 30-31 Oct, 2012. https://s.dlr.de/XptRU.

[Ohneiser, 2016a] Ohneiser, O. (2016a). Improved User Acceptance During Stepwise Air Traffic Control Display Functionality Introduction. In Stanton, N. A., Landry, S., Di Bucchianico, G., and Vallicelli, A., editors, *Advances in Human Aspects of Transportation*, volume 484, pages 873–883, Orlando, FL, USA, 27-31 Jul, 2016. Springer International Publishing, Cham, Switzerland. https://doi.org/10.1007/978-3-319-41682-3_72.

[Ohneiser, 2016b] Ohneiser, O. (2016b). Transition Steps to Orthogonal Unidirectional Air Traffic Controller Monitoring Display. In *Seventh International Conference on Research in Air Transportation*, ICRAT 2016, Philadelphia, PA, USA, 20-24 Jun, 2016. https://s.dlr.de/Xej3O.

[Ohneiser, 2017] Ohneiser, O. (2017). *Migration tolerant transitions of incremental display steps for air traffic controllers, original German title: Migrationstolerante Transitionen von inkrementellen Displaystufen für Fluglotsen; DLR-FB-2017-05*. PhD thesis, Technische Universität Carolo-Wilhelmina zu Braunschweig, Braunschweig, Germany. https://s.dlr.de/qEfTG.

[Ohneiser et al., 2021a] Ohneiser, O., Adamala, J., and Salomea, I.-T. (2021a). Integrating Eye- and Mouse-Tracking with Assistant Based Speech Recognition for Interaction at Controller Working Positions. *Aerospace, Special Issue Aeronautical Informatics*, 8(9). https://doi.org/10.3390/aerospace8090245.

[Ohneiser et al., 2018a] Ohneiser, O., Ahlstrom, V., Tracy, K., and Williams, B. (2018a). Comparison of Air Traffic Controller Display Techniques for Reaching Target Times at Significant Waypoints. In *IEEE/AIAA 37th Digital Avionics Systems Conference*, DASC 2018, pages 1092–1101, London, United Kingdom, 23-27 Sep, 2018.
https://doi.org/10.1109/DASC.2018.8569365.

[Ohneiser and Ahmed, 2025] Ohneiser, O. and Ahmed, U. (2025). Text-To-Speech Application for Training of Aviation Radio Telephony Communication Operators. *IEEE Transactions on Aerospace and Electronic Systems*, 61(2):4542–4560.
https://doi.org/10.1109/TAES.2024.3504493.

[Ohneiser and Beddig, 2013] Ohneiser, O. and Beddig, S. (2013). Air Traffic Controller Support for Conformance Monitoring and Active Control of Air Traffic, original German title: Fluglotsen-Unterstützung für die Konformitätsüberwachung und aktive Führung des Luftverkehrs. In *Deutscher Luft- und Raumfahrtkongress*, DLRK 2013, Stuttgart, Germany, 10-12 Oct, 2013.
https://doi.org/10.25967/301287.

[Ohneiser et al., 2020] Ohneiser, O., Biella, M., Schmugler, A., and Wallace, M. (2020). Operational Feasibility Analysis of the Multimodal Controller Working Position "TriControl". *Aerospace*, 7(2).
https://doi.org/10.3390/aerospace7020015.

[Ohneiser et al., 2018b] Ohneiser, O., De Crescenzio, F., Di Flumeri, G., Kraemer, J., Berberian, B., Bagassi, S., Sciaraffa, N., Aricò, P., Borghini, G., and Babiloni, F. (2018b). Experimental Simulation Set-Up for Validating Out-Of-The-Loop Mitigation when Monitoring High Levels of Automation in Air Traffic Control. In *International Conference on Air Traffic Management and Aviation*, ICATMA 2018, Lisbon, Portugal, 16-17 Apr, 2018.
https://doi.org/10.5281/zenodo.1316361.

[Ohneiser and Gürlük, 2013] Ohneiser, O. and Gürlük, H. (2013). Migration Tolerant Human Computer Interaction for Air Traffic Controllers. In *Human Interface and the Management of Information: Information and Interaction for Health, Safety, Mobility and Complex Environments (Part II)*, volume 8017 of *Human Computer Interaction (HCI) International 2013*, pages 143–152, Las Vegas, NV, USA, 21-26 Jul, 2013. Springer Berlin Heidelberg.
https://doi.org/10.1007/978-3-642-39215-3_18.

[Ohneiser et al., 2019a] Ohneiser, O., Gürlük, H., Jauer, M.-L., Szöllősi, Á., and Balló, D. (2019a). Please have a Look here: Successful Guidance of Air Traffic Controller's Attention. In *9th SESAR Innovation Days*, SID 2019, Athens, Greece, 02-05 Dec, 2019.
https://s.dlr.de/qyGZM.

[Ohneiser et al., 2014a] Ohneiser, O., Heesen, M., Flemisch, F. O., and Rataj, J. (2014a). Migration Tolerant Human Machine Interface Concepts in the domains of Air Traffic Control and Automotive. In *Deutscher Luft- und Raumfahrtkongress*, DLRK 2014, Augsburg, Germany, 16-18 Sep, 2014.
https://doi.org/10.25967/340082.

[Ohneiser et al., 2014b] Ohneiser, O., Helmke, H., Ehr, H., Gürlük, H., Hössl, M., Mühlhausen, T., Oualil, Y., Schulder, M., Schmidt, A., Khan, A., and Klakow, D. (2014b). Air Traffic Controller Support by Speech Recognition. In *Advances in Human Aspects of Transportation: Part II*, Applied Human Factors and Ergonomics (AHFE) 2014, pages 492–503, Krakow, Poland, 19-23 Jul, 2014.

CRC Press.
https://s.dlr.de/6KmDT.

[Ohneiser et al., 2019b] Ohneiser, O., Helmke, H., Kleinert, M., Siol, G., Ehr, H., Hobein, S., Predescu, A.-V., and Bauer, J. (2019b). Tower Controller Command Prediction for Future Speech Recognition Applications. In *9th SESAR Innovation Days*, SID 2019, Athens, Greece, 02-05 Dec, 2019.
https://s.dlr.de/OaEPQ.

[Ohneiser et al., 2022] Ohneiser, O., Helmke, H., Shetty, S., Kleinert, M., Ehr, H., Balogh, G., Tønnesen, A., Rinaldi, W., Mansi, S., Piazzolla, G., Murauskas, Š., Pagirys, T., Kis-Pál, G., Horváth, V., Kling, F., and Usanovic, H. (2022). Understanding Tower Controller Communication for Support in Air Traffic Control Displays. In *12th SESAR Innovation Days*, SID 2022, Budapest, Hungary, 05-08 Dec, 2022.
https://s.dlr.de/TztJt.

[Ohneiser et al., 2021b] Ohneiser, O., Helmke, H., Shetty, S., Kleinert, M., Ehr, H., Murauskas, Š., and Pagirys, T. (2021b). Prediction and extraction of tower controller commands for speech recognition applications. *Journal of Air Transport Management*, 95:102089.
https://doi.org/10.1016/j.jairtraman.2021.102089.

[Ohneiser et al., 2023] Ohneiser, O., Helmke, H., Shetty, S., Kleinert, M., Ehr, H., Schier-Morgenthal, S., Sarfjoo, S., Motlíček, P., Murauskas, Š., Pagirys, T., Usanovic, H., Meštrović, M., and Černá, A. (2023). Assistant Based Speech Recognition Support for Air Traffic Controllers in a Multiple Remote Tower Environment. *Aerospace, Special Issue Automatic Speech Recognition and Understanding in Air Traffic Management*, 10(6).
https://doi.org/10.3390/aerospace10060560.

[Ohneiser et al., 2018c] Ohneiser, O., Jauer, M.-L., Gürlük, H., and Springborn, H. (2018c). Attention Guidance Prototype for a Sectorless Air Traffic Management Controller Working Position. In *Deutscher Luft- und Raumfahrtkongress*, DLRK 2018, Friedrichshafen, Germany, 04-06 Sep, 2018.
https://doi.org/10.25967/480189.

[Ohneiser et al., 2016] Ohneiser, O., Jauer, M.-L., Gürlük, H., and Uebbing-Rumke, M. (2016). TriControl – A Multimodal Air Traffic Controller Working Position. In *6th SESAR Innovation Days*, SID 2016, Delft, The Netherlands, 08-10 Nov, 2016.
https://s.dlr.de/bM1th.

[Ohneiser et al., 2018d] Ohneiser, O., Jauer, M.-L., Rein, J. R., and Wallace, M. (2018d). Faster Command Input Using the Multimodal Controller Working Position "TriControl". *Aerospace*, 5(2).
https://doi.org/10.3390/aerospace5020054.

[Ohneiser et al., 2019c] Ohneiser, O., Kleinert, M., Muth, K., Gluchshenko, O., Ehr, H., Groß, N., and Temme, M.-M. (2019c). Bad Weather Highlighting: Advanced Visualization of Severe Weather and Support in Air Traffic Control Displays. In *IEEE/AIAA 38th Digital Avionics Systems Conference*, DASC 2019, San Diego, CA, USA, 08-12 Sep, 2019.
https://doi.org/10.1109/DASC43569.2019.9081773.

[Ohneiser et al., 2021c] Ohneiser, O., Sarfjoo, S., Helmke, H., Shetty, S., Motlíček, P., Kleinert, M., Ehr, H., and Murauskas, Š. (2021c). Robust Command Recognition for Lithuanian Air Traffic Control Tower Utterances. In *22nd Annual Conference of the International Speech Communication*

Association, Proc. of Interspeech 2021, pages 3291–3295, Brno, Czechia (hybrid), 30 Aug-03 Sep, 2021. ISCA.
https://doi.org/10.21437/Interspeech.2021-935.

[Ohneiser et al., 2015] Ohneiser, O., Temme, M.-M., and Rataj, J. (2015). Trawl-Net Technology for Timely Precise Air Traffic Controller Turn-To-Base Commands. In *11th USA/Europe Air Traffic Management Research and Development Seminar*, ATM 2015, Lisbon, Portugal, 23-26 Jun, 2015.
https://s.dlr.de/vFUP0.

[Oviatt, 1999] Oviatt, S. (1999). Ten myths of multimodal interaction. *Commun. ACM*, 42(11):74–81.
https://doi.org/10.1145/319382.319398.

[Oviatt, 2002] Oviatt, S. (2002). Multimodal Interfaces. In Jacko, J. A. and Sears, A., editors, *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, pages 286–304, Hillsdale, NJ, USA. Lawrence Erlbaum Associates Inc., 2003.
https://s.dlr.de/HcNOx.

[Oviatt et al., 2004] Oviatt, S., Coulston, R., and Lunsford, R. (2004). When do we interact multimodally? Cognitive load and multimodal communication patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI 2004, pages 129–136, State College, PA, USA, 13-15 Oct, 2004, New York, NY, USA. Association for Computing Machinery.
https://doi.org/10.1145/1027933.1027957.

[Palmer et al., 2008] Palmer, E. M., Clausner, T. C., and Kellman, P. J. (2008). Enhancing air traffic displays via perceptual cues. *ACM Transactions on Applied Perception*, 5(1):1–22.
https://doi.org/10.1145/1279640.1279644.

[Parasuraman and Riley, 1997] Parasuraman, R. and Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2):230–253.
https://doi.org/10.1518/001872097778543886.

[Parasuraman et al., 2008] Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2008). Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making*, 2(2):140–160.
https://doi.org/10.1518/155534308X284417.

[Pardo et al., 2011] Pardo, J. M., Ferreiros, J., Fernández, F., Sama, V., De Córdoba, R., Macias-Guarasa, J., Montero, J. M., San-Segundo, R., D'Haro, L. F., and González, G. (2011). Automatic Understanding of ATC Speech: Study of Prospectives and Field Experiments for Several Controller Positions. *IEEE Transactions on Aerospace and Electronic Systems*, 47(4):2709–2730.
https://doi.org/10.1109/TAES.2011.6034660.

[Peißl et al., 2018] Peißl, S., Wickens, C. D., and Baruah, R. (2018). Eye-Tracking Measures in Aviation: A Selective Literature Review. *The International Journal of Aerospace Psychology*, 28(3-4):98–112.
https://doi.org/10.1080/24721840.2018.1514978.

[Pellegrini et al., 2019] Pellegrini, T., Farinas, J., Delpech, E., and Lancelot, F. (2019). The Airbus Air Traffic Control Speech Recognition 2018 Challenge: Towards ATC Automatic Transcription and Call Sign Detection. In *20th Annual Conference of the International Speech Communication Association*,

Proc. of Interspeech 2019, pages 2993–2997, Graz, Austria, 15-19 Sep, 2019. ISCA.
https://doi.org/10.21437/Interspeech.2019-1962.

[Peysakhovich et al., 2018] Peysakhovich, V., Lefrançois, O., Dehais, F., and Causse, M. (2018). The Neuroergonomics of Aircraft Cockpits: The Four Stages of Eye-Tracking Integration to Enhance Flight Safety. *Safety*, 4(1).
https://doi.org/10.3390/safety4010008.

[Pinska-Chauvin et al., 2023] Pinska-Chauvin, E., Helmke, H., Dokic, J., Hartikainen, P., Ohneiser, O., and García Lasheras, R. (2023). Ensuring Safety for Artificial-Intelligence-Based Automatic Speech Recognition in Air Traffic Control Environment. *Aerospace*, 10(11).
https://doi.org/10.3390/aerospace10110941.

[Placek, 2024] Placek, M. (2024). The pilot shortage - statistics & facts. statista.com.
https://s.dlr.de/QTdOH [Accessed: 16 Dec, 2024].

[Povey et al., 2011] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N. K., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý, K. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hilton Waikoloa Village, Big Island, HI, USA, 11-15 Dec, 2011. IEEE Signal Processing Society.
https://s.dlr.de/Y02Me.

[Powers, 2011] Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiV*.
https://doi.org/10.48550/arXiv.2010.16061.

[Prakash et al., 2016] Prakash, A., Swathi, R., Kumar, S., Ashwin, T. S., and Reddy, G. R. M. (2016). Kinect Based Real Time Gesture Recognition Tool for Air Marshallers and Traffic Policemen. In *IEEE Eighth International Conference on Technology for Education*, T4E 2016, pages 34–37, Los Alamitos, CA, USA, Mumbai, India, 02-04 Dec, 2016. IEEE Computer Society.
https://doi.org/10.1109/T4E.2016.015.

[Prasad et al., 2021] Prasad, A., Zuluaga-Gómez, J. P., Motlíček, P., Ohneiser, O., Helmke, H., Sarfjoo, S., and Nigmatulina, I. (2021). Grammar Based Identification Of Speaker Role For Improving ATCO And Pilot ASR. In *Interspeech 2021 Satellite Workshop 'Automatic Speech Recognition in Air Traffic Management'*, ASR-ATM 2021, Brno, Czechia (hybrid), 30 Aug, 2021.
https://s.dlr.de/v1SDN.

[Prasad et al., 2022] Prasad, A., Zuluaga-Gómez, J. P., Motlíček, P., Sarfjoo, S., Nigmatulina, I., Ohneiser, O., and Helmke, H. (2022). Grammar Based Speaker Role Identification for Air Traffic Control Speech Recognition. In *12th SESAR Innovation Days*, SID 2022, Budapest, Hungary, 05-08 Dec, 2022.
https://s.dlr.de/VHwPx.

[Quek et al., 2000] Quek, F., McNeill, D., Bryll, R., Kirbas, C., Arslan, H., McCullough, K. E., Furuyama, N., and Ansari, R. (2000). Gesture, speech, and gaze cues for discourse segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2 of *CVPR 2000*, pages 247–254, Hilton Head Island, SC, USA, 13-15 Jun, 2000.
https://doi.org/10.1109/CVPR.2000.854800.

[Rajendram Bashyam et al., 2023] Rajendram Bashyam, L., Blatt, A., and Klakow, D. (2023). Enabling Noisy Label Usage for Out-Of-Airspace Data in Read-Back Error Detection. In *Workshop on Automatic Speech Recognition and Understanding*, ASRU 2023, Taipei, Taiwan, 16-20 Dec, 2023.
https://s.dlr.de/kLHTB.

[Rataj et al., 2019] Rataj, J., Helmke, H., and Ohneiser, O. (2019). AcListant with Continuous Learning: Speech Recognition in Air Traffic Control. In *ENRI International Workshop on ATM/CNS*, EIWAC 2019, Tokyo, Japan, 29-31 Oct, 2019.
https://s.dlr.de/3XHDL.

[Rataj et al., 2021a] Rataj, J., Helmke, H., and Ohneiser, O. (2021a). AcListant with Continuous Learning: Speech Recognition in Air Traffic Control, Electronic Navigation Research Institute, editor. In *Lecture Notes in Electrical Engineering: Air Traffic Management and Systems IV*, volume 731, pages 93–109, Singapore. Springer.
https://doi.org/10.1007/978-981-33-4669-7_6.

[Rataj et al., 2021b] Rataj, J., Ohneiser, O., Marin, G., and Postaru, R. (2021b). Attention: Target and Actual – The Controller Focus. In *32nd Congress of the International Council of the Aeronautical Sciences*, ICAS 2021, Shanghai, China (hybrid), 06-10 Sep, 2021.
https://s.dlr.de/LYw7z.

[Roscoe, 1984] Roscoe, A. H. (1984). Assessing Pilot Workload in Flight. In *Flight Test Techniques, Proceedings of the Flight Mechanics Panel Symposium*, Lisbon, Portugal, 02-05 Apr, 1984. Advisory Group for Aerospace Research and Development, Neuilly-sur-Seine, France; Royal Aircraft Establishment, Bedford, United Kingdom.
https://s.dlr.de/R56jU.

[Rouwhorst et al., 2017] Rouwhorst, W., Verhoeven, R., Suijkerbuijk, M., Bos, T., Maij, A., Vermaat, M., and Arents, R. (2017). Use of touch screen display applications for aircraft flight control. In *IEEE/AIAA 36th Digital Avionics Systems Conference*, DASC 2017, St. Petersburg, FL, USA, 17-21 Sep, 2017.
https://doi.org/10.1109/DASC.2017.8102060.

[Rudregowda et al., 2024] Rudregowda, S., Patilkulkarni, S., Ravi, V., H.L., G., and Krichen, M. (2024). Audiovisual speech recognition based on a deep convolutional neural network. *Data Science and Management*, 7(1):25–34.
https://doi.org/10.1016/j.dsm.2023.10.002.

[Savery et al., 2013] Savery, C., Hurter, C., Lesbordes, R., Cordeil, M., and Graham, T. C. N. (2013). When Paper Meets Multi-touch: A Study of Multi-modal Interactions in Air Traffic Control. In Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., and Winckler, M., editors, *Human-Computer Interaction*, INTERACT 2013, pages 196–213. Springer Berlin Heidelberg.
https://doi.org/10.1007/978-3-642-40477-1_12.

[Schäfer, 2000] Schäfer, D. (2000). *Context-Sensitive Speech Recognition in the Air Traffic Control Simulation*. PhD thesis, Universität der Bundeswehr München, München, Germany.
https://s.dlr.de/eLMa5.

[Schildt et al., 2013] Schildt, S., Pöttner, W.-B., Ohneiser, O., and Wolf, L. (2013). NASDI – Naming and Service Discovery for DTNs in Internet Backbones. In Borcea, C., Bellavista, P., Giannelli,

C., Magedanz, T., and Schreiner, F., editors, *Mobile Wireless Middleware, Operating Systems, and Applications*, 5th International Conference, Mobilware 2012, pages 108–121, Berlin, Germany, 13-14 Nov, 2013. Springer Berlin Heidelberg.
https://doi.org/10.1007/978-3-642-36660-4_8.

[Schmidt et al., 2014] Schmidt, A., Oualil, Y., Ohneiser, O., Kleinert, M., Schulder, M., Khan, A., Helmke, H., and Klakow, D. (2014). Context-based Recognition Network Adaptation for Improving On-line ASR in Air Traffic Control. In *IEEE Spoken Technology Workshop*, SLT 2014, South Lake Tahoe, CA, USA, 07-10 Dec, 2014.
https://doi.org/10.1109/SLT.2014.7078542.

[Seelmann, 2015] Seelmann, P.-E. (2015). Evaluation of an Eye Tracking and Multi-Touch Based Operational Concept for a Future Multimodal Approach Controller Working Position, original German Title: Evaluierung eines auf Eyetracking und Multi-Touch basierten Bedienkonzeptes für einen zukünftigen multimodalen Anfluglotsenarbeitsplatz; DLR-IB 112-2015/44. Technical report, Ostfalia - University of Applied Science, Wolfenbüttel, Germany, Bachelor Thesis.
https://s.dlr.de/JDcxc.

[SESAR3JU, 2020] SESAR3JU (2020). *European ATM master plan - Digitalising Europe's Aviation Infrastructure - Executive View*. Single European Sky ATM Research 3 Joint Undertaking, Publications Office, Brussels, Belgium.
https://doi.org/10.2829/695700.

[Sheridan and Parasuraman, 2005] Sheridan, T. B. and Parasuraman, R. (2005). Human-Automation Interaction. *Reviews of Human Factors and Ergonomics*, 1(1):89–129.
https://doi.org/10.1518/155723405783703082.

[Shetty et al., 2022] Shetty, S., Helmke, H., Kleinert, M., and Ohneiser, O. (2022). Early Callsign Highlighting using Automatic Speech Recognition to Reduce Air Traffic Controller Workload. In *International Conference on Applied Human Factors and Ergonomics*, volume 60 of *AHFE 2022*, pages 584–592, New York, NY, USA (hybrid), 24-28 Jul, 2022.
https://doi.org/10.54941/ahfe1002493.

[Shetty et al., 2020] Shetty, S., Ohneiser, O., Grezl, F., Helmke, H., and Motlíček, P. (2020). Transcription and Annotation Handbook. Technical Report HAAWAII project deliverable D3.1, German Aerospace Center (DLR), Institute of Flight Guidance, Braunschweig, Germany.
https://s.dlr.de/wPzFe.

[Shore et al., 2012] Shore, T., Faubel, F., Helmke, H., and Klakow, D. (2012). Knowledge-Based Word Lattice Rescoring in a Dynamic Context. In *13th Annual Conference of the International Speech Communication Association*, Proc. of Interspeech 2012, pages 1083–1086, Portland, OR, USA, 09-13 Sep, 2012. ISCA.
https://doi.org/10.21437/Interspeech.2012-328.

[Simpson and Williams, 1980] Simpson, C. A. and Williams, D. H. (1980). Response Time Effects of Alerting Tone and Semantic Context for Synthesized Voice Cockpit Warnings. *Human factors*, 22(3):319–330.
https://doi.org/10.1177/001872088002200306.

[Singh et al., 2005] Singh, M., Mandal, M., and Basu, A. (2005). Visual gesture recognition for ground air traffic control using the Radon transform. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IROS 2005, pages 2586–2591, Edmonton, AB, Canada, 02-06 Aug, 2005. https://doi.org/10.1109/IROS.2005.1545408.

[Sklar and Sarter, 1999] Sklar, A. E. and Sarter, N. B. (1999). Good Vibrations: Tactile Feedback in Support of Attention Allocation and Human-Automation Coordination in Event-Driven Domains. *Human Factors*, 41(4):543–552. https://doi.org/10.1518/001872099779656716.

[Slotty and Rühl, 2012] Slotty, M. and Rühl, O. (2012). Speech recognition finds its way into DFS - Procedure for the introduction of speech recognition in research and training applications, original German title: Spracherkennung findet Einzug in der DFS - Vorgehensweise bei der Einführung von Spracherkennung im Forschungs- und Trainingseinsatz. *TE im Fokus*, (2):31–37. https://s.dlr.de/8I72H.

[Smídl et al., 2019] Smídl, L., Svec, J., Tihelka, D., Matousek, J., Romportl, J., and Ircing, P. (2019). Air Traffic Control Communication (ATCC) Speech Corpora and their Use for ASR and TTS Development. *Lang. Resour. Evaluation*, 53(3):449–464. https://doi.org/10.1007/s10579-019-09449-5.

[Smisek et al., 2017] Smisek, J., Sunil, E., van Paassen, M. M., Abbink, D. A., and Mulder, M. (2017). Neuromuscular-System-Based Tuning of a Haptic Shared Control Interface for UAV Teleoperation. *IEEE Transactions on Human-Machine Systems*, 47(4):449–461. https://doi.org/10.1109/THMS.2016.2616280.

[Steinmetz et al., 2024] Steinmetz, H. A., Tao, J., Clarke, S. S., and Kalyanam, K. (2024). A Natural Language Understanding Approach for Digitizing Aircraft Ground Taxi Instructions. In *AIAA AVIATION FORUM AND ASCEND*, MAVAS 2024, Las Vegas, NV, USA, 29 Jul-02 Aug, 2024. https://doi.org/10.2514/6.2024-4359.

[Stroop, 1935] Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6):643–662. https://doi.org/10.1037/h0054651.

[Temme et al., 2023] Temme, M.-M., Gluchshenko, O., Nöhren, L., Kleinert, M., Ohneiser, O., Muth, K., Ehr, H., Groß, N., Temme, A., Lagasio, M., Milelli, M., Mazzarella, V., Parodi, A., Realini, E., Federico, S., Torcasio, R. C., Kerschbaum, M., Esbrí, L., Llasat, M. C., Rigo, T., and Biondi, R. (2023). Innovative Integration of Severe Weather Forecasts into an Extended Arrival Manager. *Aerospace*, 10(3). https://doi.org/10.3390/aerospace10030210.

[Teutsch et al., 2022] Teutsch, J., Bos, T., van Apeldoorn, M., and Camara, L. (2022). Attention Guidance for Tower ATC Using Augmented Reality Devices. In *Integrated Communication, Navigation and Surveillance Conference*, ICNS 2022, Herndon, VA, USA, 05-07 Apr, 2022. https://doi.org/10.1109/ICNS54818.2022.9771479.

[Thorburn, 1971] Thorburn, D. E. (1971). Voice Warning Systems – A Cockpit Improvement that should not be overlooked, AMRL-TR-70-138. Technical report, Aerospace Medical Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, OH,

USA.
https://s.dlr.de/TN5Tb.

[Traoré and Hurter, 2016] Traoré, M. and Hurter, C. (2016). Exploratory Study with Eye Tracking devices to build Interactive Systems for Air Traffic Controllers. In *Proceedings of the International Conference on Human-Computer Interaction in Aerospace*, HCI-Aero 2016, Paris, France, 14-16 Sep, 2016, New York, NY, USA. Association for Computing Machinery.
https://doi.org/10.1145/2950112.2964584.

[Uebbing-Rumke et al., 2014] Uebbing-Rumke, M., Gürlük, H., Jauer, M., Hagemann, K., and Udovic, A. (2014). Usability Evaluation of Multi-Touch-Displays for TMA Controller Working Positions. In *4th SESAR Innovation Days*, SID 2014, Madrid, Spain, 25-27 Nov, 2014.
https://s.dlr.de/ch7k4.

[Wald, 2011] Wald, D. (2011). Programming and Evaluation of a touch based air traffic control, original German title: Programmierung und Evaluierung einer Touch basierten Flugverkehrskontrolle. Technical report, Hochschule Fulda, Fulda, Germany. Master Thesis.

[Wallace, 2024] Wallace, G. (2024). FAA still short about 3,000 air traffic controllers, new federal numbers show. CNN.com.
https://s.dlr.de/lxpVu [Accessed: 16 Dec, 2024].

[Wang et al., 2021] Wang, Y., Wang, L., Lin, S., Cong, W., Xue, J., and Ochieng, W. (2021). Effect of Working Experience on Air Traffic Controller Eye Movement. *Engineering*, 7(4):488–494.
https://doi.org/10.1016/j.eng.2020.11.006.

[Wang, 2023] Wang, Z. (2023). Research Trends and Applications of Human-Machine Interface in Air Traffic Control: A Scientometric Analysis. In Harris, D. and Li, W.-C., editors, *Engineering Psychology and Cognitive Ergonomics*, volume 14018 of *Lecture Notes in Computer Science, Human Computer Interaction (HCI) International 2023*, pages 379–390, Copenhagen, Denmark, 23-28 Jul, 2023. Springer Nature, Cham, Switzerland.
https://doi.org/10.1007/978-3-031-35389-5_26.

[Wechsung, 2014] Wechsung, I. (2014). What Are Multimodal Systems? Why Do They Need Evaluation? — Theoretical Background. In *An Evaluation Framework for Multimodal Interaction: Determining Quality Aspects and Modality Choice*, pages 7–22. Springer International Publishing, Cham, Switzerland.
https://doi.org/10.1007/978-3-319-03810-0_2.

[Wee et al., 2017] Wee, H. J., Lye, S. W., and Pinheiro, J.-P. (2017). Real Time Eye Tracking Interface for Visual Monitoring of Radar Controllers. In *AIAA Modeling and Simulation Technologies Conference*, Grapevine, TX, USA, 09-13 Jan, 2017.
https://doi.org/10.2514/6.2017-1317.

[Westin et al., 2019] Westin, C., Vrotsou, K., Nordman, A., and Lundberg, J. (2019). Visual Scan Patterns in Tower Control: Foundations for an Instructor Support Tool. In *9th SESAR Innovation Days*, SID 2019, Athens, Greece, 02-05 Dec, 2019.
https://s.dlr.de/aGqTU.

[Wickens et al., 1997] Wickens, C. D., Mavor, A. S., and McGee, J. P. (1997). *Flight to the future: Human factors in air traffic control.* National Academy Press, Washington D.C., USA.

https://s.dlr.de/JMtdh.

[Xu et al., 2024] Xu, L., Ma, S., Shen, Z., Huang, S., and Nan, Y. (2024). Analyzing Multi-Mode Fatigue Information from Speech and Gaze Data from Air Traffic Controllers. *Aerospace*, 11(1). https://doi.org/10.3390/aerospace11010015.

[Zhang et al., 2022] Zhang, J., Zhang, P., Guo, D., Zhou, Y., Wu, Y., Yang, B., and Lin, Y. (2022). Automatic repetition instruction generation for air traffic control training using multi-task learning with an improved copy network. *Knowledge-Based Systems*, 241:108232. https://doi.org/10.1016/j.knosys.2022.108232.

[Ziv, 2016] Ziv, G. (2016). Gaze Behavior and Visual Attention: A Review of Eye Tracking Studies in Aviation. *The International Journal of Aviation Psychology*, 26(3-4):75–104. https://doi.org/10.1080/10508414.2017.1313096.

[Zuluaga-Gómez et al., 2020] Zuluaga-Gómez, J. P., Motlíček, P., Zhan, Q., Braun, R., and Vesely, K. (2020). Automatic Speech Recognition Benchmark for Air-Traffic Communications. In *21st Annual Conference of the International Speech Communication Association*, Proc. of Interspeech 2020, pages 2297–2301, Virtual, 25-29 Oct, 2020. ISCA. https://doi.org/10.21437/Interspeech.2020-2173.

[Zuluaga-Gómez et al., 2023a] Zuluaga-Gómez, J. P., Nigmatulina, I., Prasad, A., Motlíček, P., Khalil, D., Madikeri, S., Tart, A., Szoke, I., Lenders, V., Rigault, M., and Choukri, K. (2023a). Lessons Learned in Transcribing 5000 h of Air Traffic Control Communications for Robust Automatic Speech Understanding. *Aerospace*, 10(10). https://doi.org/10.3390/aerospace10100898.

[Zuluaga-Gómez et al., 2023b] Zuluaga-Gómez, J. P., Prasad, A., Nigmatulina, I., Motlíček, P., and Kleinert, M. (2023b). A Virtual Simulation-Pilot Agent for Training of Air Traffic Controllers. *Aerospace*, 10(5). https://doi.org/10.3390/aerospace10050490.

[Zuluaga-Gómez et al., 2023c] Zuluaga-Gómez, J. P., Prasad, A., Nigmatulina, I., Sarfjoo, S., Motlíček, P., Kleinert, M., Helmke, H., Ohneiser, O., and Zhan, Q. (2023c). How Does Pre-trained Wav2Vec 2.0 Perform on Domain Shifted ASR? An Extensive Benchmark on Air Traffic Control Communications. In *2022 IEEE Spoken Language Technology Workshop*, SLT 2022, pages 205–212, Doha, Qatar, 09-12 Jan, 2023. https://doi.org/10.1109/SLT54892.2023.10022724.

[Zuluaga-Gómez et al., 2023d] Zuluaga-Gómez, J. P., Sarfjoo, S. S., Prasad, A., Nigmatulina, I., Motlíček, P., Ohneiser, O., and Helmke, H. (2023d). BERTraffic: BERT-based Joint Speaker Role and Speaker Change Detection for Air Traffic Control Communications. In *2022 IEEE Spoken Language Technology Workshop*, SLT 2022, pages 633–640, Doha, Qatar, 09-12 Jan, 2023. https://doi.org/10.1109/SLT54892.2023.10022718.

List of Own Publications and Description of Habilitand Contribution

The following list shows all habilitation relevant scientific publications of the habilitand that have been published after the submission of his doctoral thesis in 2016 until 2024[1]. The bibliometric list highlights in each entry in bold font the name of the habilitand to spot the authorship position. Furthermore, the journal articles as well as six "best paper" awards for conference papers, two "best presentation" poll achievements, and an article that has been selected for the cover story of the *Aerospace* journal are highlighted in bold font. The list consists of 52 publications including 15 journal articles. The journal articles have been published in seven different journals with impact factors between 2.1 and 6.0 – with slight deviations depending on year and metrics analysis organization, ranking in the best quartiles of their field. It also comprises of 28 full triple-peer reviewed and 8 abstract-reviewed conference papers. Among the mentioned 52 papers, the habilitand has been first author of 15 and second author of 7 papers. Those 22 papers as first or second author are listed with an aircraft icon bullet (✈) compared to the hand/pen icon bullet (✍). The aircraft indicated papers have some additional explanation of their content and information about the habilitand's portion of authorship regarding DLR contribution as listed below.

Among the papers written as first author, 6 papers are journal articles and 7 papers are full triple-peer reviewed conference papers – some of the latter having won awards as detailed below. Eleven of those thirteen papers in first authorship are the core contribution to the cumulative habilitation thesis. Hence, the habilitand's contribution to them is explained more in detail in the first part of the list of own publications even if at least all journal articles contain a basic CRediT (Contributor Roles Taxonomy) authorship contribution statement. Furthermore, some copyright details for reprinting the eleven papers are noted. These papers' journals and conferences have an average impact factor of 2.2. They were written with four and a half co-authors in average and have more than five citations per paper. The order of the eleven publication copies is again based on the human modalities, i.e., auditory, tactile, and visual modality as well as a combination of these modalities. The other 41 papers with habilitand contribution are listed in anti-chronological order and are partly detailed in the second part. The total of 52 papers has been written together with 125 authors from 51 institutions from academia and industry – e.g., research organizations, universities, air navigation service providers, air traffic management system suppliers, and airports – being located in 19 countries on 3 continents.

The habilitand was also one of two guest editors for the special issue "Automatic Speech Recognition and Understanding in Air Traffic Management" (ASRU in ATM) of the journal *Aerospace* [Helmke et al., 2024b] that has been printed as a hard cover book. The edited book comprises 12 research articles from 54 different authors working for 23 institutions in 13 countries on 4 continents. It investigates advantages and disadvantages of ASRU in ATM and emphasizes the need for an industry transfer for operational use of this technology. The habilitand has been identified as one of the Top-5 authors world-wide regarding research on human-machine interface in air traffic control [Wang, 2023].

[1] One journal paper has been accepted for publication in 2024 and has finally been published in 2025. One technical report of minor importance has already been published in December 2015.

✈ **Ohneiser, O.**, Sarfjoo, S., Helmke, H., Shetty, S., Motlíček, P., Kleinert, M., Ehr, H., and Murauskas, S. Robust Command Recognition for Lithuanian Air Traffic Control Tower Utterances. In *22nd Annual Conference of the International Speech Communication Association, Proc. of Interspeech 2021*, pages 3291–3295, Brno, Czechia (hybrid), 30 Aug - 03 Sep, 2021. ISCA.
`https://doi.org/10.21437/Interspeech.2021-935` reprinted as article →**1** in the annex

This paper describes the quality of speech recognition and understanding output applied on ATC tower utterances. The habilitand was responsible for the relevant ASRU parts when conducting the human-in-the-loop simulation trials within DLR's Air Traffic Validation Center. He supported the conceptualization and implementation of ATC concept extraction together with the primary programming activities from Helmke, Shetty, and Kleinert, as well as its testing and integration mainly done by Ehr. Sarfjoo and Motlíček were responsible for the speech recognition part of ASRU. Murauskas provided expertise from the operational air navigation service provider side. The habilitand did the complete data analysis and reporting including evaluation of ASRU metrics on his own. The paper has 65 % habilitand authorship. The paper acceptance rate of the conference with impact factor 1.3 was just below 50 %.

This paper can be reused without permission for integration in a thesis by the main author due to *Interspeech copyright*: "For any article published in Interspeech proceedings, ISCA grants each author permission to use the article in that author's dissertation or in institutional repositories (paper and/or electronic versions), provided that the article is correctly referenced (including page numbers and/or paper number)".

✈ **Ohneiser, O.**, Helmke, H., Shetty, S., Kleinert, M., Ehr, H., Murauskas, S., and Pagirys, T. Prediction and extraction of tower controller commands for speech recognition applications. ***Journal of Air Transport Management***, 95:102089, 2021.
`https://doi.org/10.1016/j.jairtraman.2021.102089` reprinted as article →**2** in the annex

This journal paper explains predicting and extracting ATC tower/ground concepts of an ASRU system and outlines command recognition rates, command recognition error rates, and command prediction error rates on air traffic controller voice utterances in a laboratory multiple remote tower environment. The habilitand was responsible for the relevant ASRU parts when conducting the human-in-the-loop simulation trials within DLR's Air Traffic Validation Center. He supported the conceptualization and implementation of ATC concept prediction and extraction together with the primary programming activities from Helmke, Shetty, and Kleinert, as well as its testing and integration mainly done by Ehr. Murauskas and Pagirys provided expertise from the operational air navigation service provider side. The habilitand did the complete data analysis and reporting including evaluation of ASRU metrics on his own. The paper in the journal with impact factor 4.1 has 65 % habilitand authorship.

This paper can be reused without permission due to *Elsevier copyright policy* in "theses and dissertations which contain embedded final published articles as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publication on ScienceDirect".

✈ **Ohneiser, O.**, Helmke, H., Shetty, S., Kleinert, M., Ehr, H., Schier-Morgenthal, S., Sarfjoo, S., Motlíček, P., Murauskas, S., Pagirys, T., Usanovic, H., Meštrović, M., and Černá, A. Assistant based speech recognition support for air traffic controllers in a multiple remote tower environment. ***Aerospace***, Special Issue *Automatic Speech Recognition and Understanding in Air Traffic Management*, 10(6), 2023.
`https://doi.org/10.3390/aerospace10060560` reprinted as article →**3** in the annex

This journal paper presents the results of a human-in-the-loop simulation with controllers in a multiple remote tower environment being supported by ASRU in the solution condition instead of using an electronic touch pen for flight strip maintenance in the baseline condition. The habilitand was responsible for conducting the complete human-in-the-loop simulation trials within DLR's Air Traffic Validation Center. He supported the conceptualization and implementation of ATC concept prediction and extraction together with the primary programming activities from Helmke, Shetty, and Kleinert, as well as its testing and integration mainly done by Ehr. Schier-Morgenthal supported the conduction of the human-in-the-loop simulation from a technical perspective. Sarfjoo and Motlíček were responsible for the speech recognition part of ASRU. Murauskas, Pagirys, Usanovic, Meštrović, and Černá provided expertise from the operational air navigation service provider side. The habilitand did the complete data analysis and reporting including evaluation of ASRU metrics on his own and organized the transcription and annotation work for ATC utterances – as well doing a huge portion on his own. The paper in the journal with impact factor 2.7 has 70 % habilitand authorship.

✈ **Ohneiser, O.**, Ahmed, U. Text-To-Speech Application for Training of Aviation Radio Telephony Communication Operators. ***IEEE Transactions on Aerospace and Electronic Systems***, 61(2), pages 4542–4560, 2025.
`https://doi.org/10.1109/TAES.2024.3504493` reprinted as article →**4** in the annex

This journal paper explores using a text-to-speech (TTS) application to simulate aviation radio telephony communication. The application utilizes open-source pre-trained TTS models fine-tuned using publicly available ATC communication-specific datasets and synthesizes textual ATC utterances to simulate ATCo instructions and pilot responses. The habilitand was responsible for concept, evaluation, and reporting during supervision of the related master thesis of Ahmed. The paper in the journal with impact factor 5.1 has almost 100 % habilitand authorship.

✈ **Ohneiser, O.**, Ahlstrom, V., Tracy, K., and Williams, B. Comparison of Air Traffic Controller Display Techniques for Reaching Target Times at Significant Waypoints. In *IEEE/AIAA 37th Digital Avionics Systems Conference, DASC 2018*, pages 1092–1101. London, UK, 23-27 Sep, 2018.
https://doi.org/10.1109/DASC.2018.8569365 reprinted as article →**5** in the annex

This paper received the **Best Paper Award** for the "Air Traffic Control & Flight Planning" session. It investigates five different visual display aids to support approach controllers in giving ATC instructions in a time-based merging task of arrival streams. The habilitand was responsible for conceptualization with review from Ahlstrom, human-in-the-loop simulation, data analysis, and reporting. He supervised the implementation of the graphical user interface for the study by Tracy and Williams as part of his research semester with the FAA William J. Hughes Technical Center in Atlantic City, NJ, USA. The paper of the conference with impact factor 1.0 has almost 100 % habilitand authorship.

✈ **Ohneiser, O.**, Kleinert, M., Muth, K., Gluchshenko, O., Ehr, H., Groß, N., and Temme, M.-M. Bad Weather Highlighting: Advanced Visualization of Severe Weather and Support in Air Traffic Control Displays. In *IEEE/AIAA 38th Digital Avionics Systems Conference, DASC 2019*. San Diego, CA, USA, 08-12 Sep, 2019.
https://doi.org/10.1109/DASC43569.2019.9081773 reprinted as article →**6** in the annex

This paper received the **Best Paper Award** for the "Human Factors" session. It describes visual weather highlighting in an ATC situation data display to increase weather situation awareness of controllers. The habilitand supported the conceptualization of the weather visualization and the re-routing calculation that was mainly done by Gluchshenko and Temme. He implemented the re-routing advisories and supported the implementation of weather visualization together with Kleinert, Muth, and Groß as well as supported testing with Ehr. The paper of the conference with impact factor 1.0 has 40 % habilitand authorship.

✈ **Ohneiser, O.**, Gürlük, H., Jauer, M.-L., Szöllősi, Á., and Balló, D. Please have a Look here: Successful Guidance of Air Traffic Controller's Attention. In *9th SESAR Innovation Days, SID 2019*. Athens, Greece, 02-05 Dec, 2019.
`https://s.dlr.de/qyGZM` reprinted as article →**7** in the annex

This paper was **Best Presentation Poll** session winner. It reports on a human-in-the-loop validation study where eye tracking based attention guidance mechanisms helped en-route controllers to focus on relevant ATC events. The habilitand was responsible for conceptualization together with Jauer, the attention guidance relevant parts of the human-in-the-loop simulation, data analysis, and reporting. Szöllősi and Balló supported the validation study and provided expertise from the operational air navigation service provider side. He supervised the implementation of the visual display elements and eye tracking modules for the study. The paper has almost 100 % habilitand authorship. The paper acceptance rate of the conference with impact factor 0.7 was 53.1 %.

This paper can be reused without permission for integration in a thesis by the main author due to *SESAR 3 JU copyright*: "The SESAR 3 JU authorises the redistribution or reproduction of part or all of the contents in any form of this website, including documents and illustrations it contains, for non-commercial purposes provided this website is acknowledged as the source of the material: ©SESAR 3 JU 2022".

✈ **Ohneiser, O.**, Adamala, J., and Salomea, I.-T. Integrating Eye- and Mouse-Tracking with Assistant Based Speech Recognition for Interaction at Controller Working Positions. ***Aerospace***, Special Issue *Aeronautical Informatics*, 8(9), 2021.
`https://doi.org/10.3390/aerospace8090245` reprinted as article →**8** in the annex

This journal paper describes the use of eye-tracking (1) to improve ATC concept prediction for ASRU systems and (2) to verify human operators' visual checks of ASRU output. The habilitand was responsible for concepts, the human-in-the-loop simulation, data analysis, and reporting. He supervised the implementation of the visual display elements and eye tracking modules for the study. He supervised the master thesis of Adamala and the bachelor thesis of Salomea. The paper in the journal with impact factor 2.7 has almost 100 % habilitand authorship.

✈ **Ohneiser, O.**, Jauer, M.-L., Gürlük, H., and Uebbing-Rumke, M. TriControl – A Multimodal Air Traffic Controller Working Position. In *6th SESAR Innovation Days, SID 2016*. Delft, The Netherlands, 08-10 Nov, 2016.
`https://s.dlr.de/bM1th` reprinted as article →**9** in the annex

This paper details the setup of a multimodal controller working position integrating ASRU, gesture recognition, and eye tracking for entering ATC commands. The habilitand supported the conceptualization together with Gürlük, Uebbing-Rumke, and Jauer as well as supported implementation and testing of the prototype. He also conducted an initial system usability study and did the data analysis. The paper has 90 % habilitand authorship. The paper acceptance rate of the conference with impact factor 0.7 was 52.5 %.

✈ **Ohneiser, O.**, Jauer, M.-L., Rein, J. R., and Wallace, M. Faster Command Input Using the Multimodal Controller Working Position "TriControl". ***Aerospace***, 5(2), 2018. `https://doi.org/10.3390/aerospace5020054` reprinted as article →**10** in the annex

This journal paper analyzes the potential speed gain through use of ASRU, eye tracking, and gesture recognition in entering ATC commands at a multimodal controller working position. The habilitand was responsible for the human-in-the-loop simulation as well as major parts of data analysis and reporting. He supported the concept development and the implementation of the controller working position components for the study together with Jauer. Rein supported the data analysis while being a guest scientist at DLR under supervision of Ohneiser. Wallace coordinated the study subjects and provided expertise from the operational air navigation service provider side. The paper in the journal with impact factor 2.7 has 90 % habilitand authorship.

✈ **Ohneiser, O.**, Biella, M., Schmugler, A., and Wallace, M. Operational Feasibility Analysis of the Multimodal Controller Working Position "TriControl". ***Aerospace***, 7(2), 2020. `https://doi.org/10.3390/aerospace7020015` reprinted as article →**11** in the annex

This journal paper assesses various feasibility aspects of the aforementioned multimodal controller working position prototype at the given technology readiness level. The habilitand was responsible for the human-in-the-loop simulation and reporting. He co-supervised the concept development and data analysis mainly done by Schmugler under the supervision of Biella. Wallace coordinated the study subjects and provided expertise from the operational air navigation service provider side. The paper in the journal with impact factor 2.7 has 98 % habilitand authorship.

The decision to publish with the selected journals and conferences was positively influenced by, e.g., the involvement of relevant researchers in the field as authors, reviewers, speakers, and in the editorial boards, the impact factor and rank quartile in the field, and open access. The publishing decision was hardly influenced by later publications being indexed in the Computer Science relevant *DBLP*[2] database as Aeronautical Informatics is a small field – with often costly validation activities – and papers or venues from this field are hardly listed in this index. However, the listing and automatic tagging with, e.g., *Computer Science* and *Engineering* as paper content, is much better elaborated at *Semantic Scholar*[3], *Google Scholar*[4], or *ORCID*[5].

[2]DBLP indexed habilitand papers: `https://dblp.org/pid/125/3592.html`
[3]*Semantic Scholar* profile of habilitand: `https://www.semanticscholar.org/author/Oliver-Ohneiser/2093957049`
[4]*Google Scholar* profile of habilitand: `https://scholar.google.com/citations?user=RngdarYAAAAJ&hl=de&oi=ao`
[5]*ORCID* profile of habilitand: `https://orcid.org/0000-0002-5411-691X`

In case some of the selected papers have more than a handful of co-authors, the reason is mainly the addition of ANSP colleagues who have the operational expertise to bring the developed prototypes to TRL6. Furthermore, the number of co-authors shows the habilitand's international network and the interdisciplinary work combining Computer Science, Aviation, Human Factors, etc. The following list presents the further relevant papers and the book of the habilitand as outlined at the beginning of this section.

- ✍ Bhattacharjee, M., Motlíček, P., Madikeri, S., Helmke, H., **Ohneiser, O.**, Kleinert, M., and Ehr, H. Minimum effort adaptation of automatic speech recognition system in air traffic management. ***European Journal of Transport and Infrastructure Research (EJTIR)***, 24(4), pages 133–153, 2024. `https://doi.org/10.59490/ejtir.2024.24.4.7531`

- ✍ Meier, J., Finke, M., **Ohneiser, O.**, Jameel, M. Flexible Air Traffic Controller Deployment with Artificial Intelligence based Decision Support: Literature Survey and Evaluation Framework. In *Deutscher Luft- und Raumfahrtkongress, DLRK 2024*. Hamburg, Germany, 30 Sep-02 Oct, 2024. `https://doi.org/10.25967/630258`

- ✍ Helmke, H., Kleinert, M., **Ohneiser, O.**, Ahrenhold, N., Klamert, L., and Motlíček, P. Safety and Workload Benefits of Automatic Speech Understanding for Radar Label Updates. ***Journal of Air Transportation***, 32(4), pages 155–168, 2024. `https://doi.org/10.2514/1.D0419`

- ✍ Motlíček, P., Prasad, A., Nigmatulina, I., Helmke, H., **Ohneiser, O.**, and Kleinert, M. Automatic Speech Analysis Framework for ATC Communication in HAAWAII. In *13th SESAR Innovation Days, SID 2023*. Seville, Spain, 27-30 Nov, 2023. `https://s.dlr.de/ost2A`

- ✍ Bhattacharjee, M., Motlíček, P., Nigmatulina, I., Helmke, H., **Ohneiser, O.**, Kleinert, M., and Ehr, H. Customization of automatic speech recognition engines for rare word detection without costly model re-training. In *13th SESAR Innovation Days, SID 2023*. Seville, Spain, 27-30 Nov, 2023. `https://s.dlr.de/kSSsy`

- ✍ Pinska-Chauvin, E., Helmke, H., Dokic, J., Hartikainen, P., **Ohneiser, O.**, and Lasheras, R. G. Ensuring safety for artificial-intelligence-based automatic speech recognition in air traffic control environment. ***Aerospace***, 10(11), 2023. `https://doi.org/10.3390/aerospace10110941` has been selected for the **cover story** of the *Aerospace* journal's issue 11 "November 2023" of volume 10[6] from more than sixty papers.

- ✍ Ahrenhold, N., Helmke, H., Mühlhausen, T., Kleinert, M., **Ohneiser, O.**, and Ehr, H. Influence of Automatic Speech Recognition and Understanding on Flight Efficiency and Throughput – A Human-in-the-Loop Study. In *IEEE/AIAA 42nd Digital Avionics Systems Conference, DASC 2023*. Barcelona, Spain, 01-05 Oct, 2023. `https://doi.org/10.1109/DASC58513.2023.10311293`

- ✈ Kleinert, M., **Ohneiser, O.**, Helmke, H., Shetty, S., Ehr, H., Maier, M., Schacht, S., and Wiese, H. Safety aspects of supporting apron controllers with automatic speech recognition and understanding integrated into an advanced surface movement guidance and control system. ***Aerospace***, 10(7), 2023. `https://doi.org/10.3390/aerospace10070596` investigates ASRU for automatic system input for apron controllers at Frankfurt airport; 15 % authorship

[6] *Aerospace* journal, November 2023 with ASRU as cover story: `https://www.mdpi.com/2226-4310/10/11`

✍ Helmke, H., Kleinert, M., Ahrenhold, N., Ehr, H., Mühlhausen, T., **Ohneiser, O.**, Klamert, L., Motlíček, P., Prasad, A., Zuluaga-Gómez, J. P., Dokic, J., and Pinska Chauvin, E. Automatic Speech Recognition and Understanding for Radar Label Maintenance Support Increases Safety and Reduces Air Traffic Controllers' Workload. In *15th USA/Europe Air Traffic Management Research and Development Seminar, ATM 2023*. Savannah, GA, USA, 05-09 Jun, 2023. `https://s.dlr.de/100Zw` received **Best Paper Award** for "Human Factors" track

✈ Helmke, H., **Ohneiser, O.**, Kleinert, M., Chen, S., Kopald, H., and Tarakan, R. M. Transatlantic Approaches for Automatic Speech Understanding in Air Traffic Management. In *15th USA/Europe Air Traffic Management Research and Development Seminar, ATM 2023*. Savannah, GA, USA, 05-09 Jun, 2023. `https://s.dlr.de/AOUnj` compares the European with a US American ontology for annotation of ATC voice commands; 30 % authorship

✍ Ahrenhold, N., Helmke, H., Mühlhausen, T., **Ohneiser, O.**, Kleinert, M., Ehr, H., Klamert, L., and Zuluaga-Gómez, J. P. Validating Automatic Speech Recognition and Understanding for Pre-Filling Radar Labels – Increasing Safety While Reducing Air Traffic Controllers' Workload. ***Aerospace***, 10(6), 2023. `https://doi.org/10.3390/aerospace10060538`

✍ Chen, S., Helmke, H., Tarakan, R. M., **Ohneiser, O.**, Kopald, H., and Kleinert, M. Effects of language ontology on transatlantic automatic speech understanding research collaboration in the air traffic management domain. ***Aerospace***, 10(6), 2023. `https://doi.org/10.3390/aerospace10060526`

✍ Temme, M.-M., Gluchshenko, O., Nöhren, L., Kleinert, M., **Ohneiser, O.**, Muth, K., Ehr, H., Groß, N., Temme, A., Lagasio, M., Milelli, M., Mazzarella, V., Parodi, A., Realini, E., Federico, S., Torcasio, R. C., Kerschbaum, M., Esbrí, L., Llasat, M. C., Rigo, T., and Biondi, R. Innovative integration of severe weather forecasts into an extended arrival manager. ***Aerospace***, 10(3), 2023. `https://doi.org/10.3390/aerospace10030210`

✍ Zuluaga-Gómez, J. P., Prasad, A., Nigmatulina, I., Sarfjoo, S., Motlíček, P., Kleinert, M., Helmke, H., **Ohneiser, O.**, and Zhan, Q. How Does Pre-trained Wav2Vec 2.0 Perform on Domain Shifted ASR? An Extensive Benchmark on Air Traffic Control Communications. In *2022 IEEE Spoken Language Technology Workshop, SLT 2022*, pages 205–212. Doha, Qatar, 09-12 Jan, 2023. `https://doi.org/10.1109/SLT54892.2023.10022724`

✍ Zuluaga-Gómez, J. P., Sarfjoo, S. S., Prasad, A., Nigmatulina, I., Motlíček, P., **Ohneiser, O.**, and Helmke, H. BERTraffic: BERT-based Joint Speaker Role and Speaker Change Detection for Air Traffic Control Communications. In *2022 IEEE Spoken Language Technology Workshop, SLT 2022*, pages 633–640. Doha, Qatar, 09-12 Jan, 2023. `https://doi.org/10.1109/SLT54892.2023.10022718`

✈ **Ohneiser, O.**, Helmke, H., Shetty, S., Kleinert, M., Ehr, H., Balogh, G., Tønnesen, A., Rinaldi, W., Mansi, S., Piazzolla, G., Murauskas, S., Pagirys, T., Kis-Pál, G., Horváth, V., Kling, F., and Usanovic, H. Understanding Tower Controller Communication for Support in Air Traffic Control Displays. In *12th SESAR Innovation Days, SID 2022*. Budapest, Hungary, 05-08 Dec, 2022. `https://s.dlr.de/TztJt`. This paper compares the results of three speech recognition and understanding validation exercises from the SESAR2020 solution *PJ.05-W2-97-ASR*. The habilitand was responsible for the relevant ASRU parts of the DLR exercise when conducting the human-in-the-loop simulation trials within DLR's Air Traffic Validation Center. He supported

the conceptualization and implementation of ATC concept extraction together with the primary programming activities from Helmke, Shetty, and Kleinert, as well as its testing and integration mainly done by Ehr. Murauskas, Pagirys, and Usanovic provided expertise from the operational air navigation service provider side to the DLR exercise. The habilitand did the complete data analysis and reporting including evaluation of ASRU metrics of the DLR exercise on his own. Balogh, Tønnesen, Rinaldi, Mansi, Piazzolla, Kis-Pál, Horváth, and Kling were responsible for the two other exercises and their data analysis that made up roughly half the effort of the technological and evaluation work for the content of this paper. The paper has 70 % habilitand authorship.

✍ Helmke, H., Ondřej, K., Shetty, S., Arilíusson, H., Simiganoschi, T. S., Kleinert, M., **Ohneiser, O.**, Ehr, H., and Zuluaga-Gómez, J. P. Readback Error Detection by Automatic Speech Recognition and Understanding – Results of HAAWAII project for Isavia's Enroute Airspace. In *12th SESAR Innovation Days, SID 2022*. Budapest, Hungary, 05-08 Dec, 2022. `https://s.dlr.de/9NrZG`

✍ Kleinert, M., Shetty, S., Helmke, H., **Ohneiser, O.**, Wiese, H., Maier, M., Schacht, S., Nigmatulina, I., Sarfjoo, S. S., and Motlíček, P. Apron Controller Support by Integration of Automatic Speech Recognition with an Advanced Surface Movement Guidance and Control System. In *12th SESAR Innovation Days, SID 2022*. Budapest, Hungary, 05-08 Dec, 2022. `https://s.dlr.de/EPpjr`

✍ Prasad, A., Zuluaga-Gómez, J. P., Motlíček, P., Sarfjoo, S., Nigmatulina, I., **Ohneiser, O.**, and Helmke, H. Grammar Based Speaker Role Identification for Air Traffic Control Speech Recognition. In *12th SESAR Innovation Days, SID 2022*. Budapest, Hungary, 05-08 Dec, 2022. `https://s.dlr.de/VHwPx`

✍ Shetty, S., Helmke, H., Kleinert, M., and **Ohneiser, O.** Early Callsign Highlighting using Automatic Speech Recognition to Reduce Air Traffic Controller Workload. In *International Conference on Applied Human Factors and Ergonomics, volume 60 of AHFE 2022*, pages 584–592. New York, NY, USA (hybrid), 24-28 Jul, 2022. `https://doi.org/10.54941/ahfe1002493`

✍ Helmke, H., Shetty, S., Kleinert, M., **Ohneiser, O.**, Prasad, A., Motlíček, P., Černá, A., and Windisch, C. Measuring Speech Recognition And Understanding Performance in Air Traffic Control Domain Beyond Word Error Rates. In *11th SESAR Innovation Days, SID 2021*. Virtual, 07-09 Dec, 2021. `https://s.dlr.de/Rt8dv`

✍ Kleinert, M., Venkatarathinam, N., Helmke, H., **Ohneiser, O.**, Strake, M., and Fingscheidt, T. Easy Adaptation of Speech Recognition to Different Air Traffic Control Environments using the DeepSpeech Engine. In *11th SESAR Innovation Days, SID 2021*. Virtual, 07-09 Dec, 2021. `https://s.dlr.de/hAG7U`

✍ Kleinert, M., Helmke, H., Shetty, S., **Ohneiser, O.**, Ehr, H., Prasad, A., Motlíček, P., and Harfmann, J. Automated Interpretation of Air Traffic Control Communication: The Journey from Spoken Words to a Deeper Understanding of the Meaning. In *IEEE/AIAA 40th Digital Avionics Systems Conference, DASC 2021*. San Antonio, TX, USA, 03-07 Oct, 2021. `https://doi.org/10.1109/DASC52595.2021.9594387`

✍ Helmke, H., Kleinert, M., Shetty, S., **Ohneiser, O.**, Ehr, H., Arilíusson, H., Simiganoschi, T. S., Prasad, A., Motlíček, P., Veselý, K., Ondřej, K., Smrz, P., Harfmann, J., and Windisch, C. Readback Error Detection by Automatic Speech Recognition to Increase ATM Safety. In *14th*

USA/Europe Air Traffic Management Research and Development Seminar, ATM 2021. Virtual, 20-24 Sep, 2021. https://s.dlr.de/LtMbB.

✈ Rataj, J., **Ohneiser, O.**, Marin, G., and Postaru, R. Attention: Target and Actual – The Controller Focus. In *32nd Congress of the International Council of the Aeronautical Sciences, ICAS 2021.* Shanghai, China (hybrid), 06-10 Sep, 2021. https://s.dlr.de/LYw7z reviews the basic considerations for attention guidance in air trafic control; 69 % authorship

✍ Helmke, H., Shetty, S., Kleinert, M., **Ohneiser, O.**, Prasad, A., Motlíček, P., Černá, A., and Windisch, C. How to Measure Speech Recognition Performance in the Air Traffic Control Domain? The Word Error Rate is only half of the truth. In *Interspeech 2021 Satellite Workshop 'Automatic Speech Recognition in Air Traffic Management', ASR-ATM 2021.* Brno, Czechia (hybrid), 30 Aug, 2021. https://s.dlr.de/3Ydz6

✍ Prasad, A., Zuluaga-Gómez, J. P., Motlíček, P., **Ohneiser, O.**, Helmke, H., Sarfjoo, S., and Nigmatulina, I. Grammar Based Identification Of Speaker Role For Improving ATCO And Pilot ASR. In *Interspeech 2021 Satellite Workshop 'Automatic Speech Recognition in Air Traffic Management', ASR-ATM 2021.* Brno, Czechia (hybrid), 30 Aug, 2021. https://s.dlr.de/v1SDN

✍ Rataj, J., Helmke, H., and **Ohneiser, O.** AcListant with Continuous Learning: Speech Recognition in Air Traffic Control, In Electronic Navigation Research Institute, editor, volume 731 of *Lecture Notes in Electrical Engineering: Air Traffic Management and Systems IV*, pages 93–109. Springer, Singapore, 2021. https://doi.org/10.1007/978-981-33-4669-7_6

✍ Helmke, H., Kleinert, M., **Ohneiser, O.**, Ehr, H., and Shetty, S. Machine Learning of Air Traffic Controller Command Extraction Models for Speech Recognition Applications. In *IEEE/AIAA 39th Digital Avionics Systems Conference, DASC 2020.* Virtual, 11-16 Oct, 2020. https://doi.org/10.1109/DASC50938.2020.9256484

✈ **Ohneiser, O.**, Helmke, H., Kleinert, M., Siol, G., Ehr, H., Hobein, S., Predescu, A.-V., and Bauer, J. Tower Controller Command Prediction for Future Speech Recognition Applications. In *9th SESAR Innovation Days, SID 2019.* Athens, Greece, 02-05 Dec, 2019. https://s.dlr.de/OaEPQ. This paper was **Best Presentation Poll** session winner. It outlines the command prediction methodology for speech recognition and understanding in a multiple remote tower environment. The habilitand was responsible for the relevant ASRU parts when conducting the human-in-the-loop simulation trials within DLR's Air Traffic Validation Center. He supported the conceptualization and implementation of air traffic control (ATC) concept prediction together with the primary programming activities from Helmke, Kleinert, and Siol as well as its testing and integration mainly done by Ehr and Hobein. The habilitand did the complete data analysis and reporting including evaluation of ASRU metrics on his own. He organized the transcription and annotation work for ATC utterances – doing a huge portion on his own – and supervised the bachelor students Predescu and Bauer. The paper has 80 % habilitand authorship. The paper acceptance rate of the conference was 53.1 %.

✍ Kleinert, M., Helmke, H., Moos, S., Hlousek, P., Windisch, C., **Ohneiser, O.**, Ehr, H., and Labreuil, A. Reducing Controller Workload by Automatic Speech Recognition Assisted Radar Label Maintenance. In *9th SESAR Innovation Days, SID 2019.* Athens, Greece, 02-05 Dec, 2019. https://s.dlr.de/P3TMA

✍ Rataj, J., Helmke, H., and **Ohneiser, O.** AcListant with Continuous Learning: Speech Recognition in Air Traffic Control. In *ENRI International Workshop on ATM/CNS, EIWAC 2019*. Tokyo, Japan, 29-31 Oct, 2019. https://s.dlr.de/3XHDL

✍ Di Flumeri, G., De Crescenzio, F., Berberian, B., **Ohneiser, O.**, Kramer, J., Aricò, P., Borghini, G., Babiloni, F., Bagassi, S., and Piastra, S. Brain-Computer Interface-Based Adaptive Automation to Prevent Out-Of-The-Loop Phenomenon in Air Traffic Controllers Dealing With Highly Automated Systems. ***Frontiers in Human Neuroscience***, 13, 2019. https://doi.org/10.3389/fnhum.2019.00296

✍ Helmke, H., Slotty, M., Poiger, M., Herrer, D. F., **Ohneiser, O.**, Vink, N., Černá, A., Hartikainen, P., Josefsson, B., Langr, D., Lasheras, R. G., Marin, G., Mevatne, O. G., Moos, S., Nilsson, M. N., and Pérez, M. B. Ontology for Transcription of ATC Speech Commands of SESAR 2020 Solution PJ.16-04. In *IEEE/AIAA 37th Digital Avionics Systems Conference, DASC 2018*. London, UK, 23-27 Sep, 2018. https://doi.org/10.1109/DASC.2018.8569238

✈ **Ohneiser, O.**, Jauer, M.-L., Gürlük, H., and Springborn, H. Attention Guidance Prototype for a Sectorless Air Traffic Management Controller Working Position. In *Deutscher Luft- und Raumfahrtkongress, DLRK 2018*. Friedrichshafen, Germany, 04-06 Sep, 2018. https://doi.org/10.25967/480189 describes the setup and basic functionality of an attention guidance prototype in ATC; 98 % authorship

✈ **Ohneiser, O.**, De Crescenzio, F., Di Flumeri, G., Kraemer, J., Berberian, B., Bagassi, S., Sciaraffa, N., Aricò, P., Borghini, G., and Babiloni, F. Experimental Simulation Set-Up for Validating Out-Of-The-Loop Mitigation when Monitoring High Levels of Automation in Air Traffic Control. In *International Conference on Air Traffic Management and Aviation, ICATMA 2018*. Lisbon, Portugal, 16-17 Apr, 2018. https://doi.org/10.5281/zenodo.1316361 describes the validation setup for influencing controller's vigilance and attention; this paper received a **Best Paper Award**; 90 % authorship

✈ Berberian, B., **Ohneiser, O.**, De Crescenzio, F., Babiloni, F., Di Flumeri, G., and Hasselberg, A. MINIMA Project: Detecting and Mitigating the Negative Impact of Automation. In D. Harris, editor, Engineering Psychology and Cognitive Ergonomics: Performance, Emotion and Situation Awareness, *14th International Conference, EPCE 2017, Held as Part of HCI International 2017*, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part 1, pages 87–105. Springer International Publishing, Cham, Switzerland, 2017. https://doi.org/10.1007/978-3-319-58472-0_8 outlines a vigilance and attention controller for ATC tasks; 20 % authorship

✈ Helmke, H., **Ohneiser, O.**, Buxbaum, J., and Kern, C. Increasing ATM Efficiency with Assistant Based Speech Recognition. In *12th USA/Europe Air Traffic Management Research and Development Seminar, ATM 2017*. Seattle, WA, USA, 27-30 Jun, 2017. https://s.dlr.de/Vm1Eg presents the results on a study for radar label maintenance; this paper received the **Best Paper Award** for "Human Factors" track; 20 % authorship

✈ Helmke, H., **Ohneiser, O.**, Mühlhausen, T., and Wies, M. Reducing Controller Workload with Automatic Speech Recognition. In *IEEE/AIAA 35th Digital Avionics Systems Conference, DASC 2016*. Sacramento, CA, USA, 25-29 Sep, 2016. https://doi.org/10.1109/DASC.2016.7778024 shows the benefits of automatic speech recognition for an arrival management system; this

paper received **Best Paper Award** for "Interaction Methods and Devices" session as well as "Human Factors" track; 20 % authorship

- ✈ Ahlstrom, U., **Ohneiser, O.**, and Caddigan, E. Portable Weather Applications for General Aviation Pilots. ***Human Factors***, 58(6):864-885, 2016. https://doi.org/10.1177/0018720816641783 analyzes pilot's reaction on visualized weather aspects with functional near-infrared spectroscopy methods; 30 % authorship
- ✍ Ahlstrom, U., Caddigan, E., Schulz, K., **Ohneiser, O.**, Bastholm, R., and Dworsky, M. The Effect of Weather State-change Notifications on General Aviation Pilots' Behavior, Cognitive Engagement, and Weather Situation Awareness. Tech. Rep. DOT/FAA/TC-15/64, Federal Aviation Administration, Atlantic City, NJ, USA, 2015. https://s.dlr.de/KoZ7y

Earlier works of the habilitand are sketched below for completeness. Ealy investigations on assistant-based speech recognition (ABSR) in laboratory environments have reported on workload reduction for ATCos through ABSR support [Helmke et al., 2015], [Gürlük et al., 2015], usage of contextual knowledge to improve ABSR accuracy [Schmidt et al., 2014], as well as tools to support transcription and annotation of ATCo utterances [Ohneiser et al., 2014b]. Initial investigations on visual display aids for approach ATCos to support timely commands in the TMA are presented in [Ohneiser et al., 2015] and [Ohneiser, 2012] for turn-to-base commands, [Förster et al., 2011] for dynamic re-routing with real-time information, and [Ohneiser and Beddig, 2013] for conformance monitoring of aircraft trajectories. [Schildt et al., 2013] deals with system communication. The topic of migration tolerance, i.e., to introduce display changes for controllers in small steps is handled in [Ohneiser, 2017], [Ohneiser, 2016b], [Ohneiser, 2016a], [Ohneiser et al., 2014a], and [Ohneiser and Gürlük, 2013]. Those earlier works do not contribute to the cumulative habilitation thesis unlike the 52 articles mentioned above and especially unlike the eleven articles in first authorship of the habilitand that are reprinted in the following.

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Robust Command Recognition for Lithuanian Air Traffic Control Tower Utterances

Oliver Ohneiser[1], *Saeed Sarfjoo*[2], *Hartmut Helmke*[1],
Shruthi Shetty[1], *Petr Motlicek*[2], *Matthias Kleinert*[1], *Heiko Ehr*[1], *Šarūnas Murauskas*[3]

[1]German Aerospace Center (DLR), Institute of Flight Guidance, Braunschweig, Germany
[2]Idiap Research Institute, Martigny, Switzerland
[3]State Enterprise ”Oro navigacija“ (ON), Air Navigation Service Provider of Lithuania, Lithuania
{firstname.lastname}@dlr.de, {firstname.lastname}@idiap.ch, murauskas.s@ans.lt

Abstract

The maturity of automatic speech recognition (ASR) systems at controller working positions is currently a highly relevant technological topic in air traffic control (ATC). However, ATC service providers are less interested in pure word error rate (WER). They want to see benefits of ASR applications for ATC. Such applications transform recognized word sequences into semantic meanings, i.e., a number of related concepts such as callsign, type, value, unit, etc., which are combined to form commands. Digitized concepts or recognized commands can enter ATC systems based on an ontology for utterance annotation agreed between European ATC stakeholders. Command recognition (CR) has already been performed in approach control. However, spoken utterances of tower controllers are longer, include more free speech, and contain other command types than in approach. An automatic CR rate of 95.8% is achievable on perfect word recognition, i.e., manually transcribed audio recordings (gold transcriptions), taken from Lithuanian controllers in a multiple remote tower environment. This paper presents CR results for various speech-to-text models with different WERs on tower utterances. Although WERs were around 9%, we achieve CR rates of 85%. CR rates only slightly decrease with higher WERs, which enables to bring ASR applications closer to operational ATC environment.

Index Terms: speech recognition, speech understanding, command recognition rate, air traffic control, tower utterances

1. Introduction

Automatic speech recognition (ASR) in air traffic control (ATC) existed decades ago [1],[2]. However, it got more powerful in the last decade due to improved computing power for model training and accelerating digitization in the ATC domain. Normally, the step that follows ASR is language understanding – in ATC, also called as spoken instruction understanding [3]. Different projects have shown possible applications [4] such as runway incursion detection [5], decision support input [6], radar label maintenance [7],[8], etc., which ultimately results in benefits such as workload reduction for air traffic controllers [9]. For language understanding, multiple words are analyzed to extract the semantic meaning (concept extraction) of utterances, which includes the extractions of ATC concepts, such as callsigns, command types, command values, units, conditions, etc. The extraction of these ATC concepts is supported by machine learning algorithms [10]. The ATC concepts can be annotated by applying the rules of an ontology, agreed by 14 European air navigation service and system providers [11]. Concept extraction has already been applied to ATC utterances from the approach domain and to manually transcribed (gold) ATC utterances from the tower domain [12]. Our approach in this paper is among the first applications to apply command recognition on partly erroneous recognized speech text from the tower domain[1]. With this approach, we investigate the effect of using unsupervised data for training a robust acoustic model for the ATC domain. The improvement of word error rate (WER) and the partly dependent enhancement of command recognition rate (CRR) are important steps to achieve higher technology readiness levels because the ATC end users are interested in low error rates on semantic level. The next section presents related work on language modeling, transcription rules, and the annotation ontology. Section 3 describes the ATC concept extraction to recognize commands as well as trials for data acquisition and analysis. The recognition experiments and results are shown in section 4. Section 5 concludes and gives an outlook on future work.

2. Related Work

2.1. Language Modeling

Several LM adaptation or interpolation techniques were proposed for mapping the language model (LM) to the specific domain, e.g., linear interpolation, Bayesian interpolation and count merging. Bayesian interpolation was introduced in [13]. [14] and [15] showed that count merging with two data sources is a specific style of maximizing a posteriori (MAP) adaptation. [16] shows the theoretical connections between the mentioned LM interpolation techniques.

2.2. Transcription Rules and Annotation Ontology

Different transcription rules for ATC utterances have been defined and used for existing audio corpora [17]-[20] such as:

- Spelled letters – not pronounced using the International Civil Aviation Organization (ICAO) alphabet such as alfa, bravo, etc. – e.g., “~k~l~m”/“KLM”/“K L M”,
- Truncated/broken word parts, e.g., “luf=”/“luf*”/“luf-” if “lufthansa” was not uttered fully till the end,
- Non-understandable words (“[unk]” / “[UNKNOWN]”) and human noise/thinking loud (“[hes]” / “[HNOISE]”),
- Non-English words, e.g., “<FL></FL>” / “[NE][/NE]”.

[1] For funding information please refer to [38],[11],[30].

http://dx.doi.org/10.21437/Interspeech.2021-935

Also, for the annotation of semantic meanings of the ATC transcriptions different ontologies or rule sets exist. An early ontology developed by NATS for the terminal environment comprised of callsign, standard type, non-standard type, value, and type unit [21]. Similarly, the ontology introduced by the *AcListant®* project [22] proposed to use four different elements: callsign, type, value, and unit of a command [23],[24]. A further approach suggested to use keywords like callsign, flightlevel, altimeter for the corresponding values [25]. Another proposition was to have ten class labels for annotation of word sequences such as callsign, fix, number, etc. [26],[27]. The *AcListant®* ontology was enhanced during the *MALORCA* project [28] in which various command types for "information", "reports", and "expects" were added next to conditional clearances [29]. This ontology has been further enhanced for en-route and tower commands during the *CWP HMI* project [11]. Furthermore, the ontology with more than 100 different command types has been agreed between major European partners from the air traffic management (ATM) domain including air navigation service providers, ATM system providers, and the coordinating partner DLR. The *HAAWAII* project [30] further enhanced the ontology for pilot utterances including their requests and reports. Also, other European ASR projects such as *HMI Interaction modes for Approach control*, *HMI Interaction Modes for Airport Tower*, and *Safety and Artificial Intelligence Speech Recognition* continuously contribute to the improvement of the ontology. The global scheme for each instruction to annotate ATC utterances is shown in Figure 1. Each ATC utterance can contain multiple instructions.

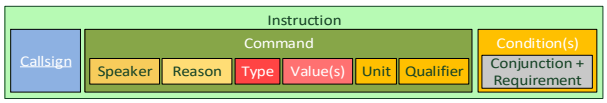


Figure 1: *Elements of an air traffic control instruction including the ATC concepts 'callsign', 'command' with sub-elements, and optional 'conditions'.*

The callsign is a mandatory element for each instruction and might be NO_CALLSIGN if not uttered. This is followed by a mandatory command and may be followed by optional conditions. The command again can have a speaker (PILOT or empty for default air traffic controller), a reason (REPORTING, REQUEST or empty), a type (REDUCE, DESCEND, VACATE, CONTACT_FREQUENCY, CLEARED VIA, etc.), one or multiple values ("200", "A B D1", "118.300", etc.), a unit (FL, ft, kt, none, etc.), and a qualifier (RIGHT, OR_LESS, etc.). The conditions have a conjunction and a requirement ("UNTIL 4 NM FINAL", "WHEN AIRBORNE", etc.). An ontology for annotations supports different purposes. It is needed as an interface to enable interoperability of different ASR applications with ATC systems. It is also necessary for evaluating automatically recognized commands against manual (gold) annotations. The name "command recognition rate" (CRR), taken from [6] has historical reasons. According to Figure 1, the term "instruction error rate" would be correct. For the calculation of the CRR, each command, for example consisting of the ATC concepts callsign, type, value, qualifier, condition, etc. is considered as one (big) word to compute the Levenshtein distance [31]. This means that a recognized command is correct only if all concepts (command parts) are correct, i.e., "DLH7HT HEADING 360 LEFT" and "DLH7HT HEADING 360 none" are not equal and would be counted as a full command recognition error. The CRR is defined as the number of controller commands correctly recognized by the ASR (and not rejected due to implausibility) divided by the total number of commands given or in other words: the percentage of given commands correctly shown on the controllers' display. An example transcription and resulting annotation is given in Table 1. A configuration file defines allowed values for taxiways, holding points, etc. to map "holding point three four to "HP_34" here.

Table 1: *Transcription and annotation example.*

Transcription	Annotation
[NE French] bonjour [/NE] hotel alfa charlie india zulu [unk] taxi to holding point three four via taxiway [hes] alfa runway in use three four and nex*	HACIZ TAXI TO HP_34 HACIZ TAXI VIA A HACIZ INFORMATION ACTIVE_RWY RW34

3. ATC command recognition and remote tower simulation trials

3.1. ATC concept extraction for command recognition

The command recognition algorithm consists of several steps, where different ATC concepts are extracted iteratively and put into relation to recognize them as single or multiple commands of an utterance (for more details see [10]). First, we try to extract a callsign from an ATC utterance by considering the callsign information from the available surveillance data (for controller utterances, only the first words are considered). Then, keywords or keyword sequences are extracted which initiate a command type. This step includes the extraction of a command type followed by value(s), unit, qualifier, etc. if applicable. Afterwards, we look for unmatched words in the complete utterance that correspond to non-extracted ATC concepts and we also look for command hints such as "feet" being used in an ALTITUDE command. We then search again for callsigns in the remaining unmatched words and then, we finally try to extract commands from unmatched numbers in the utterance. The above example transcription from Table 1 is reused for illustrating the algorithm here. The concept extraction model searches for the presence of any of the available predicted callsigns, e.g., AFR27C, DLH9LX, HACIZ (from surveillance data) in the utterance. The latter callsign matches here. Then, the keywords "taxi to" and the value keywords "holding point three four" as well as "via" and "taxiway alfa" lead to extraction of "TAXI TO HP_34" and "TAXI VIA A", respectively. The words "runway in use" and "three four" are extracted as "INFORMATION ACTIVE_RWY RW34". All other words ("bonjour", "[unk]", "[hes]", "and nex*") are not relevant for the command recognition algorithm example.

3.2. Trials for data recording and tower considerations

In March and December 2018 multiple remote tower trials with Lithuanian controllers from Oro Navigacija speaking accented English took place in DLR TowerLab in Braunschweig, Germany. These trials were conducted as human-in-the-loop simulations in the course of the project *CWP HMI-ASR* [32]. One controller was responsible for all the traffic from three international airports (named Vilnius (EYVI), Kaunas (EYKA), and Palanga (EYPA)) at the same time. In total, 41.4 hours with silence between different utterances aligned with radar data from the air traffic control simulation have been recorded. After deleting the inter-

utterance silence, 6.86 hours of pure speech in 3,919 audio files remain out of the trials, but only slightly more than 50% of the files have been manually (gold) transcribed and annotated. The simulation pilot utterances were not considered – only those of six tower controllers. The amount and division of labelled offline ASR data is shown in Table 2.

Table 2: *Description of transcribed audio data sets.*

Set name	# files	Duration (hours)	Average duration (sec)
all	1,993	3.6	6.6
adapt	1,399	2.6	6.8
test	594	1.0	6.1

The average duration of an utterance in this (Lithuanian) multiple remote tower environment is 6.6 seconds. This is significantly longer than for Vienna approach (4.4s) or Prague approach (5.1s) in real-life data from the *MALORCA* project. Furthermore, controllers instructed roughly 2.7 commands per utterance. Again, this is much more than 1.6 and 1.7 commands per utterance from Prague and Vienna approach from *CWP HMI-ASR* simulation runs, respectively. Also, the variation of words, i.e., the total number of different words used divided by the total number of used words is higher. The Lithuanian tower controllers used 560 different words (in total 32,484) compared to 196 different words (in total 31,436) for Vienna approach and 218 different words (in total 47,426) for Prague approach in *CWP HMI-ASR* simulation runs, respectively. Higher variation shows more free speech due to visual flight rules (VFR) traffic, e.g., vague and difficult to analyze commands like “fly heading north” would probably not be given to traffic following instrument flight rules (IFR). In addition, the number of different command types for tower ATC as modeled in the ontology is larger than for approach. Finally, the amount of available speech data for the tower domain is much less, because it is harder to record them as compared to the very high frequency receivers for approach ATC speech. All above-explained characteristics make it more challenging to automatically recognize tower commands.

4. Experiments and Results

4.1. Models and different error/recognition rates

All ASR experiments are conducted using the Kaldi speech recognition toolkit. The speech recognition acoustic model was trained on 195 hours of data from seven datasets in the ATC domain (model *Supervised baseline*). Description of the training datasets can be found in [33]. Hybrid deep neural network (NN)–hidden Markov model (DNN-HMM) with lattice-free maximum mutual information (LF-MMI) loss function was trained using alignment from Gaussian mixture models (GMM) HMM. State-of-the-art ASR chain recipes with convolutional NN–factorized time-delay NN (CNN-TDNNF) architecture from Kaldi toolkit was used for training. 4-gram[1] LM in ARPA format was trained using the same training set. For LM adaptation to the Lithuanian ATC domain, linear interpolation between the general LM and the LM from adaptation set with 0.8 and 0.2 weights was performed (model + *LM-mix*) due to the limited dataset. For improving the ASR accuracy and increasing the noise robustness of the trained model, we trained a *semi-supervised* model using 400 hours of unsupervised data from LiveATC dataset [34]. Incremental method was used for training the *semi-supervised* model [35]. We divided the unlabeled data to four 100 hours subsets. Starting from one subset, in each training iteration we added one unseen subset to the previous subsets. We extracted 86 out-of-vocabulary words including waypoints, airlines, and some local terms from the transcribed data. These words were added to the decoding graph for all experiments. The WER of the trained ASR models on test set is shown in Table 3. LM interpolation improved the WER on the test set by 9%. Effective mapping of LM using the dataset with similar phraseology pattern is one main reason for observing this improvement. In addition, including unsupervised data from ATC domain improved the ASR accuracy by 3%. It shows more robustness of the *semi-supervised* acoustic model w.r.t. the *supervised* model. Analysis on the recognition errors shows the majority of errors in the *supervised* baseline model are because of deviation of the main LM w.r.t. the in-domain data. *Semi-supervised* model reduced the recognition errors of the noisy segments and majority of the substitution errors are words with similarity in the pronunciation, e.g., "flight" and "sight".

Table 3: *Applied models (with 4-gram LM), word error rate (WER), command recognition rate (CRR), command recognition error rate (CER), and callsign recognition rate (CaRR) for tower utterances from Lithuanian controllers on test set in [%].*

Model	WER	CRR	CER	CaRR
Supervised baseline	20.8	59.0	14.1	79.5
+ LM-mix	11.8	78.4	8.2	93.8
+ Semi-supervised	8.8	84.3	7.7	96.3

The CRR in Table 3 is calculated on annotations. Thus, it can only loosely be compared to the sentence accuracy calculated on transcriptions – 1 minus sentence error rate (SER) – being used to evaluate ASR applications outside ATC domain. The CRR with gold transcription input, where a WER of 0% is assumed – compared to gold annotations is 95.8% with a command recognition error rate (CER) of 2.7%. From Table 3 we see that despite the high WER of almost 21%, a CRR of 59% is reached. With improved models, the WER decreases to roughly 12% and 9% which leads to CRRs of 78% and even 84%, respectively. As an example, the best and worst CRR per speaker were less than 5% different from the reported average using the *semi-supervised* model. A lower CRR does not really affect the workload of a controller. If there is no support by the ASR system in feeding recognized commands into the ATC system, the situation is comparable to today. Of course, higher CRRs reduce controller workload. However, if the CER increases, this results in additional workload for the controller to first recognize the error, then to delete the wrong result, and then to manually correct the wrong automatic input. A CER of 7.7% means that each thirteenth command needs to be corrected by the controller. With a higher WER of 20.8%, the number of errors is almost two-and-a-half times higher than 8.8%. The CER, on the other hand, also increases with increased WER but only from 7.7% to 14% (less than twice the number of errors). Using the baseline model, each seventh command would need to be corrected by the controller. The reason is that high WERs may lead to recognizing nothing at all on concept level, i.e., the recognized concept is rejected, because, e.g., a heading command of 733 degrees is extracted.

[1] 3-gram LM WERs were 0.2-0.6% higher than 4-gram LM WERs (only the latter reported in this paper) for the models.

The callsign recognition rate (CaRR) is also listed in Table 3. It is the most important ATC concept and can heavily influence the CRR, because the callsign is part of each command. From the perspective of an ATC application, the recognized callsign should be highlighted in the controller display to ease identifying the current communicating aircraft, and to speed up checking and correcting of recognized commands. The reference CaRR, i.e., automatic callsign extraction compared to the callsigns from gold annotation is 99%. The CaRR for the three model achieves roughly 80% to over 96%. Hence, the recognition rates for callsigns are much better than for commands in general. This is due to the usage of Assistant Based Speech Recognition (ABSR), first described in [6], which relies on using context information from the corresponding radar data.

4.2. Analysis of command recognition performance

Figure 2 shows the theoretical CERs if words from automatic transcriptions would be independent of each other given the three different WERs plus the perfect WER of 0%. It also presents the corresponding four achieved CRRs for the observed average number of six words per command.

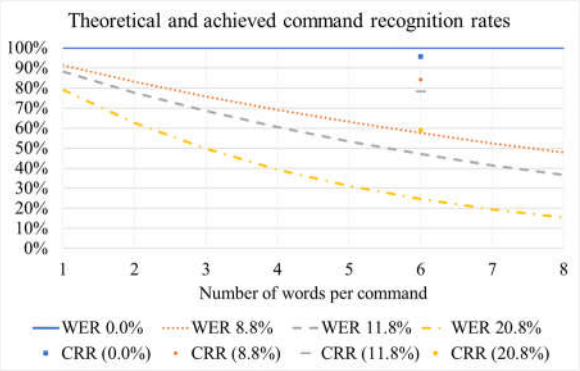


Figure 2: *Theoretical CRR for different number of words per command (sentence length) and measured CRR for average length of 6 words.*

From Figure 2 we see that with a WER of 0%, the CRR should be 100%. We, however, currently reach only 95.8% based on 68 different used ATC command types. This is on the one hand due to challenges with controller utterances being "far away" from ICAO phraseology rules [36], e.g., uttering "lufthansa two victor" if "DLH23W" is meant or instructing just the three words "two six zero", which can be a heading, speed, flight level, etc. [37]. The most recognition errors deal with command types TAXI, DIRECT_TO, and INFORMATION TRAFFIC/ACTIVE_RWY. On the other hand, the 5,367 gold annotations of the commands still contain some errors, i.e., the automatic annotation is already better than the manual annotation. However, for higher WERs, we achieve much better CRRs than theoretically achievable, if recognized words would be independent and word errors would be equally distributed. For example, a WER of 8.8% should enable a CRR of 58% for an average of 6 words per command, but we even observe above 84%. For a WER of 11.8%, we achieve a CRR of 78.4% compared to 47% based on independence assumption; and for a WER of 20.8%, we still achieve a CRR of 59% compared to 25% based on independence assumption. Hence, the WER only gives some hints to the performance of a speech recognition system in the ATC world. However, the CRR (or CER) is much more important for the end user and is more robust against higher WERs, i.e., achieve roughly 30% better recognition results than expectable due to the independence assumption.

Furthermore, it is more important to recognize longer words correctly than shorter words. If we replace each word in the speech recognition hypotheses files up to a length of 2,3,4,5,6 letters by "x", we see a steep decrease of command recognition rates when replacing words with up to three letters as shown in Table 4. If we replace the words with up to three letters, it means that we also replace the words with one and two letters. However, it is also connected to the number of replaced words, i.e., we roughly replace 0.1% (1), 5% (2), 25% (3), 51% (4), and 73% (5) of words.

Table 4: *CRRs in [%] in case of replaced words up to the listed number of letters per word (1,2,3,4,5).*

Model	1	2	3	4	5
Supervised baseline	59.0	51.6	17.1	3.0	0.4
+ LM-mix	78.4	69.6	23.6	4.9	0.6
+ Semi-supervised	84.3	75.1	27.3	5.3	0.9

This trend can be explained with the importance of certain words (given their length and number of occurrence) for the command recognition process. If words such as "a" or "A" are missing (1), there is hardly any negative effect. If words such as "to", "in", "up", "by", "or" are missing (2), there is a slight decrease in recognition. However, if meaningful words – especially numbers – such as "one", "two", "six", "via", "QNH", "KLM" are missing (3), we see a dramatic decrease. When replacing even longer words (4) such as "zero", "four", "five", "nine", "feet", "taxi", "wind", "west", "east", "left" the recognition becomes hardly usable. It is completely unusable if even longer words such as "right", "descend", "vacate", "takeoff", "knots", "degrees", "lufthansa" are replaced.

5. Conclusions and future work

This paper applies ontology-based command recognition on automatic transcriptions from ATC tower utterances of Lithuanian controllers with different WERs. Compared to the approach environment, tower utterances are longer, have more speech variety, more command types, and less available training data, i.e., recognition of words and commands is more challenging than in the approach environment. The baseline speech recognition is developed based on approach data, the first speech recognition solution uses language model adaptation, the second solution performed a semi-supervised approach leading to the best WER with around 9%. The resulting command recognition rates have proven to be robust (slight decrease) even on higher WERs. With current LM models, CRRs of 85% are possible.

In future, for alleviating the lack of transcribed speech data in the tower domain, we will focus on semi-supervised acoustic model adaptation for improving the accuracy of the ASR system on specific accents. The project *HMI Interaction Modes for Airport Tower* [38] will investigate the effect of presenting command recognition output to tower controllers in a human-in-the-loop simulation. These multiple remote tower trials will be conducted in the first quarter of 2022 in DLR TowerLab with controllers from Lithuania, Austria, and Poland. Controllers will benefit from callsign highlighting, recognized and displayed ATC concepts / commands in ontology annotation format. The presented results are already a good starting point and would enable a workload reduction compared to manually entering all given commands.

6. References

[1] D. W. Connolly, "Voice Data Entry in Air Traffic Control," Tech. Rep. N93-72621, FAA, National Aviation Facilities Experimental Center, Atlantic City, NJ, USA, 1979.

[2] C. Hamel, D. Kotick, and M. Layton, "Microcomputer System Integration for Air Control Training," Special Report SR89-01, Naval Training Systems Center, Orlando, FL, USA, 1989.

[3] Y. Lin, "Spoken Instruction Understanding in Air Traffic Control: Challenge, Technique, and Application," Aerospace 8, No. 3: 65, 2021.

[4] J. Rataj, H. Helmke, and O. Ohneiser, "AcListant with Continuous Learning: Speech Recognition in Air Traffic Control," Air Traffic Management and Systems IV – Selected Papers of the 6th ENRI International Workshop on ATM/CNS (EIWAC2019), 6, Springer, 2021.

[5] S. Chen, H. D. Kopald, A. Elessawy, Z. Levonian, and R. M. Tarakan, "Speech inputs to surface safety logic systems," IEEE/AIAA 34th Digital Avionics Systems Conference (DASC), Prague, Czech Republic, 2015.

[6] H. Helmke, J. Rataj, T. Mühlhausen, O. Ohneiser, M. Kleinert, Y. Oualil, and M. Schulder, "Assistant-based speech recognition for ATM applications," 11th USA/Europe Air Traffic Management Research and Development Seminar, Lisbon, Portugal, 2015.

[7] H. Helmke, O. Ohneiser, J. Buxbaum, and C. Kern, "Increasing ATM efficiency with assistant-based speech recognition," 12th USA/Europe Air Traffic Management Research and Development Seminar, Seattle, WA, USA, 2017.

[8] M. Kleinert, H. Helmke, S. Moos, P Hlousek, C. Windisch, O. Ohneiser, H. Ehr, and A. Labreuil, "Reducing Controller Workload by Automatic Speech Recognition Assisted Radar Label Maintenance," 9th SESAR Innovation Days, Athens, Greece, 2019.

[9] H. Helmke, O. Ohneiser, T. Mühlhausen, and M. Wies, "Reducing controller workload with automatic speech recognition," IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 2016.

[10] H. Helmke, M. Kleinert, O. Ohneiser, H. Ehr, S. Shetty, "Machine Learning of Air Traffic Controller Command Extraction Models for Speech Recognition Applications," IEEE/AIAA 39th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 2020.

[11] H. Helmke, M. Slotty, M. Poiger, D. F. Herrer, O. Ohneiser et al., "Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ.16-04," IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), London, United Kingdom, 2018. European Union's grant agreement No. 734141.

[12] O. Ohneiser, H. Helmke, S. Shetty, M. Kleinert, H. Ehr, S. Murauskas, and T. Pagirys, "Prediction and Extraction of Tower Controller Commands for Speech Recognition Applications," Journal of Air Transport Management, Elsevier, accepted 28 May 2021, expected publication June 2021.

[13] M. Weintraub, Y. Aksu, S. Dharanipragada, S. Khudanpur, H. Ney, J. Prange, A. Stolcke, F. Jelinek, and E. Shriberg, "LM95 project report: Fast training and portability," Research Note 1, Center for Language and Speech Processing, Johns Hopkins University, Tech. Rep., 1996.

[14] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP'03), Vol. 1, pp. I-224–I-227, IEEE, 2003.

[15] M. Bacchiani, M. Riley, B. Roark, and R. Sproat, "MAP adaptation of stochastic grammars," Computer speech & language, 20(1), pp. 41–68, 2006.

[16] E. Pusateri, C. Van Gysel, R. Botros, S. Badaskar, M. Hannemann, Y. Oualil, and I. Oparin, "Connecting and comparing language model interpolation techniques," Interspeech, Graz, Austria, 2019.

[17] K. Hofbauer and S. Petrik, "ATCOSIM Air Traffic Control Simulation Speech Corpus," Tech. Rep., TR TUG-SPSC-2007-11, Graz, Austria, 2008.

[18] E. Delpech, M. Laignelet, C. Pimm, C. Raynal, M. Trzos, A. Arnold, and D. Pronto, "A real-life, french-accented corpus of Air Traffic Control communications," Proc. LREC, Miyazaki, pp. 2866–2870, 2018.

[19] J. J. Godfrey, "Air Traffic Control Complete corpus," 1994, https://catalog.ldc.upenn.edu/LDC94S14A.

[20] S. Shetty, O. Ohneiser, F. Grezl, H. Helmke, and P. Motlicek, "Transcription and Annotation Handbook," HAAWAII deliverable D3.1, Braunschweig, Germany, 2020.

[21] D. Randall, "Direct Voice Input (DVI) Technology readiness and status introduction," Whitely, Fareham, UK, 2006.

[22] AcListant homepage: www.AcListant.de, AcListant = Active Listening Assistant, n.d.

[23] A. Schmidt, "Integrating situational context information into an online ASR system for Air Traffic Control," Master Thesis, Saarland University (UdS), Germany, 2014.

[24] Y. Oualil, M. Schulder, H. Helmke, A. Schmidt, and D. Klakow, "Real-time integration of dynamic context information for improving automatic speech recognition," Interspeech, Dresden, Germany, 2015.

[25] D. R. Johnson, V. I. Nenov, and G. Espinoza, "Automatic speech semantic recognition and verification in Air Traffic Control," IEEE/AIAA, 32rd Digital Avionics Systems Conference (DASC), East Syracuse, NY, USA, 2016.

[26] V. N. Nguyen and H. Holone, "N-best list re-ranking using syntactic score: A solution for improving speech recognition accuracy in Air Traffic Control," 16th Int. Conf. on Control, Automation and Systems (ICCAS 2016), Gyeongju, Korea, pp. 1309–1314, 2016.

[27] V. N. Nguyen and H. Holone, "N-best list re-ranking using syntactic relatedness and syntactic score: An approach for improving speech recognition accuracy in Air Traffic Control," 16th Int. Conf. on Control, Automation and Systems (ICCAS 2016), Gyeongju, Korea, pp. 1315–1319, 2016.

[28] MALORCA homepage: www.malorca-project.de, Machine Learning of Recognition Models for Controller Assistance, n.d.

[29] A. Srinivasamurthy, P. Motlicek, I. Himawan, G. Szaszák, Y. Oualil, and H. Helmke, "Semisupervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," Interspeech, Stockholm, Sweden, 2017.

[30] HAAWAII homepage: www.haawaii-project.de, Highly Automatic Air Traffic Controller Workstation with Artificial Intelligence Integration, n.d., This project has received funding from the SESAR Joint Undertaking under Grant Agreement No. 884287, under European Union's Horizon 2020 Research and Innovation programme. Idiap used funding parts for this work.

[31] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Soviet Physics -- Doklady 10.8, Feb. 1966.

[32] O. Ohneiser, H. Helmke, M. Kleinert, G. Siol, H. Ehr, S. Hobein, A.-V. Predescu, and J. Bauer, "Tower Controller Command Prediction for Future Speech Recognition Applications," 9th SESAR Innovation Days, Athens, Greece, 2019.

[33] J. Zuluaga-Gomez, P. Motlicek, Q. Zhan, K. Vesely, and R. Braun, "Automatic Speech Recognition Benchmark for Air-Traffic Communications," Interspeech, Shanghai, China, 2020.

[34] LiveATC-Homepage, https://www.liveatc.net/ , n.d.

[35] B. Khonglah, S. Madikeri, S. Dey, H. Bourlard, P. Motlicek, and J. Billa, "Incremental semi-supervised learning for multi-genre speech recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7419-7423, 2020.

[36] ICAO, "Doc 4444, Procedures for Air Navigation Services, Air Traffic Management," ICAO, Montréal, Canada, 2016.

[37] H. Said, M. Guillemette, J. Gillespie, C. Couchman, and R. Stilwell, "Pilots & Air Traffic Control Phraseology Study," International Air Transport Association, 2011.

[38] PJ.05-97-W2 SESAR2020 funded industrial research project under the European Union's grant agreement No. 874464, see also https://www.remote-tower.eu/wp/?page_id=888, n.d.

Journal of Air Transport Management 95 (2021) 102089

Contents lists available at ScienceDirect

Journal of Air Transport Management

journal homepage: www.elsevier.com/locate/jairtraman

Prediction and extraction of tower controller commands for speech recognition applications

Oliver Ohneiser [a,*], Hartmut Helmke [a], Shruthi Shetty [a], Matthias Kleinert [a], Heiko Ehr [a], Šarūnas Murauskas [b], Tomas Pagirys [b]

[a] German Aerospace Center (DLR), Institute of Flight Guidance, Lilienthalplatz 7, 38108, Braunschweig, Germany
[b] State Enterprise "Oro Navigacija" (ON), Air Navigation Service Provider of Lithuania, B. Karvelio str. 25, 02184, Vilnius, Lithuania

ARTICLE INFO

Keywords:
Air traffic controller
Multiple remote tower commands
Command prediction
Command extraction
Assistant based speech recognition

ABSTRACT

Air traffic controllers' (ATCos) workload often is a limiting factor for air traffic capacity. Thus, electronic support systems intend to reduce ATCos' workload. Automatic speech recognition can extract controller command elements from verbal clearances to deliver automatic input for air traffic control systems, thereby avoiding manual input. Assistant Based Speech Recognition (ABSR) with high command recognition rates and low error rates has proven to dramatically reduce ATCos' workload and increase capacity in approach scenarios. However, ABSR needs accurate hypotheses on expected commands and accurate extractions of command annotations from utterance transcriptions to achieve the required performance. Based on the experience of implementation for approach control, a hypotheses generator and a command extractor have been developed for speech recognition applications regarding tower control communication to face current and future challenges in the aerodrome environment. Three human-in-the-loop multiple remote tower simulation studies were performed with 16 ATCos from Hungary, Lithuania, and Finland at DLR Braunschweig from 2017 to 2019. Roughly 100 h of speech with corresponding radar data were recorded. Around 6000 speech utterances resulting in 16,000 commands have been manually transcribed and annotated. Some parts of the data have been used for training prediction models and command extraction algorithms. Other parts were used for evaluation of command prediction and command extraction. The automatic command extractor achieved a command extraction rate of 96.7%. The hypotheses generator showed operational feasibility with a sufficiently low command prediction error rate of 7.3%.

1. INTRODUCTION

Although temporary Corona crisis reduces the amount of air traffic, a growing number of worldwide flights every year is expected. This goes along with challenges regarding safety, efficiency, capacity, and environmental impact for air traffic management. A steadily increasing degree of automation and digitization currently seems to be the best method to face cost pressure and enhance air traffic management performance as also outlined by SESAR (SESAR Joint Undertaking, 2020) and NextGen (Federal Aviation Administration, 2019).

It can be assumed that digital controller pilot data link communications (CPDLC) will not completely and quickly replace analogue voice communication especially in the tower environment in the next decades. Hence, the transformation of such analogue speech data of air traffic controllers (ATCos) into the digital world is a valuable input for all following electronic air traffic control (ATC) systems. ATCos' workload is an important factor to increase such systems' overall performance.

ATCos issue clearances via voice and radio communication to aircraft pilots for controlling all relevant flights under responsibility. The flight crew is expected to confirm voice clearances by a readback or by acknowledging the information – this means instant feedback to the ATCo. For their effective operation, ATC systems need accurate data in a timely manner as well. The issued clearances are one of the relevant inputs for ATC systems. These inputs are done manually by the ATCo (1) in former times and still at some tower working positions on paper flight strips or (2) within an electronic system using a mouse or another input device through the interaction with an electronic flight strip or with an electronic flight label in case of a stripless system. However, these electronic input devices generate a higher, but measurable workload for the ATCo (Tobaruela et al., 2014).

The necessary information for ATC system input is doubled because it already exists within the voice clearance in analogue format.

* Corresponding author.
E-mail address: oliver.ohneiser@dlr.de (O. Ohneiser).

https://doi.org/10.1016/j.jairtraman.2021.102089
Received 7 September 2020; Received in revised form 2 March 2021; Accepted 28 May 2021

Automatic speech recognition (ASR) and understanding can support ATCos by extracting the semantic meanings of the issued clearances as *concepts* and automatically feeding the relevant ones to the digital ATC system. Hereinafter, the term *concept* is used instead of *abstraction*, *content*, *intent*, *interpretation*, *meaning*, *pattern*, *semantic*, *sense*, etc., because it fits best to the superset of radiotelephony utterances. *Concepts* are, e.g., callsigns, command types, QNH-values, squawk settings, and many more. Several concepts together are combined to a command. Throughout this paper we, therefore, use the terms *command recognition rate* and *command recognition error rate* to evaluate the performance of an ASR applied in the ATC environment. Detailed definitions are provided in (Helmke et al., 2015).

The digitized concepts can be used as input for further ATC support functionalities, thereby leading to reduced workload, manual input error prevention, and increased safety. However, this requires high reliability on automatic speech recognition (speech-to-text) and extraction of command elements (text-to-concepts). Hypotheses about the content of controller utterances support the speech recognition engine to choose from a reduced set of possible concepts.

Such controller command hypotheses can be derived with the support of an assistant system (e.g., an arrival manager) considering surveillance data (radar data and flight plans), meteorological data, airspace and airport layouts (Aeronautical Information Publication), active configurations, etc. A speech recognition system that is integrated with such an assistant is called an assistant based speech recognition (ABSR) system. ABSR has already proven to decrease the command recognition error rate and increase the command recognition rate for the approach area (Helmke et al., 2015).

One chain of effects resulting from high controller command recognition rates starts with a reduction of ATCo workload to enter clearances (Helmke et al., 2016), resulting in more timely and accurate commands. This, in turn, can already be a safety gain and can further lead to shorter flight routes and shorter flight times that go along with reduced fuel consumption and carbon dioxide emissions (Helmke et al., 2017). Furthermore, manual input errors, i.e., forgotten or wrong command information into the aircraft radar labels, can be reduced (Helmke et al., 2016). Also, the visualization of uttered clearance elements in a controller display for better awareness and further tracking of conforming aircraft trajectory changes can overcome controller-pilot communication problems (Skaltsas et al., 2013) and increase safety.

To achieve these possible benefits also in the aerodrome ATC environment, two software modules have been developed for usage in a later ABSR system for the tower environment. This comprises a tower command hypotheses generator to predict controller commands and a tower command extractor to convert text-to-concepts (annotation) after the speech recognition engine's speech-to-text conversion (transcription). Transcription is defined as the word-by-word equivalent of a verbal utterance, e.g., "good morning KLM three nine [hes] charlie altitude five thousand feet reduce speed two hundred knots or less and turn right heading zero seven zero i will ca*". In this example "[hes]" represents a hesitation such as "ah/hmm". The "*" symbol indicates that the (probable) word "call" has not been uttered fully. Annotation is defined as the machine-readable semantic contents (ATC concepts) of a verbal utterance like "KLM39C 5000 ft, 200 kt, 070 right".

A multiple remote tower simulation is used to compare the actual given annotated controller commands with automatically generated command hypotheses (predictions) and with automatically extracted command annotations obtained from manually transcribed utterances. It should be noted that there are different steps made in the ABSR process that all can be evaluated with their "error rates" on command level. *Command Prediction* means to generate hypotheses which commands the controller will give in the near future. *Command Recognition* means that the automatic annotation of commands is generated based on transcription. This can be compared to the parsing step in natural language processing. For recent ASR projects in ATC, the command recognition rates of automatic annotation are based on automatic transcription. For this paper, the automatic command recognition bases on correct (manual) transcription and is hereinafter referred to as *command extraction* (similar to (Chen et al., 2017)) for better emphasis.

This paper covers related work with respect to command prediction for automatic speech recognition and command extraction in section II. Section III outlines the concept for a tower command hypotheses generator. The human-in-the-loop study setup for data recording and some implications for machine learning are explained in section IV. Section V describes the command extraction algorithms as well as their recognition and error rates. Section VI presents the results regarding quality of command hypotheses. Chapter VII summarizes, concludes, and gives an outlook on future work.

2. Related work on speech recognition and controller command hypotheses

2.1. History of speech recognition in ATC

ASR systems convert spoken words into machine-useable digital data and thus serve as an alternative input modality. Today, voice recognition is used in various areas of human life such as navigation systems or smartphone applications. The first ASR systems for ATC were developed (Young et al., 1989a), (Young et al., 1989b) about three decades ago and integrated for ATC training (Hamel et al., 1989). Years later, this led to the possibility of replacing or reducing simulation pilots and to enhance simulator infrastructure (e.g., DLR (Schäfer, 2001), MITRE (Tarakan et al., 2008), FAA (FAA, 2012), and DFS (Ciupka, 2012)). ASR also supports safety improvements, e.g., for detecting closed runway incursions (Chen et al., 2015) or pilot readback errors (Chen et al., 2017), and to perform ATCo workload assessment (Cordero et al., 2012), (Cordero et al., 2013).

However, an ABSR system (Shore et al., 2012) can also significantly reduce ATCos' workload as shown in the projects AcListant® and AcListant®-Strips (Helmke et al., 2016). In addition, air traffic management efficiency can be increased with fuel savings of 50–65 L per flight (Helmke et al., 2017). DLR and its speech recognition partner Saarland University used KALDI as the ASR platform. DLR developed a hypotheses generator making predictions about the next possible controller commands (Helmke et al., 2015). For example, if an arriving aircraft is at FL 100, it is more likely that the ATCo will issue a descent to FL 80 than a climb to FL 140 or even a non-reasonable descent to FL 140. The most probable hypotheses are sent to the speech recognition engine to reduce its search space and improve recognition quality. With this, command recognition error rates below 1.7% were achieved in 2015 (Helmke et al., 2015).

One main issue to transfer ABSR from the laboratory to operational systems is the costs of deployment, because modern speech recognition models require manual adaptation to local requirements and environments (language accents, phraseology deviations, environmental constraints, etc.) (Rataj et al., 2021). AcListant® needed more than 1 Mio € for development and validation for Düsseldorf approach area.

The SESAR exploratory research project MALORCA (Machine Learning of Speech Recognition Models for Controller Assistance) proposes a general, cheap, and effective solution to automate this re-learning, adaptation and customization process by automatically learning local speech recognition and controller models from radar and speech data recordings (Kleinert et al., 2018a).

Command recognition error rates of 3.2% and 0.6% were achieved for Vienna and Prague approach, respectively in MALORCA (Helmke et al., 2019). Those low command recognition error rates were reached by using command hypotheses and plausibility checking components as they reduce the command recognition error rate by roughly 12% (Vienna) and 6% (Prague) (Kleinert et al., 2018a). Table 1 shows the command prediction error rates, i.e., the quality of predicted commands, for MALORCA and SESAR2020 industrial research project PJ.16-04-ASR CWP HMI (Controller Working Position Human Machine Interface). The

Table 1
Overview of command prediction error rates of former projects.

Name of ASR Project and Environment	Command Prediction Error Rate
MALORCA – Prague Approach	2.3%
MALORCA – Vienna Approach	3.2%
SESAR2020 PJ.16-04-ASR – Prague Approach	0.3%
SESAR2020 PJ.16-04-ASR – Vienna Approach	4.8%

command prediction error rate is defined as the number of given controller commands that were not forecasted divided by the number of all given controller commands. Or in other words, it is the number of actually given commands by the controller that are not part of the set of predicted commands divided by the total number of commands actually given by the controller.

Again, the use of an assistant system with command predictions in PJ.16-04-ASR dramatically decreases the command recognition error rate, whereas only slightly decreasing the command recognition rate (Kleinert et al., 2019).

2.2. Controller command annotation formats

A necessary step for the evaluation of command prediction accuracy is to extract the concepts of actual given commands. Thus, ATCo utterances need to be transcribed and annotated in a common format – manually, automatically or with a mixture of such methods. Annotation is defined as the extraction of the meaning of word sequences from utterance transcriptions. Therefore, a set of rules – an ontology – has been developed, which the annotations must conform to. In the AcListant® project (The project AcListant® (A), a first version of an ontology used by the ATC concept extraction module was created, which consists of four elements: 1) callsign, 2) command type, 3) command value, and 4) unit (Schmidt, 2014), (Oualil et al., 2015) with mandatory and optional elements. The example from the introduction "KLM39C 5000 ft, 200 kt, 070 right" is annotated as "KLM39C ALTITUDE 5000 ALT KLM39C REDUCE_OR_BELOW 200 KLM39C TURN_RIGHT_HEADING 070". More than 30 approach command types such as ALTITUDE, DESCEND, TURN_RIGHT_HEADING were supported.

The approach reached its limits in the MALORCA project (The projectA (Mach), (Kleinert et al., 2017), when it was extended to roughly four dozen different approach command types for command annotation for live traffic for Vienna and Prague approach (Kleinert et al., 2018b) also including departure and overflight traffic. More command types were needed (e.g., QNH, INFORMATION, REPORT_SPEED, EXPECT_-RUNWAY) and the necessity to handle conditional clearances occurred (Srinivasamurthy et al., 2017). For the tower environment, even controller command types used only on ground such as PUSHBACK, TAXI, and LINEUP need to be considered. A first proprietary model for a tower command ontology already existed within the tower flight data processing system (TFDPS) of the German air navigation service provider DFS. It defines roughly twenty different rule-based states that an aircraft can be in, e.g., FIRST_CONTACT, DOWNWIND, LOW_-APPROACH, TAXI_IN, TAXI_OUT, READY_FOR_DEPARTURE (Schier and Manske, 2015).

According to the above-described challenges, all major European air traffic management (ATM) system providers and European air navigation service providers agreed on a common enhanced ontology – with roughly 100 command types for all flight phases – suggested and coordinated by DLR to provide a common catalogue for en-route, approach, and tower voice commands (Helmke et al., 2018). It can be used for annotation of ATCo and pilot utterances as well as for command predictions. In the above example, the new annotation would be "KLM39C ALTITUDE 5000 ft KLM39C REDUCE 200 kt OR_LESS KLM39C HEADING 070 RIGHT".

The ASR projects HAAWAII (Highly Automated Air Traffic Controller Workstations with Artificial Intelligence Integration (The projectI (High)) and STARFiSH (Safety and Artificial Intelligence Speech Recognition (Project description of)) are further enhancing this ontology mainly with respect to ground and en-route commands in coordination with PJ.05-97-ASR and PJ.10-96-ASR. The ontology now comprising more than 120 different command types. These projects will also foster the automatic extraction of commands based on machine learning techniques.

The next section describes the concept, the evaluation metrics, and the implementation of the command prediction for the tower environment that was implemented for the first time.

3. Tower command hypotheses generator concept, implementation and validation goals

The tower command hypotheses generator is a new system developed by DLR using the experience of former projects regarding hypotheses information generation in the approach area. However, the command types used in tower environment are different from those of approach controllers of former projects. Additionally, the tower area comprises of many more command types than the implementation for approach. The tower command hypotheses generator predicts possible commands for the near future considering available data such as radar data, flight plans, and meteorological data (Ohneiser et al., 2019). This prediction is not a single forecasted command, but a set of possible commands (context). Examples for predictions in different air traffic situations according to the defined ontology are: "AEE2019 STARTUP", "BAW123 PUSHBACK", "AFR456 TAXI VIA A″, or "DLH789 CLEARED TAKEOFF RW13R".

The technical validation plan for the tower command hypotheses generator in the project activity PJ.16-04-ASR *exercise 240* evaluation foresaw two objectives and three criteria goals. The first objective was to assess the stability of the (ASR) system performance. The second objective was to assess the operational feasibility of the integration of the (ASR) system and its sub-systems into operations. Furthermore, three target numbers regarding the prediction quality should be reached. The relevant numbers are the command prediction error rate with its standard deviation (SD), the context prediction time, and the context portion predicted.

The lower the command prediction error rate the better, because an ABSR system can rely on accurate forecasts to avoid falsely rejecting as few commands as possible due to stated non-conformity to the context. The PJ.16-04-ASR project requires an average *command prediction error rate* below 10% with a standard deviation of less than 2.5%. It was assumed that the command prediction error rate for a first tower command prediction – particularly due to a greater variety of commands than in the approach environment – will slightly be higher than command prediction error rates of already advanced approach command predictions. The context prediction time should be below 5 s to enable a prediction at least for each radar data update cycle.

The third metric *context portion predicted is* defined as the number of forecasted commands divided by the total number of commands per callsign, which an ATCo theoretically could give. Multiple hundreds of commands are possible per aircraft (commands for speed, altitude, direction, ground clearances, etc. With reasonable values). Or in other words, it is the total number of predicted commands divided by the number of commands, which are theoretically modeled and are possible, e.g., the number of predicted heading commands is normally in the range of 10–40 per callsign, whereas the total number in our model is 144 (005, 010, 015, …355, 360 multiplied by two because the qualifiers LEFT and RIGHT are possible). The total number of heading commands is even higher, i.e., 720, if heading commands of a step size with one degree are considered. The lower the context portion predicted, the better, because a lower number of command hypotheses helps the ABSR system to faster choose the best fitting command hypotheses for a given utterance and to increase the command recognition rate in case of correct forecasts.

The command hypotheses needed to be generated for three remote airports at the same time due to the validation study and scenario layout. Therefore, there were three geographical regions defined to forecast commands with respect to aircraft within those airport regions. One further global geographic area covers the airspace between and around the airports, e.g., to predict commands for flights that fly from one to another of these three airports. After implementation and integration of the tower command hypotheses generator in a multiple remote tower environment, the prediction quality was assessed using data from a series of four successive human-in-the-loop studies.

4. Human-in-the-loop study with tower command hypotheses generator

4.1. Validation setup and simulation run conditions

The project PJ.16-04-ASR contained – amongst others – a validation exercise for the tower command hypotheses generator. The PJ.16-04-ASR exercise 240 "Controller Command Prediction for Remote Tower Environment" was hosted at DLR's Multiple Remote Tower Experimental Setup in Braunschweig, Germany. The series of four human-in-the-loop studies to evaluate the tower command hypotheses generator prototype took place in 2017 and 2018. The ATCos – as study subjects – had three rows of monitors presenting the camera image of the respective three airports and a head-down ATM system unit to monitor and safely influence the given traffic (see Fig. 1).

However, the command hypotheses did not influence the ATCo's or controller support system's active work. The simulated remote airports were run in parallel with different traffic. In the study with ATCos from HungaroControl the airports were located in Hungary: Budapest (LHBP), Debrecen (LHDC), Papá (LHPA), with ATCos from Oro Navigacija in Lithuania the airports were Vilnius (EYVI), Kaunas (EYKA), Palanga (EYPA). Five different traffic scenarios were used for Hungary and four for Lithuania. They comprised Instrument and Visual Flight Rules (IFR/VFR) traffic, but VFR traffic was never more than 20%. All simulation scenarios lasted 50 min and took place at day light conditions. The majority of traffic had to be controlled from the first listed tower, i.e., LHBP and EYVI. There were a few special situations that ATCos were faced with, e.g.,

- four simultaneous movements, i.e., departure or arrival (two at the main airport, one at the smaller airports each),
- VFR arrival and departure crossing,
- Remotely Piloted Aerial System (RPAS) in airspace,
- responsibility for ground movements.

The exercise was conducted jointly by DLR, HungaroControl, and Oro Navigacija under the umbrella of SESAR2020 solution PJ.05-02. This solution was responsible for the validation platform itself – without the tower command hypotheses generator – and the remote tower concept validation. The communication between ATCos and simulation pilots was done via radiotelephony (Yada console) on three different frequencies. The resulting wav files with controller utterances and radar data were captured on a Linux laptop.

Fig. 1. Multiple remote tower environment at DLR Braunschweig.

4.2. Data recordings

Pre-trials with seven Hungarian ATCos running four different air traffic scenarios were performed from November 13 to 21, 2017. Pre-trials with six Lithuanian ATCos running also four different scenarios took place from March 19 to 27, 2018. Pre-trials were used to collect data to develop the models for command prediction of the tower command hypotheses generator. Final trials with seven Hungarian ATCos running five different scenarios were performed from November 12 to 22, 2018. Final trials with six Lithuanian ATCos running four different scenarios each took place from December 3 to 11, 2018. The data recordings of those trials were used for evaluation of tower command hypotheses generator prediction accuracy. With pre-trial data (before summer 2018), machine learning algorithms were implemented to improve the accuracy of command hypotheses. The scenarios of pre-trials and trials were very similar, i.e., the callsigns, aircraft characteristics, airports, traffic density, etc. Were mainly the same. In the main trials slight changes to some aircraft movements were implemented. Furthermore, an additional scenario with a runway configuration change was evaluated in the main trials.

The complete training data set comprised 52 simulation runs with a duration of roughly 39.4 h. This included about 4700 voice utterances (wav files). 100% of the Hungarian and 30% of the Lithuanian tower utterances have been manually transcribed (speech-to-text), annotated (text-to concepts, i.e., transformation of word sequences to ATC concepts, callsigns, and commands), and checked for this learning approach. This sums up to more than 3400 transcription files and the same number of annotation files. When ignoring the "silence" between different wav-file occurrences, there were roughly 7 h (26 h "with silence") of annotated tower commands available for learning.

The data resulting from the final trials (after summer 2018) was used to test and perform the evaluation of command prediction accuracy. The reported portion of actually given annotated controller commands was compared to the tower command hypotheses from the trials. The complete evaluation data set comprises 59 simulation runs with a duration of roughly 45.6 h. This included about 4600 voice utterances. 25% of the Hungarian and 35% of the Lithuanian tower utterances have been manually transcribed, annotated, and checked for the testing and evaluation (9 simulation runs each). These data sum up to more than 1000 transcription and annotation files, each consisting of more than 2 h (12 h "with silence") of annotated tower commands available for evaluation purposes.

4.3. Determining parameters for machine learning

The determination of parameters for the machine learning algorithms is a pre-result that needs to be found first. Therefore, the methodology of how to find this result is shortly outlined in the following. A split of 80%/20% for training and test data can be used, which is very typical in the machine learning community to estimate the accuracy of the learned model.

As a first step, an appropriate window size for the machine learning approach needed to be found (for more background of the "window" use and the machine learning algorithms, refer to (Kleinert et al., 2017)). The "window" is a raster size (a certain rectangle in terms of latitude and longitude) and is used to cluster airspace areas, where certain controller commands can be expected (hypotheses). If the window size is huge, command types are predicted everywhere in the airspace. However, a

lineup far away from an airport does not make sense. Furthermore, a speech recognition engine would receive too many hypotheses to choose from.

If the window size is small, valid command types might not be forecasted, e.g., because the aircraft was just a few meters away from the forecast region being too small. Besides, a speech recognition engine would not receive an accurate set of hypotheses (context) including the actually given ones of the ATCo. Thus, a trade-off needs to be found for the window size. The window size is completely different to the "context size". The context size comprises of all command predictions at a given time. The absolute context size indicating the number of predicted commands per controller utterance occasion is a very important parameter to the context portion predicted and helps to find reasonable values for machine learning.

For determining the best window size, the Hungary-2017-11 data was used to train the command prediction model. This model was then used to test the Hungary-2017-11 data that were split into two halves (A/B). As this was done with all available data from the end of 2017 and is only a pre-result for applying the machine learning algorithms on later data, only Hungary-2017-11 data was used for determining the parameters. Different window sizes from 1 to 14 were used for this test. For the global geographic area between the three airport regions, the window raster bases on roughly one nautical mile times one nautical mile depending on the latitude/longitude coordinates, i.e., 1.5 arc minutes "width" and 1 arc minute "height" for central Europe. On the airports itself a finer window size was used, i.e., a window size of 1x1 for the three airport regions can roughly be converted to the sixtieth part of a nautical mile as a square, i.e., 1.5 arc seconds "width" and 1 arc second "height".

The command prediction error rate and the context size (number of forecasted commands) should be as low as possible. However, big context size normally results in low command prediction error rates and vice versa. Hence, it was decided to choose the window size parameter that does not show great differences in the results of the two aforementioned values compared to the parameter step before. The analysis result is shown for command prediction error rate in Fig. 2 and for context size in Fig. 3.

The command prediction error rate with a window size of 11 was roughly 97% of the error rate with a window size of 10 for data half B (red line). For data half A (blue line), the command prediction error rate from window size 10 to 11 only changed in the second value after the decimal. Hence, a window size of 11x11 seems to be a good size regarding the command prediction error rate.

The context size with a window size of 11 was 99% of the context size

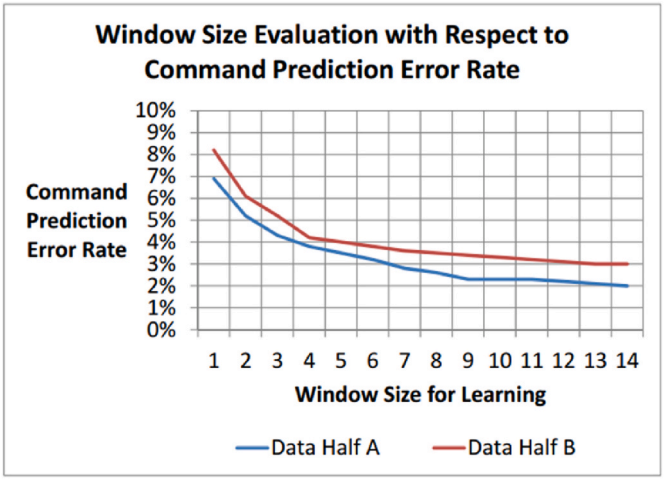


Fig. 2. Comparison of command prediction error rates for different window sizes.

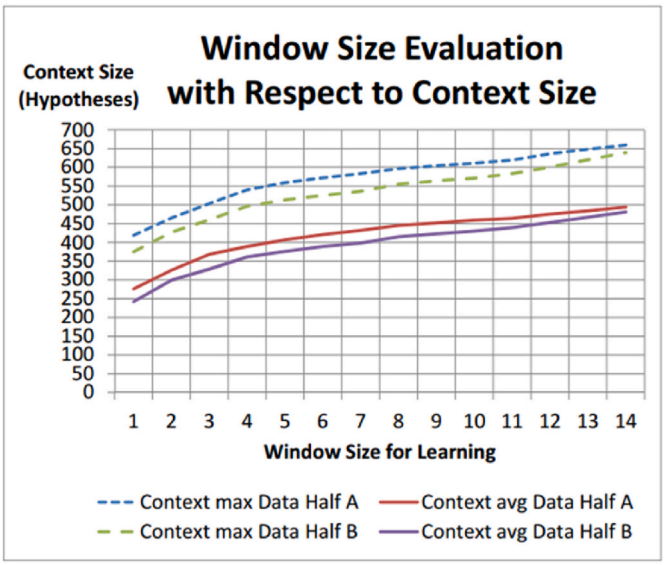


Fig. 3. Comparison of context size for different window sizes.

with a window size of 12 for data half A (blue dotted line). The context size with a window size of 11 was 98% of the context size with a window size of 12 for data half B (green dotted line). Thus, a window size of 11x11 was chosen as the best compromise between low command prediction error rates and smaller context size. Further increasing the window sizes marginally improves the command prediction error rate, but further increases the context size. The above reported window size is valid when analyzing all controller command types together.

However, there might be better fitting window sizes for single command types that have other characteristics with respect to airspace regions that are usually instructed by an ATCo. Table 2 shows the top twenty commands actually given by ATCos in descending order of their number of appearances (the most often used command at rank 1 appeared 1631 times, rank 20 appeared only 37 times in all Hungary-2017-11 data).

Therefore, an analysis of selected commands on one half of the data shows which window size could be chosen best for them individually as shown in Table 3.

The results in Table 3 show that a REPORT or TAXI command can occur nearly everywhere, but the area when a CLIMB or PUSHBACK command is given, can be determined very precisely. In the chosen

Table 2
Most frequently used command types of tower controllers.

Controller Command Type
INFORMATION (WINDDIRECTION, WINDSPEED, ATIS, TRAFFIC, …)
CLEARED (LANDING, TAKEOFF, TOUCH_GO, TO, VIA, etc.)
INIT_RESPONSE
TAXI
CLIMB
SQUAWK
REPORT (FINAL, BASE, etc.)
CONTACT_FREQUENCY
CONTACT
STARTUP
LINEUP
INFORMATION QNH
REPORT_MISCELLANEOUS
CONTINUE
VACATE (also with TO, VIA)
CALL_YOU_BACK
PUSHBACK
VFR_CLEARANCE
DIRECT_TO
ENTER_CTR

Table 3
Differences of best window sizes for machine learning of tower command types.

Controller Command Type	Best Window Size
REPORT	12
TAXI	11
INIT_RESPONSE	11
INFORMATION	9
VACATE	9
CLEARED	8
CONTACT_FREQUENCY	7
CLIMB	2
PUSHBACK	1

scenarios the ATCo only gave CLIMB commands either together with the initial clearance, when the aircraft is still at the parking position or shortly after taking off.

This analysis serves as an input for optimization of the tower command hypotheses generator towards future technology readiness level 6. For the further analysis, the above reported determined window size of 11x11 is used.

5. Automatic extraction of command annotation from transcription

5.1. Transcription and annotation of controller utterances

Manual transcription and annotation of controller utterances is a very time-consuming process. The software tool CoCoLoToCoCo (Controller Command Logging Tool for Context Comparison) concentrating on efficient usage has been developed to accelerate this process by DLR (see Fig. 4). This tool also performs a check of the set of predicted commands (context) to evaluate whether the given commands were forecasted or not. It also performs automatic plausibility checks for transcriptions and annotations with respect to the defined ontology (section II.B) format, air traffic rules, common typing errors, etc. (Shetty et al., 2020). Furthermore, the json format of the used file types that are explained in the following is checked. The graphical user interface of CoCoLoToCoCo offers several possibilities to view and manipulate the ATC utterance analysis files.

The upper middle file list displays all audio wav-files ("wave") with timestamps in different colors based on the transcription and annotation progress levels along with the comments. The field at the bottom allows to edit the word-by-word transcriptions in jcor-files ("json" and "correct") with frequently used phrases to be quickly copied from the lower right field. The pre-transcription can also be auto-generated if an annotation is already available. The middle part of the tool window is used for annotation via the six column menus (Callsign, Type, 2nd Type, Value, Qualifier, Condition). On the middle right side, the jcmd-file ("command") with the annotation of the current wav-file, e.g., "NAX6TW CLEARED TO ESSA" is shown – with context check (green is in context; red is not). The annotation can also be auto-extracted out of the transcription for further manual checking. On the upper right side, CoCoLoToCoCo maintains nfo-files ("inform") for comments and err-files ("error") listing the automatically detected error messages. The used json-format means that all file content is encapsulated into json tags for better readability and interoperability as shown in Fig. 5.

5.2. Description of command extraction algorithm

The module of the ABSR system, which is responsible for extracting commands from utterances is the *Command Extractor*. The tower command extractor uses a *Command Extraction Model* and the tower command hypotheses generator to carry out the extractions. Our Command Extraction Model aims at identifying commonly occurring patterns in ATCo utterances in order to automatically extract the command

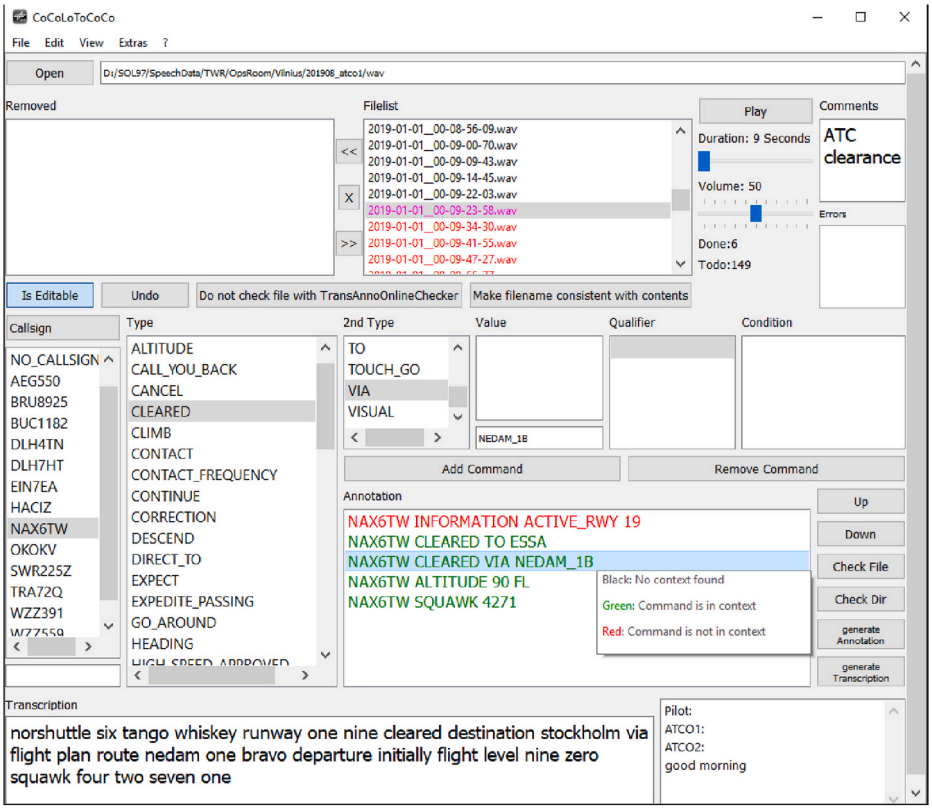


Fig. 4. Software tool for efficient transcription and annotation of controller utterances using standardized ontology terms and performing integrated hypotheses and plausibility checks.

```
01 {
02  "filename": "2020-08-31__09-51-02-47.wav",
03  "abstraction_layer_word_sequence":
04     "air_france seven two five turn left direct dexon bye"
05  "commands": [
06   {
07       "csgn": "AFR725",
08       "type": "DIRECT_TO",
09       "valu": "DEXON",
10       "qual": "LEFT"
11   }
12  ]
13 }
```

Fig. 5. Json-format for a file containing a single ATC command.

annotations associated with an ATCo utterance. The algorithm – explained in details in (Helmke et al., 2020) – consists of three general steps: (1) callsign extraction, (2) command extraction using keyword sequences, and (3) extraction of incomplete commands from known ATC concepts and unrecognized words of the utterance. Fig. 6 gives an overview on the complete command extraction algorithm:

An example utterance explains the algorithm well: "*good evening lufthansa seven hotel tango palanga tower wind zero five at eight knots runway zero five right cleared to land*". The Command Extraction Model uses the set of plausible commands predicted by the hypotheses generator, through which the set of plausible callsigns (or predicted callsigns) is also delivered, e.g., AFR783, DLH12C, DLH7HT. We begin by trying to extract the callsign from the first words of the transcribed utterance as described in line 1 of the algorithm (Fig. 6). In the given example, DLH7HT will be the extraction result from "*lufthansa seven hotel tango*".

Next, we try to identify command types by using keyword sequences. For example, keyword sequences such as "*wind*" or "*wind is*" could indicate a wind information command, and keyword sequences such "*cleared to land*" and "*cleared landing*" could indicate and trigger the extraction of the CLEARED LANDING command as defined in the ontology. We therefore identify that we may have commands INFORMATION WINDDIRECTION and INFORMATION WINDSPEED when we encounter "*wind*" in the utterance. Here, INFORMATION represents Command *Type*, and WINDDIRECTION and WINDSPEED represent Command *Second Type*. Moreover, the ontology defines that

```
01 Extract (predicted) callsign from first words of utterance;
02 while (NOT end of utterance reached){
03    Extract command if matching keyword sequence found;
04    if (type, value, unit, qualifier ... can be extracted){
05       Add command to list of already extracted commands;
06    }
07 }
08 while (NOT end of utterance reached){
09    Find ATC concept in unmatched words of utterance;
10    if (type, value, unit, qualifier ... can be extracted){
11       Add command to list of already extracted commands;
12    }
13    Find command hints in unmatched words of utterance;
14    if (type, value, unit, qualifier ... can be extracted){
15       Add command to list of already extracted commands;
16    }
17 }
18 while (NOT end of utterance reached){
19    Find first or further callsigns in unmatched words of
utterance;
20 }
```

Fig. 6. Pseudo-code algorithm to extract callsign and command annotations from air traffic controller utterance transcriptions (Helmke et al., 2020).

INFORMATION WINDDIRECTION commands must always have a value, which must be a 3-digit number such as 010, 360, etc. (representing the angle in degrees or the value "VARIABLE"). The ontology also specifies that command INFORMATION WINDSPEED also has a value representing wind speed (or "*calm*" that is extracted as "*0*") and a unit. For commands INFORMATION WINDDIRECTION and INFORMATION WINDSPEED, we try to extract their values and unit after the keyword "*wind*" in the utterance. For the given example, we successfully extract both INFORMATION commands resulting in "DLH7HT INFORMATION WINDDIRECTION 055" and "DLH7HT INFORMATION WINDSPEED 8 kt", respectively. If the extraction of the mandatory *value*, e.g., for INFORMATION WINDDIRECTION fails, the next matching keyword sequence is processed. Besides, commands can also have optional qualifier or condition fields.

Similarly, keyword sequence "*cleared to land*" results in the extraction of *Command Type* CLEARED and *Command Second Type* LANDING, with an optional runway value of RW05R found before the matched keyword sequence. If a valid runway is not found, the value is set to "*none*". In our example, we successfully extract "DLH7HT CLEARED LANDING RW05R". The command extraction is illustrated between lines 2 and 7 in Fig. 6. The given example can be rewritten and color-coded as:

"good evening lufthansa seven hotel tango palanga tower wind zero five five at eight knots runway zero five. right cleared to land"

Here, colors *red*, *green*, and *blue* mark the extracted *callsign*, *INFORMATION WINDDIRECTION*, *INFORMATION WINDSPEED*, and *CLEARED LANDING* commands, respectively. The black color-coded words have not yet been used for command extraction. These words do not match any keyword sequences. However, they can still be used to extract commands by identifying unmatched words or known ATC concepts. This next step is implemented between lines 8 and 17 of the algorithm in Fig. 6. ATC concepts contain runways, taxiways, taxipoints, waypoints, frequencies, etc. In the given example, "*palanga tower*" represents a valid position name PALANGA_TOWER. We extract a STATION command from this ATC concept directly, resulting in the extraction of "DLH7HT STATION PALANGA_TOWER". The remaining unmatched words are further analyzed for indications with respect to Command Types. For example, "*flight level*" and "decimal" could suggest ALTITUDE and CONTACT_FREQUENCY commands, respectively. In addition, fillers such as greetings (e.g., "*good bye*", "*bonjour*", etc.) are also recognized.

Lines 18 to 20 illustrate a second round of callsign extraction in order to extract callsign(s) which were not previously extracted. The algorithm applies Levenshtein distance calculations (Levenshtein, 1966) for evaluating the best matching callsign. Using the Levenshtein distance calculations, callsign DLH7HT is identified as the best match even for "*lufthansa six hotel tango*". This second callsign extraction step also examines if there are callsigns in the same utterance separated by the "*break break*" keyword sequence. The second round of callsign extraction after extracting all commands helps to avoid mixing up command and callsign symbols that would lead to wrong callsign/command extractions. For example, in "*speedbird two juliett x-ray papa tower …*", the word "*papa*" could either indicate an International Civil Aviation Organization (ICAO) alphabet letter (ICAO, 2017) from the callsign or could specify the Hungarian air base Pápa. The callsign extraction process for unrecognized callsigns at the end ensures that "*papa*" is correctly associated with the respective callsign/command (STATION command in our example).

5.3. Challenges in the command extraction process

Here, we shall discuss some of the challenges faced while implementing the tower command extractor. Even though ATCos must follow the ICAO standard phraseology (ICAO, 2016) while issuing instructions to pilots, they often tend to deviate from the standard phraseology,

thereby making the command extraction process more challenging. Some of the challenges faced are listed below:

- *Ambiguities in ATCo utterances "taxi to holding point runway one three right"* can be easily recognized as "TAXI TO HP_13R", but in *"taxi to runway one three right"*, the *"runway one three right"* could either be runway 13R or taxipoint HP_13R, making it difficult for the software to identify the correct ATC concept.
- *Mistakes in callsign utterances going beyond defined Levenshtein distance threshold* Saying "*hotel alfa kilo **whiskey***" instead of "*hotel alfa **sierra** kilo **victor***" makes extraction of the correct callsign HASKV challenging.
- *Misrecognition of callsign or wind values as runways* Parts of callsign and wind information that match runway values could be misrecognized as runways. For example, the "two three" in "*iberia three **two three** hold for landing traffic*" could also be recognized as runway value (RW23) associated with the HOLD_SHORT command.

The last example again shows the added value of command prediction, i.e., of a good prediction. If the callsign IBE323 is predicted, it is very likely that the "two three" does not correspond to the runway. If no IBE323 is predicted, but an IBE2A3, it is likely that "two three" is used for an abbreviated HOLD_SHORT command. The example also shows that sticking to phraseology increases safety, because word sequences which could be ambiguous for an automatic speech recognition system could also be challenging for a pilot especially if the pilot is not a frequent user of the corresponding airspace/airport/air traffic control unit. And last, but not least, it must be mentioned that we only described the challenges of concept and command extraction when the extraction is performed on manually transcribed utterances. More challenging is the process, when the word sequences are the output of a speech-to-text module. In the first case we can assume an ideal word error rate (WER) of 0%, but in the latter case, the command extraction must be able to extract something meaningful for WERs between 2% and 20%. Command Extraction Rates of 75% are still possible even if WER goes much beyond 10%, as it is shown in (Helmke et al., 2020) for our implementation.

5.4. Results of automatic command extraction on tower data

The data used for command extraction consisted of more than 16,000 commands from almost 6000 transcribed utterances. The data was collected from multiple remote tower simulation trials of Hungarian, Lithuanian, and Finnish controllers in the DLR TowerLab from 2017 to 2019. The number of commands per utterance ranged between 1 and 9. The results of command extraction for tower are shown in Table 4.

The extraction rate (ExtrR) is defined as the number of correctly recognized commands (Cmds) divided by the total number of actually given commands. Hereinafter, *gold annotations* refer to the manual, correct annotation of a human expert, i.e. the manual transformation of the uttered word sequences to the ATC concepts defined by the ontology. A command is said to be correctly recognized only if the callsign, command type including second type, value, unit, qualifier, and the condition are all correctly extracted. Hence, "DLH7HT CLEARED LANDING RW05R″ is different from "DLH7HT CLEARED LANDING none". Error rate (ErrR) is the percentage of wrongly extracted commands, i.e., the number of commands extracted wrongly divided by the total number of commands actually given. Rejection rate (RejR) is the percentage of gold annotations which are not extracted. A wrong command extraction is considered as a rejection, if for a given gold annotation (i) nothing is extracted, or (ii) command type NO_CONCEPT is extracted, or (iii) the correct command type is extracted, but with the callsign NO_CALLSIGN and this differs from the given gold annotation. Table 4 Also shows the callsign extraction rate (CsgnExtrR), which represents the percentage of correct callsign extractions (99.4%).

Table 4
Command extraction rates for tower.

Runs	#Extr Cmds	#Utterances	ExtrR	ErrR	RejR	CsgnExtrR
Hungary	11416	4061	97.3%	2.3%	1.7%	99.5%
Lithuania	4410	1665	96.0%	2.3%	2.5%	99.1%
Finland	287	131	94.2%	3.5%	3.4%	100%
All	16113	5857	96.7%	2.4%	2.0%	99.4%

Note 1: The sum of the command recognition, command error, and command rejection rate is normally slightly bigger than 100% (ExtrR + ErrR + RejR $\geq$ 100%): If exactly one command is said by the ATCo (gold annotation), but three commands are extracted, the error rate is at least 200% for this utterance. If, however, three commands are said and no command, except NO_CONCEPT, is extracted, the rejection rate is 100%.

Note 2: During implementation of the tower command extractor and the succeeding manual checks, errors in manual transcriptions and especially in manual annotations (gold annotations) have been found. This means that, e.g., some transcribed uttered words were misspelled, missing, etc. And that gold annotations included wrong command types or callsigns, commands were missing, inconsistencies existed due to different manual annotators, etc. Hence, a number of gold annotations and transcriptions have been updated compared to the first manual annotation and transcription in order to correct them and to also fit them to the enhanced ontology. This also means that there are still incorrect gold annotations existing, i.e., manual gold annotations with few errors are compared to automatically extracted annotations with few errors.

From Table 4 We see that an overall extraction rate and error rate of 96.7% and 2.4% respectively are achieved with the tower command extractor. The extraction rates obtained for individual runs corresponding to Hungary, Lithuania, and Finland airports are 97.3%, 96% and 94.2%, respectively. Here we see that our Command Extraction Model works best on Hungary data (best simulation run with 99.7% extraction rate) and worst on Finland data (worst simulation run with 90.8% extraction rate). The lower recognition rates for Finland could be attributed to the lower number of modeled specifics for Finland environment. An extraction rate of 96.7% in average means that about 365 commands are still not correctly recognized. Some of the reasons why we could not achieve an extraction rate of 100% are: (i) remaining errors in gold annotations, i.e., we assume that half of the extraction errors can just be corrected by correcting the gold transcriptions and/or annotations, (ii) large and unpredictable phraseology deviations, (iii) corrections in utterances like "runway zero correction one three right", (iv) mistakes in utterances, for example saying invalid runway values like "runway zero four" instead of "runway zero five". The remaining callsign extraction errors also have different reasons such as (i) only the airline without flight number was uttered, (ii) the wrong airline was uttered, (iii) two letters of the callsign are mixed up, (iv) and non-ICAO rule conform abbreviations of the callsign have been used.

6. Results of tower command hypotheses generator validation exercise

6.1. Applying machine learning techniques and evaluation of command prediction quality

There are four different data sets for tower command hypotheses generator evaluation called Hungary-2017-11, Lithuania-2018-03, Hungary-2018-11, and Lithuania-2018-12. As described above, an evaluation analysis consists of training the prediction models via machine learning and testing afterwards. Four different training and test set combinations have been used as per the data history. Those training/test sets are called as listed in Table 5.

For each of the annotated runs there was a set of predicted commands per timetick called context (ctx) every time the ATCo uttered something. The results of the command hypotheses evaluation are reported in historical order in the following sub-sections after the overview in Table 6.

Table 5
Overview on training and test data sets with time of recording and air navigation service provider information.

Name of Evaluation	Training with Dataset	Test with Dataset
HUNGARY	Hungary-2017-11	Hungary-2018-11
LITHUANIA	Lithuania-2018-03	Lithuania-2018-12
BOTH_COUNTRIES	Hungary-2017-11, Lithuania-2018-03	Hungary-2018-11, Lithuania-2018-12
COMPLETE	Hungary-2017-11, Lithuania-2018-03, Hungary-2018-11	Lithuania-2018-12

Table 6
Overview of Command Prediction Error Rate and Context Size for the four different Evaluations.

Name of Evaluation	Command Prediction Error Rate	Standard Deviation	ctx_avg	ctx_max
HUNGARY	7.8%	2.57%	450	593
LITHUANIA	12.5%	3.6%	340	498
BOTH_COUNTRIES	7.3%	2.46%	548	760
COMPLETE	7.5%	3.7%	629	900

Command prediction error rate (underlined if significantly below 10%), standard deviation of command prediction error rate as well as average and maximum number of hypotheses relevant for context size in ctx-files (ctx_avg, ctx_max).

6.1.1. Hungary

For the six annotated Hungarian simulation runs containing a scenario without runway configuration change, the average command prediction error rate per run is 7.8% (standard deviation SD: 2.57%).

As there was no runway configuration change in the training data (Hungary-2017-11 runs), this could not be learned. Thus, different runway configurations for commands have not been predicted for 2018 data. Hence, such a new aspect negatively influences the command prediction quality i.e., a change in the runway configuration affected the prediction quality of commands such as TAXI, VACATE, CLEARED LANDING/TAKEOFF, etc. Taking three further simulation runs with runway configuration changes of test data also into account, the command prediction error rate is 9.1% (SD: 2.85%). 450 commands have been predicted on an average (ctx_avg). When analyzing the biggest set of predicted commands per run (ctx_max), this averages to 593 for the Hungary-2018-11 runs.

All of the top 14 commands (that were at least used in 2.2% of all commands from the Hungary-2018-11 scenarios without runway change) showed command prediction error rates below 8.3%. However, for the "TAXI TO"-command this is only true if a generalization of stands is made (so specific stands such as "R107" were not forecasted, but only "TAXI TO STAND").

6.1.2. Lithuania

For the nine annotated Lithuanian simulation runs, the average command prediction error rate per run is 12.5% (SD: 3.6%). Many scenarios had aircraft repeating a touch-and-go or a go-around respectively which are more difficult to predict. When ignoring such command predictions, the command prediction error rate would be around 7%. 340 commands have been predicted as context on an average (ctx_avg). When analyzing the ctx_max, this averages to 498 for the Lithuania-2018-12 runs.

6.1.3. BOTH_COUNTRIES

For the six annotated Hungarian simulation runs – however, machine learning performed on Hungary-2017-11 and Lithuania-2018-03 – the average command prediction error rate per run is 7.3% (SD: 2.46%). These numbers are reported as the main result as the technical validation plan foresaw validation trials with Hungarian ATCos and a command prediction model that learned from all available data before.

Taking also the nine annotated Lithuanian simulation runs into account, the average command prediction error rate per run is 7.9% (SD: 3.2%). Taking all 18 annotated simulation runs (with runway configuration changes) into account, the command prediction error rate is still below 10% – in a range between 8 and 9%. For the Autumn-2018 runs the ctx_avg and ctx_max were 548 and 760, respectively.

6.1.4. COMPLETE

For the nine annotated Lithuanian simulation runs – however, machine learning performed on Hungary-2017-11, Lithuania-2018-03, and Hungary-2018-11 – the average command prediction error rate per run is 7.5% (SD: 3.7%). The ctx_avg was 629, the ctx_max was 900 for the Lithuania-2018-12 runs.

6.1.5. Significance of results

The confidence in the results of the exercise is high due to the number of simulation runs, i.e., the command prediction error rates have high statistical significance. The performed *t*-test tested against the required average value of 10%. The obtained p-values are 1.18% and 3.29% for the "BOTH_COUNTRIES-data" and the core "HUNGARY-data", respectively. Normally, a statistical significance of below 5% is required to significantly support the underlying assumption. Hence, we can conclude that the average command prediction error rate per run is very surely below 10%. However, for the reported "LITHUANIA-data", we assume that the command prediction error rate is above 10% due to the *t*-test on the other side of the threshold value with a p-value of 2.46%. But again, for the "COMPLETE"-data, testing the same Lithuanian files with the reported model learned on more data, we can assume that the command prediction error rate is below 10% with a p-value of 2.91%.

6.1.6. General notes

The context portion predicted was below 10% all the times. However, the ctx_avg shows that more training data lead to more command predictions. This results in less command prediction errors, but is increasing the search space for a speech recognition engine. Furthermore, it has to be noted, that there are still unintended human-made transcription and annotation errors in the data. The CoCoLoToCoCo tool notifies the human annotator about possible errors. This tool continuously improves; however, it is not able to detect all errors. Besides, the simulation runs chosen to be annotated and analyzed might have a small influence on the errors detected. In general, it can be stated that a command prediction error rate below 10% was achieved and can be further enhanced.

6.2. Real-time aspects of command prediction with respect to given commands

In average, software-based generation of command hypotheses took 119 ms (analyzed from log files of 59 simulation runs of final trials in autumn 2018) which is a factor of 40 better than required.

Context (set of command hypotheses) has been generated more than 21,000 times. As traffic density during tower trials was rather medium to low (compared to former ATC approach trials, in which context was used), context was generated only every 10 s. However, the measurements show that it is possible to do it 80 times more often if needed. The context generation frequency was also reasonable, because the average duration of radiotelephony (RT) calls from tower controllers is just slightly shorter than 10 s as shown in Fig. 7. These five to 7 s of simple ATCo command communication have also been reported for en-route sectors (Rodríguez and Cordero, 2014).

The portion of time used for commands and the number of commands is visualized in Fig. 8. This emphasizes the potential for workload reduction by usage of ASR, because each command nowadays means that ATCos also need to perform manual input into the ATC system.

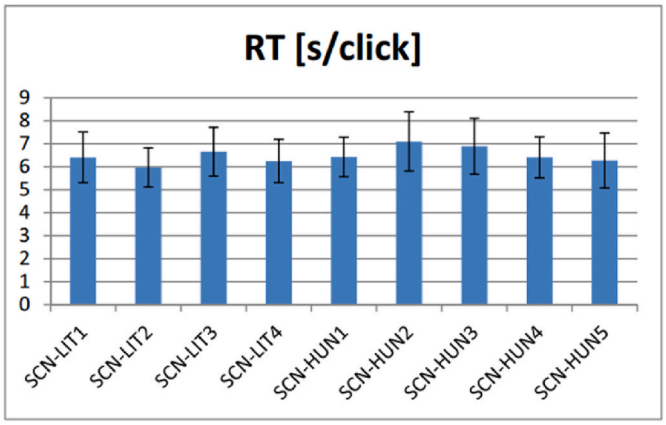


Fig. 7. Average duration of radiotelephony transmissions in seconds from tower controller to pilot (with positive and negative standard deviation).

Within the number of roughly 21,000 command prediction attempts, only four of them lasted longer than 1 s. However, it was always faster than 2 s. Further analyzes showed that the command prediction itself was not slow, but fetching input data from databases and storing it slows down the predictions. All other attempts succeeded in less than 1 s. These numbers are highly reliable regarding the number and duration of runs.

6.3. Debriefing comments

Each ATCo also joined a debriefing session. This included a discussion about strengths and weaknesses of transcription and annotation focusing on the appropriateness of the ontology and the CoCoLoToCoCo support tool. Furthermore, the ATCos were asked for their opinion about an ABSR system in the daily life CWP. Paraphrased answers are reported in the following.

ABSR do not seem to be of huge interest for one of the controllers, who does not need to enter many things in the CWP HMI today. A list of last clearances in written form would be good for another ATCo as he also uses the hearback replay button in his CWP. ABSR would be a good support for improvement of the HMI, for departure clearances, as well as flight levels and squawks. Also, regarding future ATC systems for the ground – that will be introduced in the next two years – ABSR support would be helpful as ATCos need to enter all waypoints, taxi points, routes, etc. for applications such as "follow-the-greens". Other ATCos also mentioned safety critical aspects that could be supported by ABSR. ABSR should recognize "RWY blocked" and show it on the HMI if ground personnel enter a runway. Furthermore, a reliable readback failure presentation would be great, especially for digits in clearances, frequencies, and for the attributes such as "left" and "right".

Thus, the central recommendation for HMI design is to display the semantics (annotation) of utterances at the CWP. By doing so, ATCos have the chance to check and compare implications from the past or for the future. Hence, this will be implemented for the following human-in-the-loop study with a completely integrated prototypic ABSR system at a multiple remote tower CWP setup.

7. Summary and outlook

7.1. Summary of command hypotheses validation results

The complete trials generated 107 recorded simulation runs for data analysis. The results of the simulation runs with respect to the tower command hypotheses generator are positive and encouraging. The hypotheses generator fulfilled both validation objectives, i.e., operational feasibility as well as performance stability was validated. One aim was to have a command prediction time at least as fast as the update rate of radar data, i.e., the prediction time should be below 5 s. Command predictions were forecasted timely and could be generated on an average every 120 ms and were always below 2 s. The command prediction error rate achieved its targets to stay below 10% with a standard deviation below 2.5%. Furthermore, the context portion predicted was below 10% for all simulation runs, however, having limited expressiveness as a valuable result. Nevertheless, the context portion predicted was in a comparable dimension as for the approach hypotheses generator.

AcListant® project showed that ATC command hypotheses improve the recognition quality of an ABSR system in the approach environment. Hence, the positive evaluation of the forecast quality of the tower command hypotheses generator is a central factor for a later tower ABSR system. This future system could present the recognized commands to the controller and might be used similar to the actions in the other four PJ.16-04-ASR exercises.

7.2. Summary of command extraction validation results

16,000 tower controller commands from 6,000 utterances have been manually transcribed and annotated to serve as the "gold" (correct) baseline for comparison. The developed tower command extractor automatically extracted commands from the manually transcribed utterances with an average accuracy of 96.7%. The callsign extraction rate even reached 99.4%. This good extraction rate again helps to improve the command prediction, because it can learn from more accurate data. The developed tower command extractor can be used on any tower command transcription independent of the ASR engine's automatic or manual transcription.

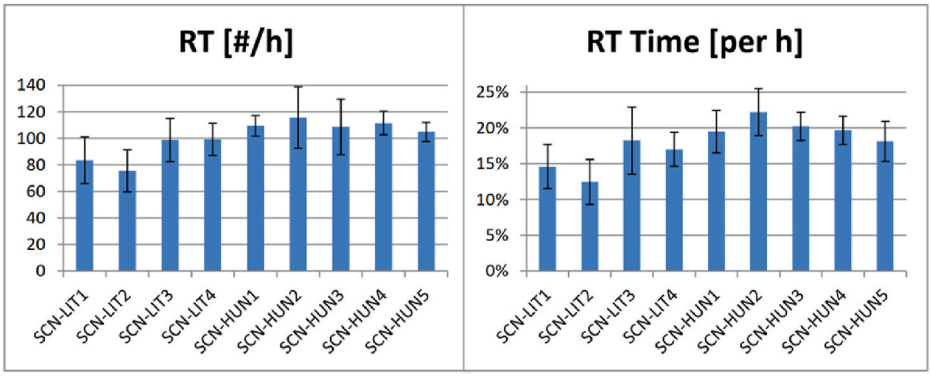


Fig. 8. Percentage of time and number of tower controller utterances per hour using radiotelephony (with positive and negative standard deviation).

7.3. Outlook on future work

There will be further industrial research on ABSR in SESAR's Wave 2 project PJ.05 "Digital Technologies for Tower" more precisely solution 97 "HMI Interaction modes for Airport Tower". Meanwhile, an enhanced tower command hypotheses generator and command extractor from DLR was integrated with a speech recognizer from the Swiss research institute Idiap to generate the first ABSR system for towers transitioning from technology readiness level 2 to 4. A real-time simulation exercise with controllers will be conducted in Remote TowerLab at DLR Braunschweig early 2022. The trials are supported by the air navigation service provider partners of B4 (ON from Lithuania, PANSA from Poland, and ANS-CR from Czech Republic) as well as COOPANS (ACG from Austria and CCL from Croatia).

So far, the command prediction error rate and the command extraction error rate on gold transcription for aerodrome ATC are known. However, there are – conform to planning – still many aspects which are unknown, compared to the results from AcListant® and the approach environment such as (1) the performance of an adapted basic speech recognition engine on tower utterances used by partly different air navigation service providers' ATCos with other English accents, (2) the extent to which tower command hypotheses, including more types of clearances, e.g., for VFR flights than just the approach ATC, improve the command recognition error rate on automatic, partly erroneous, transcriptions, and (3) possible workload reductions and their effects on aerodrome ATC performance.

Regarding the implementation of the enhanced tower command hypotheses generator, there are different aspects of improvement for the context quality. The set of hypotheses should be minimized and must fulfil more requirements valid on ground. A state machine approach – complementing the machine learning approach – could deliver even more background knowledge for the hypotheses generator. There are more single actions and command types that succeed each other in a certain order in the tower than in the approach environment. If, e.g., once a taxi clearance follows a pushback clearance, then the likelihood of a startup or landing clearance is almost zero. However, the likelihood of a lineup clearance is very high. This of course depends on the accuracy of the data quality – e.g., command extraction error rate - of former clearances. If they are derived from the ABSR system output, the follow-up states of the aircraft and thus of clearances is connected to a certain probability. Individual window sizes per controller command type prediction as described above can further improve machine learning results and thus command hypotheses accuracy. Additionally, the growing amount of data, i.e., (1) available radar and speech data for training of all scenarios and environments that are tested and (2) annotated speech data to optimize command prediction error rate and context portion predicted helps to build a tower command hypotheses generator on higher technology readiness levels.

An evaluation of speech data will further be done with real life data from Vilnius, Vienna, and Prague tower. The transcription and annotation tool CoCoLoToCoCo will provide further functionality of recording readback errors. This will enable the training of automatic readback error detection from controller as well as pilot utterances. In order to do this, the tower command extractor must be enhanced for real-life controller and pilot utterances as well. Furthermore, the ABSR systems with tower command hypotheses generator and command extractor should be brought closer to the operation's room, i.e., real towers or remote towers to support various applications of using speech information.

CRediT authorship contribution statement

Oliver Ohneiser: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Project administration. **Hartmut Helmke:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft. **Shruthi Shetty:** Software, Formal analysis, Writing – original draft. **Matthias Kleinert:** Software, Writing – review & editing. **Heiko Ehr:** Validation, Resources, Data curation, Writing – review & editing. **Šarūnas Murauskas:** Resources, Writing – review & editing. **Tomas Pagirys:** Resources, Writing – review & editing.

ACKNOWLEDGMENT

Thanks to all ATCos contributing to the validation trials with recorded speech utterances. Three SESAR2020 industrial research projects PJ.16-04 CWP HMI (Wave-1), PJ.10-96 (Wave-2), and PJ.05-97 (Wave-2), all including an automatic speech recognition (ASR) activity, and the exploratory research project HAAWAII have received funding from the SESAR Joint Undertaking under the European Union's grant agreement No. 734141, 874464, 874470, 884287 as well as the project STARFiSH funded by the German Federal Ministry of Education and Research. The DLR tower command hypotheses generator has first been developed in PJ.16-04-ASR EXE-240. The ontology for controller command annotation has been developed and enhanced in several above-mentioned projects on Assistant Based Speech Recognition.

REFERENCES

Chen, S., Kopald, H.D., Elessawy, A., Levonian, Z., Tarakan, R.M., 2015. Speech Inputs to Surface Safety Logic Systems. IEEE/AIAA 34th Digital Avionics Systems Conference (DASC), Prague, Czech Republic.

Chen, S., Kopald, H.D., Chong, R., Wei, Y., Levonian, Z., 2017. Read back error detection using automatic speech recognition. In: 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017).

Ciupka, S., 2012. "Siris big sister captures DFS," original German title: "Siris große Schwester erobert die DFS". transmission 1.

Cordero, J.M., Dorado, M., de Pablo, J.M., 2012. Automated speech recognition in ATC environment. In: Proceedings of the 2nd International Conference on Application and Theory of Automation in Command and Control Systems (ATACCS'12). IRIT Press, Toulouse, France, pp. 46–53.

Cordero, J.M., Rodríguez, N., de Pablo, J.M., Dorado, M., 2013. Automated Speech Recognition in Controller Communications Applied to Workload Measurement. 3rd SESAR Innovation Days, Stockholm, Sweden.

FAA, 2012. National Aviation Research Plan (NARP). FAA.

Federal Aviation Administration, 2019. NextGen Implementation Plan – 2018-19.

Hamel, C., Kotick, D., Layton, M., 1989. "Microcomputer System Integration for Air Control Training," Special Report SR89-01. Naval Training Systems Center, Orlando, Florida, USA.

Helmke, H., Rataj, J., Mühlhausen, T., Ohneiser, O., Ehr, H., Kleinert, M., Oualil, Y., Schulder, M., 2015. Assistant-based speech recognition for ATM applications. In: 11th USA/Europe Air Traffic Management Research and Development Seminar (ATM2015).

Helmke, H., Ohneiser, O., Mühlhausen, T., Wies, M., 2016. Reducing controller workload with automatic speech recognition. In: IEEE/AIAA 35th Digital Avionics Systems Conference (DASC).

Helmke, H., Ohneiser, O., Buxbaum, J., Kern, C., 2017. Increasing ATM efficiency with assistant-based speech recognition. In: 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017).

Helmke, H., Slotty, M., Poiger, M., Herrer, D.F., Ohneiser, O., et al., 2018. Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ.16-04. In: IEEE/AIAA 37th Digital Avionics Systems Conference (DASC).

Helmke, H., Kleinert, M., Rataj, J., Motlicek, P., Klakow, D., Kern, C., Hlousek, P., 2019. "Cost reductions enabled by machine learning in ATM - how can automatic speech recognition enrich human operators' performance?. In: 13th USA/Europe Air Traffic Management Research and Development Seminar (ATM2019).

Helmke, H., Kleinert, M., Ohneiser, O., Ehr, H., Shetty, S., 2020. Machine learning of air traffic controller command extraction models for speech recognition applications. In: IEEE/AIAA 39th Digital Avionics Systems Conference (DASC), Completely Virtual and Not as Planned in San Antonio.

ICAO, 2016. Doc 4444, Procedures for Air Navigation Services, Air Traffic Management. ICAO, Montréal, Canada.

ICAO, 2017. Doc 8585, manual on designators for aircraft operating agencies. In: Aeronautical Authorities and Services. ICAO, Montréal, Canada.

Kleinert, M., Helmke, H., Siol, G., Ehr, H., Finke, M., Srinivasamurthy, A., Oualil, Y., 2017. Machine learning of controller command prediction models from recorded radar data and controller speech utterances. In: 7th SESAR Innovation Days.

Kleinert, M., Helmke, H., Ehr, H., Kern, C., Klakow, D., Motlicek, P., Singh, M., Siol, G., 2018. Building Blocks of Assistant Based Speech Recognition for Air Traffic Management Applications. 8th SESAR Innovation Days, Salzburg, Austria.

Kleinert, M., Helmke, H., Siol, G., Ehr, H., Cerna, A., Kern, C., Klakow, D., Motlicek, P., et al., 2018. Semi-supervised adaptation of assistant based speech recognition models for different approach areas. In: IEEE/AIAA 37th Digital Avionics Systems Conference (DASC).

Kleinert, M., Helmke, H., Moos, S., Hlousek, P., Windisch, C., Ohneiser, O., Ehr, H., Labreuil, A., 2019. Reducing Controller Workload by Automatic Speech Recognition Assisted Radar Label Maintenance. 9th SESAR Innovation Days, Athens, Greece.

Levenshtein, V.I., Feb. 1966. "Binary codes capable of correcting deletions, insertions, and reversals," in. In: Soviet Physics – Doklady 10, vol. 8.

Ohneiser, O., Helmke, H., Kleinert, M., Siol, G., Ehr, H., Hobein, S., Predescu, A.-V., Bauer, J., 2019. Tower Controller Command Prediction for Future Speech Recognition Applications. 9th SESAR Innovation Days, Athens, Greece.

Oualil, Y., Schulder, M., Helmke, H., Schmidt, A., Klakow, D., 2015. Real-Time Integration of Dynamic Context Information for Improving Automatic Speech Recognition. Interspeech, Dresden, Germany.

Project Description of STARFiSH (Safety and Artificial Intelligence Speech Recognition) Listed in German. n.d. https://www.softwaresysteme.pt-dlr.de/de/ki-in-der-praxis.php.

Rataj, J., Helmke, H., Ohneiser, O., 2021. AcListant with continuous learning: speech recognition in Air Traffic Control. Air Traffic Management and Systems IV - Selected Papers of the 6th ENRI International Workshop on ATM/CNS (EIWAC2019). Electronic Navigation Research Institute, Springer, Singapore.

Rodríguez, N., Cordero, J.M., 2014. Relationship between workload andduration of ATC voice communications. In: 6 Th International Conferenceon Research in Air Transportation.

Schäfer, D., 2001. "Context-sensitive Speech Recognition in the Air Traffic Control Simulation," Eurocontrol EEC Note No. 02/2001. PhD Thesis of the University of Armed Forces, Munich.

Schier, S., Manske, P., 2015. "VisiTop II – Briefing-Unterlagen," Section 4.2. DLR-internal report, Braunschweig, Germany.

Schmidt, A., 2014. "Integrating Situational Context Information into an Online ASR System for Air Traffic Control," Master Thesis. Saarland University (UdS), Germany.

SESAR Joint Undertaking, 2020. European ATM Master Plan Executive View – Digitalising Europe's Aviation Infrastructure.

Shetty, S., Ohneiser, O., Grezl, F., Helmke, H., Motlicek, P., 2020. "Transcription and Annotation Handbook," HAAWAII Deliverable D3. HAAWAII project.

Shore, T., Faubel, F., Helmke, H., Klakow, D., 2012. Knowledge-based Word Lattice Rescoring in a Dynamic Context. Interspeech, Portland, Oregon, USA. Sep. 2012.

Skaltsas, G., Rakas, J., Karlaftis, M.G., 2013. An analysis of air traffic controller-pilot miscommunication in the NextGen environment. J. Air Transport. Manag. 27, 46–51.

Srinivasamurthy, A., Motlicek, P., Himawan, I., Szaszák, G., Oualil, Y., Helmke, H., 2017. Semisupervised learning with semantic knowledge extraction for improved speech recognition in air traffic control. In: INTERSPEECH 2017. 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden.

Tarakan, R., Baldwin, K., Rozen, R., 2008. An Automated Simulation Pilot Capability to Support Advanced Air Traffic Controller Training. 26th Congress of the International Council of the Aeronautical Sciences, Anchorage, Alaska, USA.

The Project AcListant® (Active Listening Assistant). n.d. http://www.aclistant.de/wp.

The Project MALORCA (Machine Learning of Speech Recognition Models for Controller Assistance). n.d. http://www.malorca-project.de.

The Project HAAWAII (Highly Automated Air Traffic Controller Workstations with Artificial Intelligence Integration). n.d. https://www.haawaii.de/wp.

Tobaruela, G., Schuster, W., Majumdar, A., Ochieng, W.Y., Martinez, L., Hendrickx, P., 2014. A method to estimate air traffic controller mental workload based on traffic clearances. J. Air Transport. Manag. 39, 59–71.

Young, S.R., Ward, W.H., Hauptmann, A.G., 1989. Layering predictions: flexible use of dialog expectation in speech recognition. In: Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI89). Morgan Kaufmann, pp. 1543–1549.

Young, S.R., Hauptmann, A.G., Ward, W.H., Smith, E.T., Werner, P., Feb. 1989. High level knowledge sources in useable speech recognition systems. Commun. ACM 32 (2), 183–194.

aerospace

MDPI

Article

Assistant Based Speech Recognition Support for Air Traffic Controllers in a Multiple Remote Tower Environment

Oliver Ohneiser [1,*], Hartmut Helmke [1], Shruthi Shetty [1], Matthias Kleinert [1], Heiko Ehr [1], Sebastian Schier-Morgenthal [1], Saeed Sarfjoo [2], Petr Motlicek [2], Šarūnas Murauskas [3], Tomas Pagirys [3], Haris Usanovic [4], Mirta Meštrović [5] and Aneta Černá [6]

[1] German Aerospace Center (DLR), Institute of Flight Guidance, Lilienthalplatz 7, 38108 Braunschweig, Germany; hartmut.helmke@dlr.de (H.H.); shruthi.shetty@dlr.de (S.S.); matthias.kleinert@dlr.de (M.K.); heiko.ehr@dlr.de (H.E.); sebastian.schier@dlr.de (S.S.-M.)
[2] Idiap Research Institute, Centre du Parc, Rue Marconi 19, 1920 Martigny, Switzerland; saeed.sarfjoo@gmail.com (S.S.); petr.motlicek@idiap.ch (P.M.)
[3] AB "Oro Navigacija" (ON), Air Navigation Service Provider of Lithuania, Balio Karvelio St. 25, 02184 Vilnius, Lithuania; murauskas.s@ans.lt (Š.M.)
[4] Austro Control (ACG), Österreichische Gesellschaft für Zivilluftfahrt mbH, Air Navigation Service Provider of Austria, Schnirchgasse 17, 1030 Vienna, Austria
[5] Croatia Control (CroControl), Air Navigation Service Provider of Croatia, Rudolfa Fizira 2, 10410 Velika Gorica, Croatia
[6] Air Navigation Services of the Czech Republic (ANS CR), Navigační 787, 25261 Jeneč u Prahy, Czech Republic
* Correspondence: oliver.ohneiser@dlr.de; Tel.: +49-531-295-2566

Abstract: Assistant Based Speech Recognition (ABSR) systems for air traffic control radiotelephony communication have shown their potential to reduce air traffic controllers' (ATCos) workload. Related research activities mainly focused on utterances for approach and en-route traffic. This is one of the first investigations of how ABSR could support ATCos in a tower environment. Ten ATCos from Lithuania and Austria participated in a human-in-the-loop simulation to validate ABSR support within a prototypic multiple remote tower controller working position. The ABSR supports ATCos by (1) highlighting recognized callsigns, (2) inputting recognized commands from ATCo utterances in electronic flight strips, (3) offering correction of ABSR output, (4) automatically accepting ABSR output, and (5) feeding the digital air traffic control system. This paper assesses human factors such as workload, situation awareness, and usability when ATCos are supported by ABSR. Those assessments result from a system with a relevant command recognition rate of 82.9% and a callsign recognition rate of 94.2%. Workload reductions and usability improvement with *p*-values below 0.25 are obtained for the case when the ABSR system is compared to the baseline situation without ABSR support. This motivates the technology to be brought to a higher technology readiness level, which is also confirmed by subjective feedback from questionnaires and objective measurement of workload reduction based on a performed secondary task.

Keywords: air traffic controller; multiple remote tower; assistant-based speech recognition; automatic speech recognition and understanding; electronic flight strips

Citation: Ohneiser, O.; Helmke, H.; Shetty, S.; Kleinert, M.; Ehr, H.; Schier-Morgenthal, S.; Sarfjoo, S.; Motlicek, P.; Murauskas, Š.; Pagirys, T.; et al. Assistant Based Speech Recognition Support for Air Traffic Controllers in a Multiple Remote Tower Environment. *Aerospace* **2023**, *10*, 560. https://doi.org/10.3390/aerospace10060560

Academic Editor: Judith Rosenow

Received: 5 April 2023
Revised: 27 May 2023
Accepted: 7 June 2023
Published: 14 June 2023

1. Introduction

Speech recognition and speech understanding have found their way into use in daily life. While speech recognition has become quite robust with growing amounts of data, speech understanding remains a challenge given the complexity of verbal utterances' semantics. However, high accuracy in speech understanding is needed for human operators that supervise safety-critical processes, such as in aviation. Only then, users of speech recognition and understanding systems such as controllers will accept them and can benefit from their support, e.g., through workload reduction. Nowadays, tower controllers are burdened with manually maintaining flight strips, even if the content that needs to be

entered in such flight strips is also communicated verbally in air traffic control radio telephony. This article presents one of the first prototypes of a speech recognition and understanding system to support ATCos in the tower environment in maintaining digital flight strips—in our case, even in a simulated multiple remote tower environment.

Our conducted validation study with ten air traffic controllers (1) quantifies any productivity enhancements in terms of mental workload, situation awareness, satisfaction, acceptance, trust, and usability through the advanced support functionalities in the digital system with automatic flight strip maintenance and highlighting features (independent variable); (2) quantifies the quality of speech-to-text and text-to-concept functionality; and (3) gathers feedback on the prototypes' functionality and visualization.

1.1. Related Work

1.1.1. Automatic Speech Recognition and Understanding in Air Traffic Management

During the last decades, a row of prototypes for speech recognition and understanding [1] in the air traffic management (ATM) domain has been developed. Early prototypes intended to support air traffic control (ATC) training and to reduce the number of required simulation pilots [2,3]. ATC events have been recognized from utterances to estimate controller workload [4,5]. The integration of contextual knowledge from an electronic assistant system for the speech recognition and understanding process [6] reduced recognition error rates [7]. These so-called assistant-based speech recognition (ABSR) systems initially focused on the approach environment [8]. For interoperability and comparability, rules for transcription (speech-to-text) and annotation (text-to-concepts)—so-called ontologies—have been defined and agreed upon between the major European ATM stakeholders [9]. Due to these rules, ATC utterances always comprise a callsign and at least one command that can consist of a type, unit, qualifier, and conditions. Later, ABSR systems were enhanced and enrolled on the en-route [10], apron [11,12], and tower environment [13]. This included the prediction and extraction of ATC commands [14]. Further research prototypes enhanced the ontologies, worked on speech recordings and radar data from real operations rooms, especially, but not limited to, recognizing callsigns [15–17], pre-filled aircraft radar labels that reduced the workload of ATCos [18,19], and implemented automatic readback error detection [10,20]. However, there was no validation of a sophisticated ABSR system's support for tower controllers, especially in a multiple remote tower setup using such a system in a high-fidelity laboratory environment.

1.1.2. Multiple Remote Air Traffic Control Tower and Human Operator Performance

The history of laboratory remote tower working positions started over two decades ago [21]. Recent research focused on human performance in multiple remote tower environments, i.e., where an ATCo is responsible for more than one remote airport at the same time. This started with analyzing eye-tracking data to characterize tower controllers' visual attention [22]. The research went on to investigate the changes in monitoring tasks and drafting multimodal interaction to support human operators at the controller working position (CWP) [23]. The latest research concentrated on workload assessment [24], operational feasibility and safety [25], as well as a supervisor position [26]. With fostering the technology maturity, questions regarding standardization with the European Organization for Civil Aviation Equipment (EUROCAE) and the European Union Aviation Safety Agency (EASA) guidelines have been developed [21]. Furthermore, the certification process for multiple remote tower operations has been sketched [27].

In the multiple remote tower environment, the human ATCo remains a central mean for the overall performance, with or without ABSR support. Related work on human performance assessment with standardized questionnaires is explained together with their results in the subsections of the result Section 3.

1.2. Structure of the Article

Section 2 describes the setup for the validation of ABSR support for ATCos and the conduction of this study. Section 3 presents the study results for the two aspects "Application of ABSR" and "ABSR in an ATM environment", i.e., results on speech recognition performance (Section 3.1) and speech understanding performance (Section 3.2) as well as on human factors such as mental workload, situation awareness, satisfaction, acceptance, trust, and usability (Sections 3.3–3.10), and ends with general feedback from ATCos (Section 3.11). Section 4 discusses the major study results for the fast readers who just quickly scanned Sections 2 and 3. For the very fast overview reader, Section 5 concludes and gives an outlook on future work. A list of abbreviations is provided before the Appendix. For more details and to follow some of the calculations, Appendix A lists results on speech-to-text performance, Appendix B lists results on text-to-concept performance, Appendix C lists the questionnaire statements of this study, and Appendix D details some validation setup views.

2. Materials and Methods

This section describes the hardware and software setup, as well as the methodology for the conduction of a human-in-the-loop simulation study to validate the benefits of an implemented ABSR prototype that was integrated with a prototypic electronic flight strip system for ATCos working within a simulated multiple remote tower environment. The technological validation exercise "006" was part of SESAR2020's wave 2 project PJ.05, "Digital Tower Technologies (DTT)" that received funding from the SESAR Joint Undertaking under the European Union's Horizon 2020 research and innovation program under grant agreement No 874470. More specifically, the exercise was conducted within solution 97, "HMI Interaction modes for Airport Tower," with its "Automatic Speech Recognition (ASR)" activity for "Improving controller productivity by ASR at the TWR CWP".

2.1. Hardware Setup of the Validation Study

Figure 1 shows the hardware setup of a prototypic CWP for a multiple remote tower environment in DLR's TowerLab [28]. Three horizontal rows of monitors (top of Figure 1) visualize the artificial outside view for the three configured airports. The airport layout is generic, but the three airports are named Vilnius, Kaunas, and Palanga.

Figure 1. Multiple remote tower environments with a row of monitors per each of the three airports under ATCo control, three radar screens, and the electronic flight strip system that is supported by the output of an assistant-based speech recognition system. The position for Vilnius is always top/left, Kaunas is middle, and Palanga is bottom/right.

The three monitors below on the desk (see Figure 1) depict the air traffic in the airport's vicinity. The touch display at the middle of the desk (see Figure 1) presents the electronic flight strips per airport per column. The ATCo wears a headset with speakers and a microphone that is triggered via a push-to-talk button at the headset's cable. The paper sheets on the left of the desk (see Figure 1) contained the airport layout, aircraft callsigns, and a legend for the symbols of the electronic flight strip system.

2.2. Software Setup and Simulation Environment of the Validation Study

All used software and displays are prototypic DLR developments. They consist of the most common elements that the usual controller working positions of European air navigation service providers offer. Thus, a wide range of ATCos from many different countries can use the systems of the validation study even if the details differ compared to their "usual" systems in daily-life operations. The aircraft and ground vehicle movements relevant to the tower and ground control were simultaneously simulated in three remote Lithuanian airports, i.e., Vilnius, Kaunas, and Palanga.

2.2.1. Outside View for Supervision of Movements on Ground and above the Airfield

The artificial outside view, such as out of a physical tower for those three airports, comprises the runway, taxiways, stands, and some environments, such as landscape and buildings, as shown in Figure 1. On the left and right side of each monitor row, there was a compass rose with additional information relevant to aircraft takeoff and landing (more details in Appendix D). If the validation condition "with ABSR support" was active, the ABSR output was also shown in the ATCo outside view.

2.2.2. Radar Displays to Monitor Air Traffic Close to the Airfield

A radar display for each of the three airports (see Figure 1 middle part) visualized the airspace structure with waypoints and the air traffic in the airport's vicinity. Each aircraft had a radar label displaying the aircraft callsign, weight category, current altitude, rate of descent/climb, speed, heading, and aircraft type. The biggest airport (Vilnius) also had a ground radar display showing the runway, taxiways, stands, and aircraft information, i.e., current and latest positions, aircraft callsign, relevant runway or stand, speed, and aircraft type, as well as a color indicating if the flight is an arrival or departure.

2.2.3. Electronic Flight Strip System (EFS)

The electronic flight strip system on the touch display consisted of one column per airport (see Figure 2). The column heads presented the airport's ICAO code, runways, automatic terminal information service (ATIS) letter, and radio frequency. Each of the three columns, in turn, comprised four different bays—air, runway, ground, and stand—in order to enable managing the flight progress in a procedural way.

Each flight strip (see zoomed white box in Figure 2) offered the option for hand written notes (pen symbol in upper left area), and showed aircraft callsign (BRU835), ICAO weight category (M), runway (34), stand (M1), estimated time of arrival/departure (08:39), aircraft type (A320), flight rules ("I" or "V" for instrument/visual flight rules), origin/destination airport (EDDK), standard instrument departure (such as BELED3D for aircraft GAF612 on the lower right blue flight strip), and squawk (3511).

The EFS for the ATCos further had a number of flight status icons on the right side (see Figure 2). The flight status icons depended on the flight intentions, i.e., blue departure flight strips/purple arrival flight strips, and on the progress, i.e., in which bay the flight strips currently are. Each flight status icon could be toggled, i.e., activated when a status change was initiated or deactivated, e.g., in case of activating by accident. The different flight status icons are shown in Figure 3. If they were activated through the tap of an electronic pen, they turned into a light green color in the electronic flight strip.

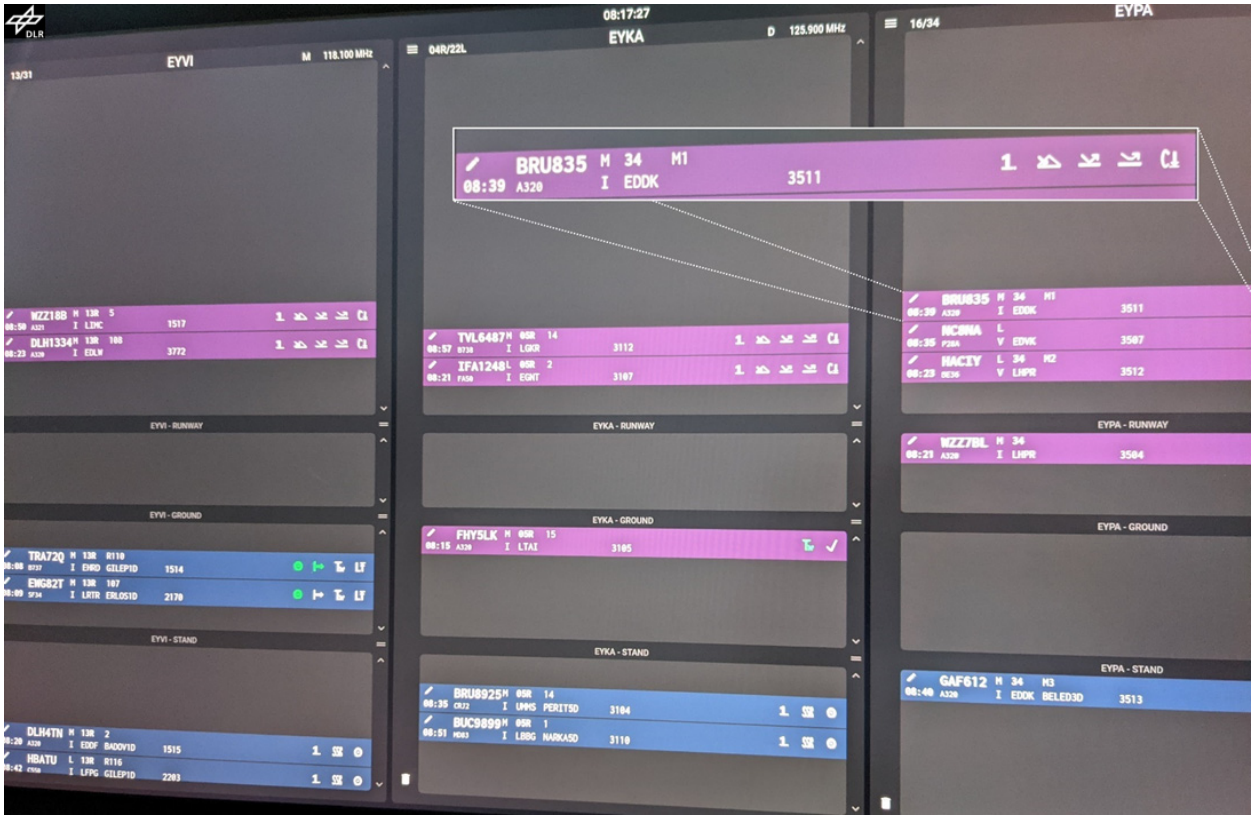


Figure 2. DLR's prototypic electronic flight strip system for aircraft at three remotely controlled airports (from left to right: Vilnius, Kaunas, Palanga).

Symbol	Name	Description
1.	FIRST_CONTACT	First radio contact established
⊙	START_UP	Aircraft has clearance for startup
↦	PUSHBACK_GIVEN	Aircraft has clearance for pushback
T↗	TAXI_OUT	Aircraft has clearance to for taxi to runway
LT	LINE_UP	Aircraft has clearance to line up on the runway
CT	TAKEOFF_CLEARANCE	Aircraft has clearance for takeoff
↑	DEPARTING	Aircraft is flying away from airport
↗	EXIT_CTR	Aircraft is leaving control zone
↙	ENTER_CTR	Aircraft is entering control zone
C↓	LANDING_CLEARANCE	Aircraft has clearance to land
↓	LANDED	Aircraft has landed
T↙	TAXI_IN	Aircraft has clearance to taxi to apron
↘	TOUCH_AND_GO	Aircraft has clearance for touch and go landing
⥿	LOW_APPROACH	Aircraft has clearance for low approach
✓	CLOSED	Flightplan has been closed
SSR	SQUAWK_SET	Transponder code has been set (event, not a state the aircraft remains in)

Figure 3. Flight status icons of electronic flight strips available depending on the current flight status [29].

The electronic flight strips changed their bays with further progress of the flight status when arriving or departing, e.g., after setting the status "LINEUP," the flight strip moved from the ground bay to the runway bay.

2.2.4. Assistant-Based Speech Recognition and Understanding Prototype

The core development for the validation study was a prototypic system for speech recognition and understanding in a multiple remote tower environment. This ABSR system is based on a number of models based on deep neural networks trained by machine learning methods, respectively. The two main steps are (1) speech recognition, i.e., automatic speech-to-text transcription from tower controller audio input, and (2) speech understanding, i.e., automatic semantic text-to-concept annotations from the transcription input (see Figure 4). The speech recognition and understanding models were trained on in-domain and out-of-domain data, specifically 200 h from seven different datasets and 4.5 h (recorded in the later study environment) of manually transcribed speech data, as well as 400 h of untranscribed data from LiveATC (Homepage: https://www.liveatc.net/ (accessed on 4 April 2023)) [30]. Further references on the development of the speech recognition engine with artificial intelligence techniques can be found in [30].

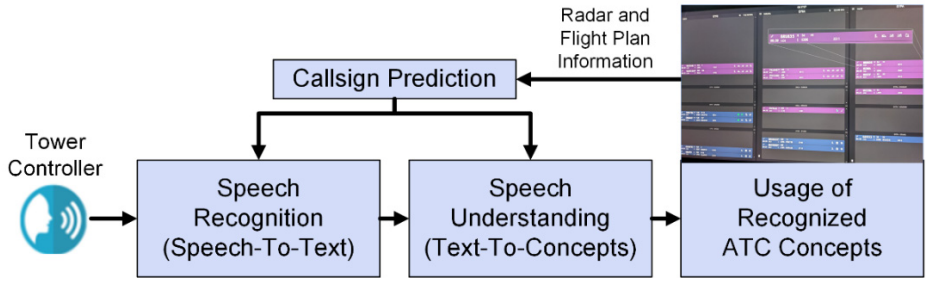


Figure 4. Components of assistant-based speech recognition (ABSR) in the multiple remote tower environment.

Both speech-to-text and text-to-concepts benefit from the use of contextual data, i.e., they consider radar data and flight plan data. The callsign prediction model is used to forecast aircraft callsigns for the next ATCo utterances, i.e., it predicts only those aircraft callsigns which are in the current area of responsibility of the ATCo. Those forecasted callsigns support the speech recognition engine in recognizing the correct word sequences and the speech understanding module in extracting the correct callsigns, especially in cases when not all words of the callsign are correctly recognized.

The command extraction model in the speech understanding module analyses the automatically transcribed ATCo utterances and extracts meaningful content, i.e., ATC concepts such as commands with callsigns, command types, values, units, etc., conform to the defined ontology. Two example transcriptions with their example annotations shall illustrate this:

- *wizz air two echo bravo good morning vilnius tower startup and pushback approved cleared to sofia* via *erlos one delta departure route seven thousand feet squawk two one seven seven QNH one zero one four*

  WZZ2EB GREETING
  WZZ2EB STATION VILNIUS_TOWER
  WZZ2EB STARTUP
  WZZ2EB PUSHBACK
  WZZ2EB CLEARED TO LBSF
  WZZ2EB CLEARED VIA ERLOS_1D
  WZZ2EB ALTITUDE 7000 ft
  WZZ2EB SQUAWK 2177
  WZZ2EB INFORMATION QNH 1014

- *hotel tango uniform when you are ready taxi to holding point runway three one correction one three right* via *[hes] golf vilnius*

  HBATU CORRECTION
  HBATU TAXI TO HP_13R WHEN READY
  HBATU TAXI VIA G C WHEN READY

The recognized ATC concepts, i.e., the annotations, are then used for highlighting purposes or supporting manual input in electronic ATC systems.

2.2.5. Visualization of ABSR Output on EFS and Outside View

The ABSR output was visible through different highlighting mechanisms in the electronic flight strips if the validation condition "with ABSR support" was active. If a callsign was recognized [31], the callsign was highlighted by displaying a rectangle in inverted colors for ten seconds at the callsign field of the flight strip (see "DLH4TN" in Figure 5). The callsign was highlighted immediately after being recognized and extracted even before the ATCo finished the utterance by releasing the push-to-talk button.

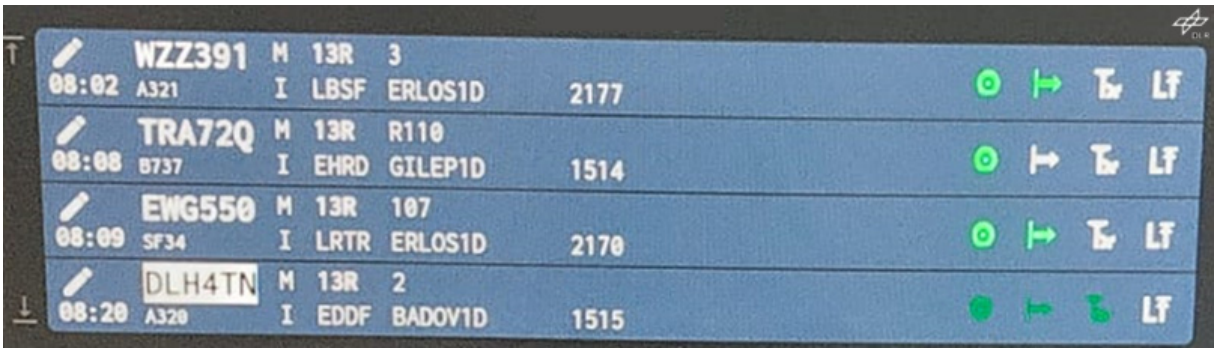


Figure 5. Prototypic electronic flight strips in the ground bay with a highlighted callsign as recognized from an ATCo utterance (DLH4TN), dark green automatically highlighted status icons for DLH4TN (STARTUP, PUSHBACK, TAXI), and five light green highlighted status icons of three other flights after being automatically accepted from the system or manually entered by the ATCo.

If one or more ATC concepts, such as commands and optionally command values, have been recognized, there was a dark green highlighting to support the ATCo in maintaining flight strips. This means the flight status icons on the right side of a flight strip or text values on the left side of a flight strip have been highlighted for ten seconds (see highlighted status icons for STARTUP, PUSHBACK, and TAXI of DLH4TN in Figure 5).

If the flight status icons in dark green mode remained unchanged by the ATCo for ten seconds, they were automatically accepted and turned into light green as with manual activation. In the case of a recognized HOLD_SHORT of runway command, the runway name was highlighted with color inversion for ten seconds as well.

2.3. ATCo Tasks in the Different Validation Conditions

Many of the tasks that ATCos needed to perform during the real-time human-in-the-loop validation study were identical under different validation conditions. Two conditions have been analyzed in the simulated multiple remote tower environment: baseline, i.e., without ABSR support and solution, i.e., with ABSR support. Section 2.3.1 describes the ATCo tasks in the baseline condition; Section 2.3.2 explains the changes induced for the ATCo when working in the solution condition.

2.3.1. ATCo Primary Tasks in Baseline Condition without ABSR Support

During the simulation runs, ATCos primarily needed to control the relevant traffic at three remote airports (tower and ground), with the above-described hardware and software setup consisting of an outside view, radar displays, and the electronic flight strip system.

Hence, they mainly gave ATC clearances, allowed for startup and pushback, instructed taxi, lineup/vacate and takeoff/landing/touch-and-go clearances for the single runway in use at each airport, as well as approved to enter/leave the control zone and to contact adjacent sectors. They also had to handle special situations on the ground with aircraft and ground vehicles being involved, such as a bird strike following a runway check and an emergency landing with the disembarkation of a sick passenger. The ATCos instructed all commands to the relevant traffic verbally in the English language via an emulated radio system.

Three simulation pilots (one for each airport) in another room communicated with the ATCo to run air and ground traffic with the support of a simulation pilot interface (see Appendix D). The ATCos were instructed to speak as usual at their working position. This also implies that some ATCos stick closer to the ICAO phraseology than others. The only continuous additional content for each ATCo utterance was the name of the station the ATCos are representing with the current utterance, i.e., "vilnius/kaunas/palanga tower," in order to fulfill safety requirements of the multiple remote tower concept.

The ATCos were asked to enter the semantic content of all utterances in terms of changed flight status into the electronic flight strip system with an electronic touch pen. Thus, they had to touch the flight status icon PUSHBACK in case they verbally instructed a pushback clearance or TAXI and the name of the taxiway if there were multiple options in case they issued to taxi via a certain taxiway (see Figure 6).

Figure 6. Prototypic electronic flight strips (blue departures; violet arrivals) in different bays (air, runway, ground, stand) with relevant information on the left (estimated time, callsign, aircraft type and weight category, flight rules, runway, destination airport, stand, departure route, squawk) and status icons on the right (e.g., CLEARED TOUCH_GO in green, ENTER_CTR, etc.).

The ontology defines 80 different command types as relevant for tower ATCos if they also include the role of ground control. All of these command types have been implemented within our command extraction algorithm.

The airport topologies were rather simple, i.e., the two smaller airports (Kaunas, Palanga) had just one taxiway each from the apron to the lineup. They vacated the single runway, and only the biggest airport (Vilnius) had two taxiway alternatives each for lining up and vacating the single runway. No runway change occurred during the simulation time. The weather conditions at all three airports remained visual meteorological conditions in the daytime throughout the simulation.

The relevant traffic in the two different one-hour simulation scenarios comprised twelve flights in Vilnius (plus two ground vehicles), six flights in Kaunas (plus one ground vehicle), and five flights in Palanga—at the latter airport, including training flights with multiple approaches—so 23 flights plus three ground vehicles (the ground vehicles make 11.5% of total relevant traffic) in total. For later evaluation, the results refer to all 26 traffic vehicles (flights plus ground vehicles) as ATC communication took place between ATCos and pilots or ground vehicle drivers, respectively. The callsigns and timing of appearance of the flights in these two scenarios were slightly different in order to reduce learning effects.

2.3.2. ATCo Tasks in Solution Condition with ABSR Support

In the solution scenario, ATCos had the same hardware setup as in the baseline scenario. The only difference was the support of the ABSR system. ATCos could majorly resign from using the electronic pen to maintain flight strips and benefit from automatic maintenance through the ABSR system, i.e., the ABSR output was used to highlight the flight status icons and callsigns in electronic flight strips automatically (see lower zoomed white box in Figure 7). The ATCos only needed to check the automatically highlighted output, i.e., representing issued commands and thus changes in the aircraft flight status, and correct if needed. A video about the simulation environment in the solution runs can be downloaded from https://www.youtube.com/watch?v=Y76kQmo_ANU&cbrd=1 (accessed on 4 April 2023). The ABSR output was only shown to the ATCos in solution scenarios. However, recording of verbal utterances, automatic transcription and automatic annotation was also performed in the background in baseline runs. The flow of using speech recognition and understanding output in the flight strips can be traced in Figure 7.

Figure 7. ATCo in front of electronic flight strip display with highlighted callsign and flight status icons, as well as outside view with transcription and annotation of ABSR output.

The complete transcription of words (first line) and the relevant annotation of commands in the agreed ontology format (second line) have been displayed in the outside view of the human-machine interface as shown in Figure 7 (zoomed white box on the upper area of the figure) if the validation condition "with ABSR support" was active.

2.4. Questionnaires and Further Tasks during and after Simulation Runs

Every five minutes, the ATCos were requested to rate their workload on a displayed graphical interface for an instantaneous self-assessment of workload (ISA) scale [32]. This interface offered values from 1 (low workload) to 5 (high workload) and appeared in the EFS system (see Figure 8).

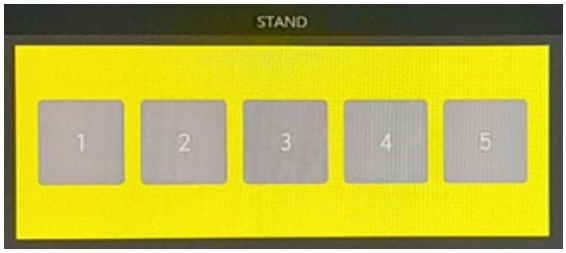


Figure 8. Instantaneous self-assessment of workload (ISA) scale to be responded to. "1" corresponds to "Under-utilized", "2" to "Relaxed", "3" to "Comfortable", "4" to "High Workload", and "5" to "Excessive Workload".

2.4.1. ATCo Secondary Tasks during Simulation Runs

Furthermore, the ATCos were asked to perform a secondary task next to their primary ATC task. After 10 and 40 min in the scenario, ATCos were requested to sort a deck of 48 cards and name one to four randomly missing cards (see Figure 9). This sorting of cards was repeated three times each or a maximum of 15 min (after 10 min) or 13 min (after 40 min), respectively. This secondary task is aimed to give a more objective impression about workload when comparing the time needed to sort and identify missing cards between baseline and solution scenarios. It is assumed that ATCos have more free cognitive capacity (less workload) if they can sort the cards quicker in one of the simulation conditions. The points in time (after 10 and 40 min) have been chosen as the ATCo workload should have been slightly increased due to the traffic situation at that time. The need to respond to ISA and to perform the card sorting remained identical in baseline and solution runs.

Figure 9. ATCo interrupts card sorting (secondary task) to check the outside view.

2.4.2. ATCo Post-Run Questionnaires after Simulation Runs

The post-run questionnaires needed to be filled by ATCos twice on each validation day, i.e., after each of the two simulation runs with the two different conditions. The well-established questionnaires cover the most important factors of air traffic controller work, such as situation awareness, workload, and trust [33] and are listed below:

- NASA-TLX (National Aeronautics and Space Administration Task Load Index) [34,35];
- Bedford Workload Scale [36];
- Three SHAPE questionnaires (Solutions for Human Automation Partnerships in European ATM) [37]:
    - AIM-s (Assessing the Impact on Mental Workload);
    - SASHA (Situation Awareness for SHAPE) ATCo;
    - SATI (SHAPE Automation Trust Index);
- CARS (Controller Acceptance Rating Scale) [38];
- SUS (System Usability Scale) [39,40].

2.4.3. Statistical Analysis Approach

When reporting the results of data that has been measured for baseline and solution runs, there will also be a statistical significance analysis, e.g., of all the above-mentioned questionnaires. Usually, there is a learning effect if ATCos perform multiple simulation runs in a row, i.e., they will perform better in the later runs, because they are used to the overall environment. Hence, better performance cannot simply be assigned to possibly different simulation run conditions such as baseline or solution. The sequence of baseline and solution runs is also an independent variable.

Therefore, two measures have been taken to compensate for the sequence effects as much as possible. First, the order of simulation runs alternate, i.e., half of ATCos start with a baseline run and end with a solution run and vice versa for the other half. The performance usually is, of course, better in the later runs, but the effect on baseline and solution runs should average out. Nevertheless, the standard deviations will be higher than they would be without sequence effects. Hence, secondly, the sequence effects will be compensated by considering the performance difference between the two runs. This sequence effect compensation technique (SECT) is described in more detail in [41]. An example shall illustrate the application of SECT. If any performance in all first runs of ATCos is 50 s and in all second runs 30 s, i.e., 20 s better, the performance difference is calculated as 50–30 = 20. Half of this difference (20/2), i.e., 10 s, is subtracted from each result of a first run and half of the difference is added to each result of a second run. Afterwards, the averages per run are the same. Furthermore, the averages of baseline and solution keep the same. We had exactly half of the ATCos having a baseline run and a solution run as the first run, respectively. However, the standard deviation will decrease, i.e., statistical significance will increase. This was already shown for earlier project result analyses such as of AcListant®-Strips when analyzing workload benefits [18].

Unpaired t-Tests can only reject hypotheses with some probability α. Therefore, the so-called null hypothesis H_0 is usually the opposite of the effect to be validated, e.g., "*ABSR support does not reduce workload as measured with a secondary task*". The test value T is calculated as the product of (1) the difference between the mean value of the performance measurement and $\mu_{0,}$ which is set to zero, and (2) the square root of the number of performance measurements, i.e., ten study subjects, divided by the standard deviation of the performance measurement. If the measurement values follow a Normal Gaussian distribution, the value T obeys a t distribution with n-1 degrees of freedom. Therefore, the resulting value T is compared with the value of the inverse t-distribution at the position $t_{n-1,1-\alpha}$ with n-1 degrees of freedom. If the calculated value T is bigger than the $t_{n-1,1-\alpha}$ threshold, we can reject the null hypothesis with probability α. As this falsifies the null hypotheses, we could assume that "ABSR support does reduce workload as measured with a secondary task." Additionally, the minimum α will be calculated, i.e.,

so that the value T threshold is still exceeded. These calculations will be performed on all single rated statements and answered questions, respectively, as well as for the group of statements/questions that belong together in a single questionnaire, e.g., the aggregating of the six items of NASA-TLX.

2.4.4. ATCo Post-Validation Overall Questionnaire

The post-validation questionnaire requested to be filled by ATCos only once after finishing all simulation runs, i.e., there is an overall rating on the ABSR prototype instead of a rating on baseline and solution each. It contained 28 statements to be rated regarding human performance, safety, operating methods, and technical feasibility. If answers on the post-validation questionnaire of the ten ATCos are reported in the following, the scale ranges from 1 (fully disagree) to 10 (fully agree), i.e., the scale mean is 5.5.

2.5. Validation Schedule and Participants

Each validation day with an ATCo began with organizational tasks such as the signature of informed consent, a briefing, and a demographics questionnaire. It was followed by 60 min training run with low to medium traffic (30 min each with baseline and solution condition, i.e., without ABSR and with ABSR support). Then, two simulation runs of 60 min each with baseline and solution conditions, respectively, and medium traffic were carried out. One run included a bird strike, and the other run included a sick passenger in an aircraft as special situations that the ATCos needed to handle and coordinate with ground vehicles. In order to average out the influence of the learning effect, baseline and solution scenarios have been alternated for ATCos throughout the validation campaign. After each run, the ATCos were requested to fill the mentioned questionnaires regarding workload, situation awareness, etc., as sketched in Section 2.4.2 and gave comments and answers in a debriefing. Finally, ATCos needed to fill out an overall tailor-made questionnaire (see Section 2.4.4) on the ABSR system after the last debriefing.

It has to be noted that the technical team of the validation campaign replaced a laptop and made a software update regarding the allowed central processing unit (CPU) load for the automatic speech recognition (ASR) engine after the eighth ATCo in the simulation campaign. However, no significant change in ABSR accuracy was noted due to this.

The validation campaign took place at DLR TowerLab in Braunschweig, Germany, from 14 February to 3 March 2022 (8:30 a.m. to 4:30 p.m.). This study was conducted with one ATCo per day for exactly ten days with five ATCos from Oro Navigacija (ON, Lithuania) and five ATCos from AustroControl (ACG, Austria). All participants were holders of an active tower ATCo license. The ten ATCos were not involved in the project in terms of participation in previous work sessions.

The nine male and one female ATCo had an arithmetic mean age of 31.9 years (standard deviation, SD: 5.5 years). The ATCos had 7.4 years of professional working experience as an ATCo (SD: 5.8 years), while ON ATCos were already longer on duty (9 years, SD: 7.3 years) compared to ACG ATCos (5.7 years, SD: 3.9 years).

3. Results

Each of the ten ATCos participated in a baseline run without ABSR support and a solution run with ABSR support, i.e., the data of twenty simulation runs with their succeeding post-run questionnaires as well as the final ten post-validation questionnaires' answers are analyzed in the following subsections. This section details:

(1) Objectively measured speech recognition performance;
(2) Objectively measured speech understanding performance;
(3) Perceived speech recognition and understanding performance;
(4) Operational and technical questions;
(5) Overall ratings on perceived workload, perceived situation awareness, satisfaction, acceptance, trust, and usability;

(6) Ratings per simulation run on perceived and more objectively measured workload, perceived situation awareness, satisfaction, acceptance, trust, and usability;
(7) General debriefing feedback.

The tailor-made statements of the questionnaires to be rated by ATCos described in the following contained the term ASR for brevity, even if automatic speech recognition and understanding was meant and experienced by the ATCos. Furthermore, the ABSR performance and the effect on subjective, as well as objective results are shown in more detail on a per-case basis by comparing ON and ACG ATCos for two reasons. First, the amount of training data differs by a factor of four between ON and ACG ATCos which influences the speech-to-text and text-to-concept performance. Second, the three controller working positions that (1) the Lithuanian ATCos are used to, (2) the Austrian ATCos are used to, and (3) is used as a prototypic environment in the simulation differ so that the familiarization with the system differs as well.

3.1. Results of Speech-to-Text Analysis

3.1.1. Audio Recordings with Transcriptions and Annotations

Verbal utterances of ATCos that were triggered with the push-to-talk button during twenty hours of simulation runs (radar data duration) have been recorded as wav-files. For each wav-file of the twenty simulation runs (baseline and solution) exists an automatic transcription and an automatic annotation. We recorded 2427 wav files with a net speech time of 4.5 h (i.e., when ATCos speak) during 20 h of radar simulation, i.e., the frequency load by ATCos was roughly 22%. The average duration per utterance was 6.6 s.

All wav-files have been manually transcribed and annotated ("gold") with DLR's Controller Command Logging Tool for Context Comparison (CoCoLoToCoCo, see Figure 10) to enable comparison and calculations about recognition and error rates on the word level and semantic level.

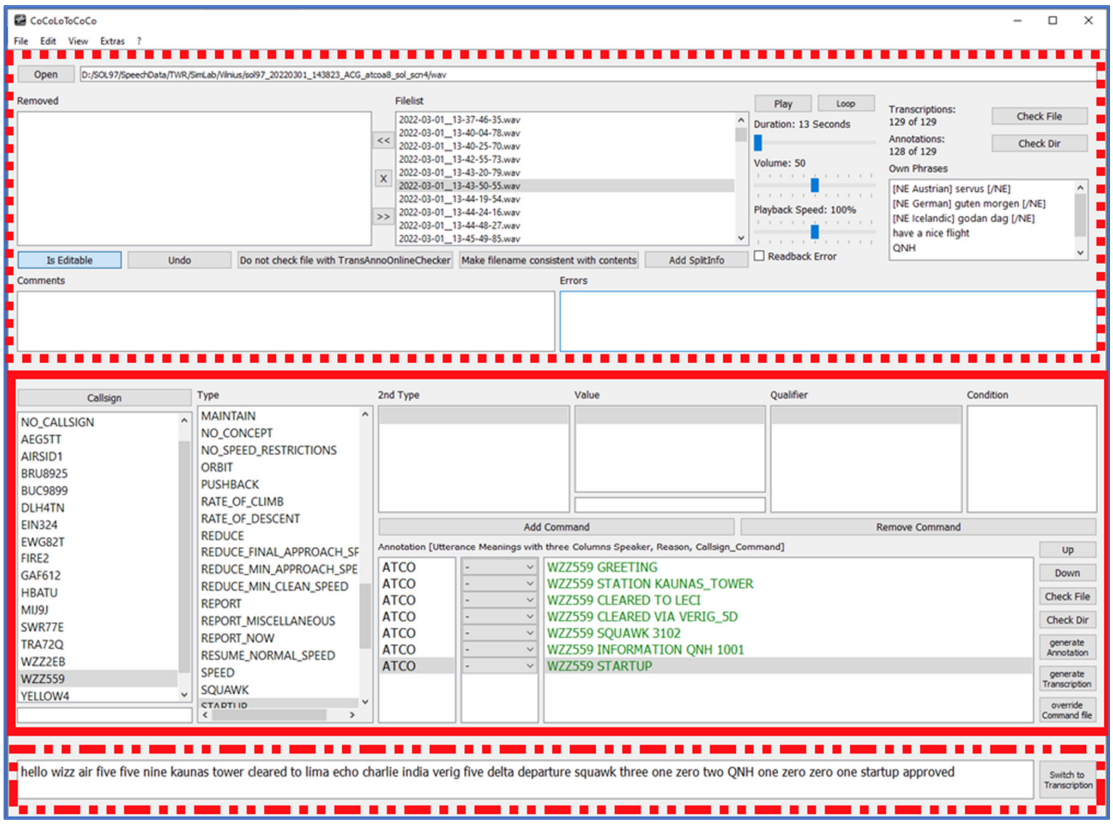


Figure 10. Software tool CoCoLoToCoCo to support transcription and annotation of ATC utterances.

The upper area of CoCoLoToCoCo (red dotted line) lists all audio files of a selected folder, has buttons and sliders to adjust the playback of the files, has a comment window and an error output window, as well as offers some further file-checking opportunities. The middle area (red solid line) shows the annotation view with a column per element of a controller command, the resulting annotation of an audio file in ontology format [9] (green font), and further buttons for rearranging and checking. The lower area (red point-dash line) visualizes the transcription of a selected audio file following defined transcription rules.

The gold transcriptions of the validation trials contain in total 37,238 words without words that are not fully uttered and thus contain a "*" such as "*lufthan**" due to our transcription rules, i.e., each ATCo utterance contains roughly 15 words. Table A3 shows the top-25 1-grams, i.e., the uttered words with their absolute and relative frequency. The most often occurring words, "one" (6.43%) and "zero" (3.97%), are usually in the top three for other ATC communication corpora as well. However, the word at rank three, "tower" (3.96%), is specific for the multiple remote tower environment, in which the transmitting entity should always be named and, therefore, appears quite often. Normally, the digits from zero to nine fill the first ten ranks in ATC communication corpora.

Furthermore, the words "*runway*," "*to*," and "*cleared*" appear in the top 12 as runway clearances and "*cleared to*" are often uttered. This latter result is confirmed by analyzing two real-life ATCo utterance corpora from Vilnius tower, as well as from Vienna tower, with roughly 7500 words in total each. This shows that the simulation setup and the challenges for the speech-to-text engine were quite realistic.

Table A4 lists the number of different words to reach a relevant portion of all uttered words, i.e., if speech-to-text performs well on the 100 most often occurring words, almost 90% of the total number of words are covered.

3.1.2. Speech-To-Text Performance

Some abbreviations that are used for analyzing purposes in the following and in the Appendices A and B are introduced:

- Onl = online (analysis as experienced by ATCos during simulation runs);
- Off = offline (analysis of audio files after the simulation runs);
- WER = Word Error Rate;
- Subs = Substitutions;
- Del = Deletions;
- Ins = Insertions;
- LevenDist = Levenshtein Distance [42] between automatic and gold transcription;

The speech-to-text accuracy is presented with details per each simulation run in the tables of Appendix A (see Tables A1 and A2). Table A1 visualizes the WER for offline recognition (Off) as evaluated after the end of the validation trials. It shows what results would be already achievable when the technical setup is improved to deliver the offline performance during the simulation runs. Table A2 visualizes the WER for online (i.e., real-time) recognition from the voice stream (Onl) as evaluated during the simulation runs, i.e., the WER are usually worse than for Off.

There were some technical problems with the ABSR setup: (1) the audio device continuously disconnected in one simulation run resulting in the loss of some data, and (2) there was partly CPU overload, especially for the first eight ATCos. The performance of the ASR engine was much worse in the online mode (as experienced by ATCos) than in the later offline analysis of recorded audio files. Worse speech-to-text performance, i.e., a higher WER being the sum of substitutions, insertions, and deletions regarding two-word sequences divided by the total number of correct words, of course also led to worse text-to-concepts performance. Some average and some specific results from these tables are analyzed deeper in the following.

The average WER for all twenty runs was 5.1% in Off mode. When just considering solution runs, the average WER even reached 4.4%, while baseline runs have an average

WER of 5.7%. When omitting the single run with audio device problems, the maximum WER was below 8% for all other 19 simulation runs in Off mode, i.e., the highest WER in that single run was 11.5%, and the lowest WER for any run was 1.3%. It needs to be admitted that the training data already contained a few speech samples from some ATCos that also participated in the final validation trials.

In Onl mode, the average WER was 13.6%, while the average WER for solution runs was 9.8% and for baseline runs 17.4% (see Table A2). There is a remarkable difference in the WER of ON ATCos (6.8%) compared to ACG ATCos (12.8%) in solution runs. This probably goes back to the amount of training data in the identical recording environment to the later validation trials, which was only 3.6 h for ON and even 0.9 h for ACG.

Four of twenty runs still achieved good performance with WER < 3%. However, three other runs that were affected by technical problems achieved a WER > 23%. Still, the Onl performance was sufficient in almost all solution runs to produce an acceptable text-to-concept quality. Nevertheless, the degradation of the speech-to-text performance is higher from offline mode to online mode than expected and offers room for improvement.

3.2. Text-To-Concept Quality

3.2.1. Description of Gold Annotation Data Set

All twenty simulation runs consist of 7560 commands (ALL), whereof 3701 are from baseline runs (BAS), and 3859 are from solution runs (SOL), respectively. Hence, there were 3.1 commands per ATCo utterance and 5.1 words per command if we assume that all words of an utterance are relevant to form a command.

However, it has to be noted that there are some word sequences annotated as commands that do neither influence the aircraft status nor include any request, report or traffic information from the ATCo side:

- First, the annotations GREETING (e.g., "hello"), FAREWELL (e.g., "bye"), and NO_CONCEPT (e.g., "thanks;" no relevant ATC command in the utterance) that are summing up to 9.8% of commands during this study. These command types can indicate that the ATCo workload might not be assumed as overwhelmingly high if they still have time for welcoming, saying goodbye, and thanking anybody.
- Second, the annotation CORRECTION and CALL_YOU_BACK (e.g., "standby") that sum up to 1% of the commands might indicate a higher workload as ATCos often correct themselves, are asking for repetition of the transmission or are telling to wait for further information. The annotation SAY_AGAIN, which also belongs to this command group, has not been used.
- Third, the annotation AFFIRM and one annotation of DISREGARD that sum up to 4.1% of the commands have ATC communication relevant content, even if they are no commands in a classical sense. The annotation NEGATIVE, that also belongs to this command group, has not been used.

Though, the above-listed annotations enable a workload analysis of human ATC operators that will be published in another paper. 15 of the 80 possible command types for tower ATCos as defined in the ontology, such as GO_AROUND and ABORT TAKEOFF, did not occur at all in the 7560 commands. This means 65 different command types have been used by the ten ATCos, e.g., PUSHBACK, TAXI TO, CLEARED TAKEOFF/LANDING, ENTER_CTR, etc. Table A5 lists the relative occurrence of all command types greater than 1%. The last type, "others", groups all command types that occurred between 0.33% and 1%, such as CONTACT, ENTER_CTR, LINEUP_BEHIND, CLIMB, and DIRECT_TO. In total, there are 36 different command types that appeared more than 25 times, i.e., more than 0.33%.

The most often used command type is—unsurprisingly—STATION, as ATCos were asked to utter it in each radio transmission. However, 1529 occurrences (20.2% of commands) in 2427 utterances mean that ATCos did not follow this multiple remote tower safety-related request in 37% of all utterances. This might not be critical if ATCos just uttered "bye," but in any case, it should be considered for the multiple remote tower

concept. The (CONTINUE) TAXI TO/VIA commands sums up to 11.5% of commands. The INFORMATION WINDSPEED/DIRECTION even sum up to 15% of the commands as they were instructed for all takeoffs and landings/touch-and-gos. The exclusive runway clearances CLEARED TAKEOFF/LANDING/TOUCH_GO/VISUAL sum up to 6.8% of commands. The runway usage clearances LINEUP, LINEUP_BEHIND, VACATE (VIA), and BACKTRACK sum up to 4% of commands.

A total of 29 of those 65 used command types occurred a maximum of 25 times for all ATCos in total such as BACKTRACK, CLEARED VISUAL, HOLD_SHORT, JOIN_TRAFFIC _CIRCUIT, LEAVE_CTR VIA, and ORBIT. For the above considerations, we neglect that only 87% of all words that are available in the gold transcriptions have been used by the automatic command recognition algorithm to classify commands (see column "*Unknown Classified Rate*" in Tables A6, A8 and A10).

It needs to be mentioned that our prototype follows a more holistic approach than some very basic prototypes of other actors in the field of speech recognition and understanding [43]. Our command extraction algorithm does not only extract callsigns (DLH4TN), basic types (TAXI), and values, but more sophisticated command types of multiple parts (TAXI TO/VIA), units, qualifiers, conditions (WHEN READY), chain commands with multiple callsigns, tackles many types of corrections through the ATCo and even robustly recognizes elements of the ontology if there are minor and major (acceptable) deviations from ICAO phraseology [44] in the utterances. Furthermore, we support a bigger number of command types (from the agreed ontology) as defined by the different actors themselves. The execution time of the command extraction per utterance in offline mode on a standard laptop, i.e., on a complete transcription, has an arithmetic mean of 2 ms and a median of 1.2 ms with a minimum execution time below 0.1 ms and a maximum execution time below 40 ms independent of performing command extraction on gold, offline or online transcription files. In addition, our prototype is—to the best of our knowledge—the first to support multiple remote towers at the same time (not just one) and delivers recognition error rates on an acceptable level despite all the above-mentioned complex add-ons.

3.2.2. Description of Results of Automatically Extracted Commands on Different Versions of Speech-To-Text Transcriptions

The following three subsections present recognition and error rates on callsign and command level, as well as the portion of words from the utterances that have not been used for ATC concept extraction while referring to Appendix B. More details on the semantic level metrics can be found in [45]. The command extraction results will also be presented by comparing the different command type groups:

- "All;"
- "Relevant" if appearing more than 25 times in all 20 runs;
- "EFS" has a visible effect on the electronic flight strips;
- "Status" that changed the aircraft status in the electronic flight strips;
- "Outside" is just shown on the monitors for the outside view;
- "Hypo-EFS" could have been highlighted in the flight strips but have not been during the trials, such as recognizing the active runway in an utterance.

3.2.3. Speech Understanding Performance on Gold Transcriptions

In total, 65 different command types have been automatically extracted from the gold transcriptions, i.e., the same number as in gold annotations. Table A6 shows how well the ontology-conform automatic recognition of ATC commands is modeled. The command recognition rate is around 96% with an error rate below 2.5%; the rejection rate (not reported herein) causes a difference to 100% in the total sum of command rates. The callsign recognition rate even achieved 99.8% with an error rate of 0.2%. The command recognition rates in solution runs were 96.6% for ON and 95.4% for ACG.

A total of 18.3% of all problematic annotations (recognized commands) go back to the three ground vehicles in the scenario that make up 11.5% of all relevant traffic. Further,

7.3% of problematic annotations go back to the emergency aircraft, even if this makes up 3.8% of the flights.

18 of the 80 defined command types from the ontology had visible effects in the flight status icons of the electronic flight strips—hereinafter referred to as command type group *Status*. Three further commands had a visual effect on the textual data of the electronic flight strips. These 21 commands that influenced the appearance of the electronic flight strips are grouped in the command type group *EFS*. Three supported commands contained weather information from the *Outside* view (QNH, INFORMATION WINDDIRECTION and WINDSPEED); the values of four further supported commands could have been displayed in the relevant field of the electronic flight strip. However, this highlighting has not been fully implemented yet (command group *Hypo-EFS*), i.e., STATION, INFORMATION ATIS, INFORMATION ACTIVE_RWY, and HOLD_SHORT for all possible airfield elements such as taxiways. The command type group *Relevant* includes all commands that have been automatically extracted more than 25 times. Table A7 shows the command recognition performance on the above-mentioned command type groups, i.e., presenting command recognition rates of 96% and more.

3.2.4. Speech Understanding Performance on Offline Transcriptions

The command recognition results of Table A8 are based on the output of the speech recognition engine, i.e., the transcription from Off mode. The command recognition rate is above 91%, with an error rate below 5%. The callsign recognition rate achieved almost 98.5% with an error rate below 1%. The command recognition rate of command type group *EFS* is beyond 93%, as Table A9 shows. 16.2% of all problematic annotations go back to the three ground vehicles that comprise 11.5% of all relevant traffic.

3.2.5. Speech Understanding Performance on Online Transcriptions

Tables A10 and A11 present the command recognition results on transcriptions from Onl mode. The command recognition rates are roughly 10% worse than in Off mode. The command recognition rate for solution runs in which the ATCos saw the ABSR output was 82.9%, with an error rate of 6.6%. However, there is a huge difference in the command recognition rate for ON ATCos (88.0% based on WER of 6.8%) compared to ACG ATCos (77.7% based on WER of 12.8%). As the command recognition rates for ON and ACG ATCos were both close to 96% on gold transcriptions, the high WER resulting from the mentioned low amount of available training data was a major impact on the online command recognition next to some deviations of ATCos from ICAO phraseology. The online callsign recognition rate achieved 94.2% with an error rate of 2.4%. This again shows the influence of the high WER on the ATC concept extraction.

The following measurements, especially the questionnaire ratings of ATCos, are based on the Onl mode, as this performance was "experienced" by ATCos during simulation runs.

3.2.6. Subjectively Perceived Speech Recognition and Understanding Performance and Functionality (Post-Validation)

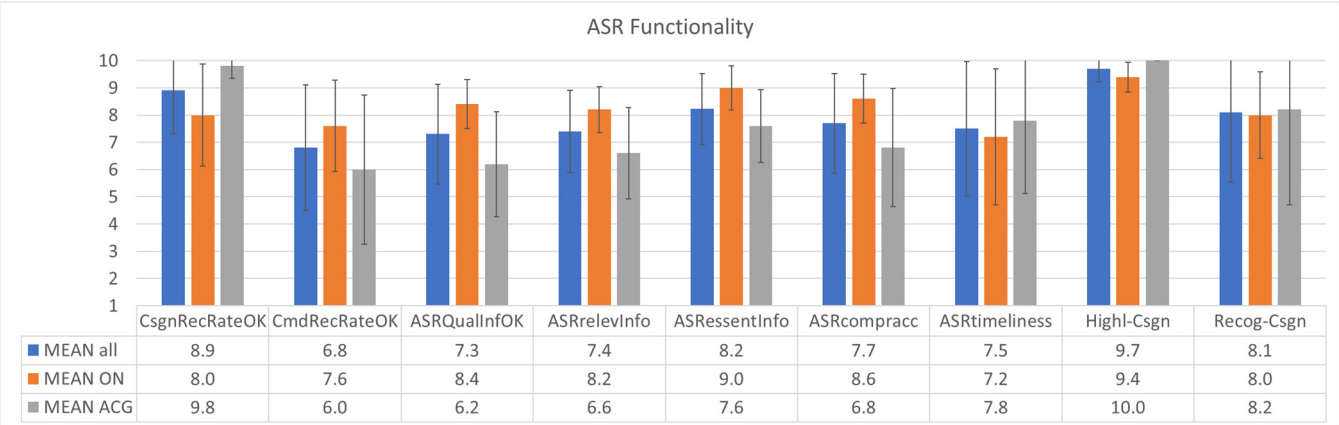
The post-validation questionnaire contained nine statements about technical feasibility with respect to the recognition and error rate of callsigns and commands as well as the ASR functionality:

1. The recognition rate and recognition error rates for callsigns by ASR were at an acceptable level. [CsgnRecRateOK];
2. The recognition rates and recognition error rates for commands by ASR were at an acceptable level. [CmdRecRateOK];
3. Overall, the level and quality of information provided by ASR were an acceptable level. [ASRQualInfOK];

The post-validation questionnaire contained four statements about the ASR interface:

4. The ASR tool interface (HMI) provides suitable access to relevant information in all situations. [ASRrelevInfo];
5. The ASR tool interface (HMI) does not display any non-essential information (clutter). [ASRessentInfo];
6. The ASR tool display is both comprehensible and acceptable. [ASRcomprehaccep];
7. The timeliness of the ASR tool display is within acceptable limits. [ASRtimeliness];
8. Automatic Speech Recognition (ASR) highlighting aircraft callsigns in the electronic flight strip display technically worked well. [Highl-Csgn];
9. Automatic Speech Recognition (ASR) highlighting aircraft callsigns in the electronic flight strip display supports recognizing which aircraft callsign has been (speech) recognized quickly. [Recog-Csgn].

The results are shown in Figure 11. ATCos rated the recognition of callsigns as almost perfect, with a mean value of around 9 on a scale from 1 to 10. The recognition rates of ATC commands were also perceived as good, with a mean value of around 7. The general quality level of information presentation from ASR was rated to be at an acceptable level with a mean value of slightly beyond 7. It has to be noted that the command recognition and overall ASR information displayed were rated much higher from ON than from ACG ATCos. This is most probably due to the underlying WER of 13% for ACG ATCos and 7% for ON ATCos, which is, however, still improvable to reach the 4% WER of offline analysis. Relevant information about the ABSR system can be assessed (mean value 7.4, but more than 1.5 points rated higher by ON than by ACG). The ASR tool seems to only present essential information with a mean value of 8.2 (again, ON rated almost 1.5 points higher than ACG). The ASR visualization is perceived as comprehensible with a mean value of 7.7 (again, ON rated almost 2 points higher than ACG). Finally, the output of the ABSR system was shown timely (mean value 7.5) due to the ATCo feedback.



	CsgnRecRateOK	CmdRecRateOK	ASRQualInfOK	ASRrelevInfo	ASRessentInfo	ASRcompracc	ASRtimeliness	Highl-Csgn	Recog-Csgn
MEAN all	8.9	6.8	7.3	7.4	8.2	7.7	7.5	9.7	8.1
MEAN ON	8.0	7.6	8.4	8.2	9.0	8.6	7.2	9.4	8.0
MEAN ACG	9.8	6.0	6.2	6.6	7.6	6.8	7.8	10.0	8.2

Figure 11. Subjective ATCo ratings on ASR accuracy and functionality.

The highlighting of callsigns in the electronic flight strip display (*Highl-Csgn*) was perceived as working technically very well, with a mean of 9.7 on a 10-point scale and a low standard deviation of 0.5. The second statement *Recog-Csgn* rated with a mean value of 8.1, helped the ATCos to detect which aircraft callsign has been recognized by the ABSR system. This information is needed to decide whether the following recognized ATC commands are highlighted for the correct callsign. The interesting part of these answers is the comparison with the objective measurements, i.e., the online callsign recognition rates, which are 92.1% for Lithuanian ATCos and 91.3% for Austrian ATCos (see Table A10). The same applies to the callsign recognition error rates, which are 3.9% for ACG, and also much higher than the 2.4% for ON ATCos. We have no real answer for this discrepancy between subjective rating and objective measurement.

3.3. Answers to Subjective Post-Validation Questionnaires

3.3.1. Operational Use of ASR (Post-Validation)

The post-validation questionnaire contained five statements about the operational feasibility of the ASR system:

1. I can apply operating methods in an accurate, efficient, and timely manner with ASR. [AccOpMeth];
2. I think that operating methods are clearly identified and consistent in all operating conditions. [OpMethConsis];
3. Procedures and operating methods are acceptable when using the ASR tool. [ProcOKwASR];
4. There are no changes needed to current working methods/procedures to fully support the use of the ASR tool. [NoChgNeed];
5. The ASR tool would be operationally acceptable under either nominal or non-nominal conditions. [OpAccAllCond].

The results are shown in Figure 12. The operating methods with ASR seem to be accurate, efficient, timely, and consistent in different conditions, with mean values of 8 and 7.4, respectively. Procedures and operating methods seem to be fine, with a mean value of 8.5 and a standard deviation of only 1.0. There are some changes to current working methods needed to fully support the use of the ASR tool, as the mean value equals the scale mean value of 5.5. However, ON ATCos rated this statement with almost 7, while ACG ATCos rated it with slightly above 4 points. The ASR seems to be operationally acceptable under different conditions, most probably under the majority of nominal and a few non-nominal conditions, as the ATCo rating was just slightly beyond the scale mean value.

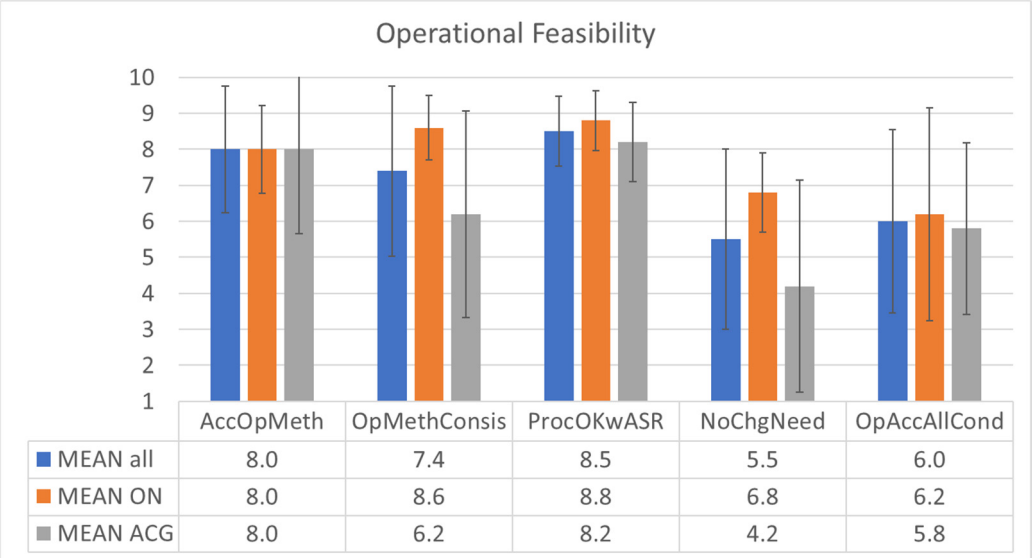


Figure 12. Subjective ATCo ratings on operational feasibility and operating methods.

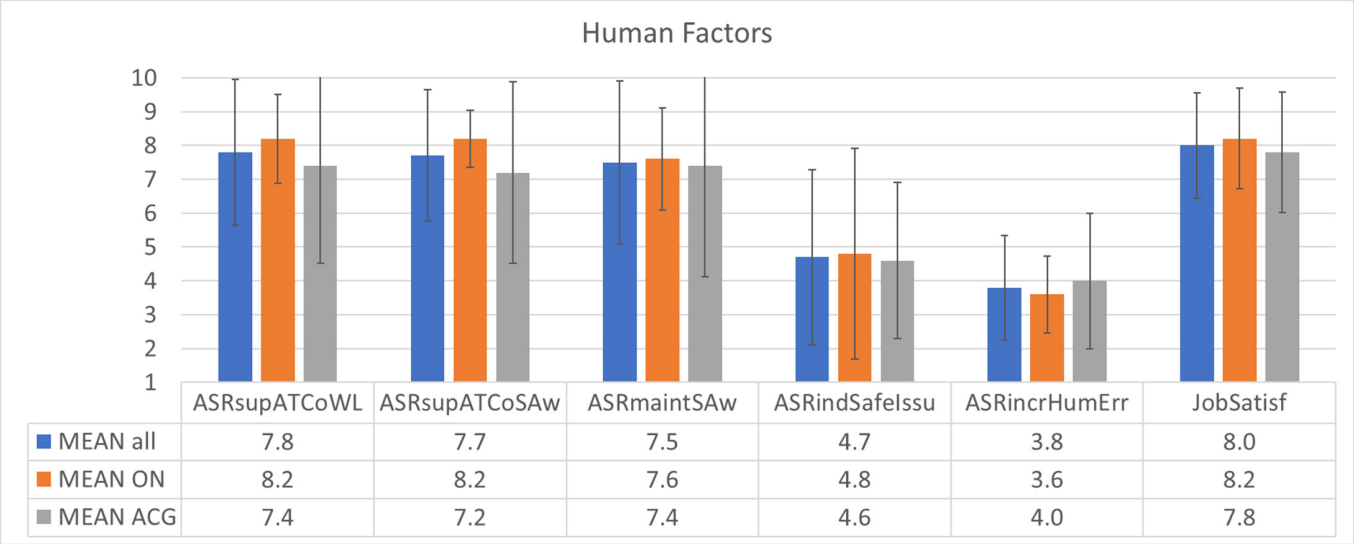
3.3.2. Human Factors Questions (Post-Validation)

The post-validation questionnaire contained six statements on human factors:

1. I think that ASR supports me in maintaining my workload at an acceptable level. [ASRsupATCoWL];
2. I think that ASR supports me in maintaining an adequate level of situational awareness. [ASRsupATCoSAw];
3. My situational awareness is maintained at an acceptable level with Automated Speech Recognition (ASR). [ASRmaintSAw];

4. I see many safety-related issues to be solved regarding automatic speech recognition implementation. [ASRindSafeIssu];
5. I think that ASR did increase the potential for human errors. [ASRincrHumErr];
6. Overall, I was satisfied performing my task with ASR. [JobSatisf].

The results are shown in Figure 13.



	ASRsupATCoWL	ASRsupATCoSAw	ASRmaintSAw	ASRindSafeIssu	ASRincrHumErr	JobSatisf
MEAN all	7.8	7.7	7.5	4.7	3.8	8.0
MEAN ON	8.2	8.2	7.6	4.8	3.6	8.2
MEAN ACG	7.4	7.2	7.4	4.6	4.0	7.8

Figure 13. Subjective ATCo ratings on human factors.

ASR seems to support maintaining situation awareness and workload of ATCos at an acceptable level with mean values of 7.5 and beyond on a 10-point scale. The *ASRsupATCoWL* statement was rated with 7.8 on a 10-point scale (90% of ATCos rated this item with 7 or above). The *ASRsupATCoSAw* statement was rated with 7.7 on a 10-point scale (90% of ATCos rated this item with 7 or above). The statement, if ASR induced safety issues or increased the potential for human errors, was rated with mean values below the scale mean of 5.5. ATCos rated their job satisfaction with using ASR high (mean value of 8 on the 10-point scale).

3.3.3. Acceptance (Post-Validation)

The post-validation questionnaire contained three statements about acceptance of and trust in the ASR system:

1. I think that the ASR system is adequately usable. [ASRadequse];
2. I would accept such an ASR system in my future tower CWP. [ASRacceptCWP];
3. My trust in the ASR system is at an acceptable level. [ASRtrust].

The results are shown in Figure 14. ATCos rated the adequate usage of ASR with a mean value of around 7. However, it has to be noted that it was rated much higher by ON than by ACG ATCos. All ATCos would accept such an ASR system in their future tower CWP with a mean value of 7.5. They trusted the ASR system with a mean value of around 7.

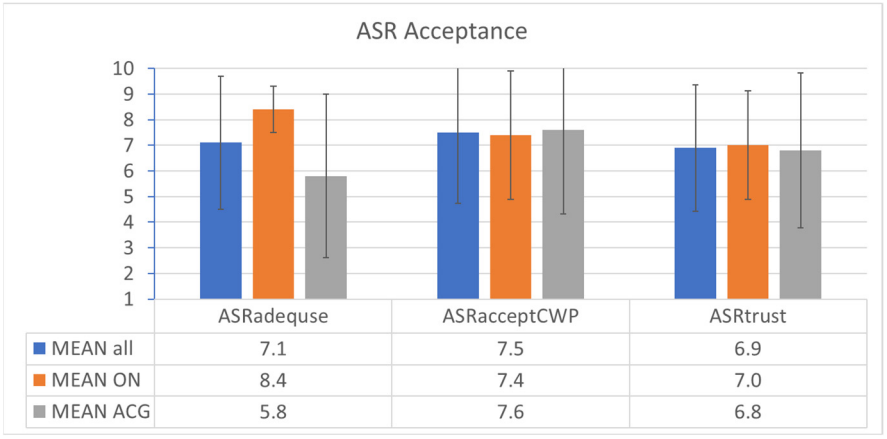


Figure 14. Subjective ATCo ratings on technical ASR acceptance.

3.4. Answers to Subjective Post-Run Questionnaires

3.4.1. Controller Acceptance Rating Scale (CARS) (Post-Run)

The post-run questionnaires contained the CARS statement to be rated on a scale from 1 to 10, with 10 being the best value, as listed in Appendix C.1. The results of the CARS questionnaire are shown in Figure 15. The acceptance was, on average, 0.6 points higher on the CARS scale for the baseline condition compared to the solution. The CARS questionnaire was filled out by each ATCo twice, once after the run with ABSR support and once after the run without ABSR support. Therefore, we are able to perform a paired *t*-test. After compensating sequence effects, the α was 0.1 to reject the inverse hypothesis that ABSR support reduces the controller acceptance due to CARS. The absolute value was 6.8 versus 6.2 (0.8 points higher for ON on average and 0.8 points lower for ACG on average).

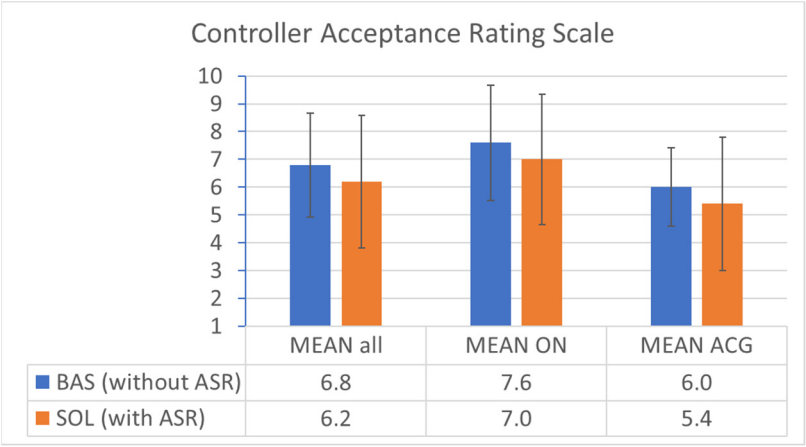


Figure 15. Subjective ATCo ratings on CARS.

3.4.2. Trust (SATI) (Post-Run)

The post-run questionnaires contained the six statements of SATI, as listed in Appendix C.2. The seven-item answer scale ranged from "Never, Seldom, Sometimes, Often, More Often, Very Often, and Always." To present the results in a bar diagram, "Never" is translated to 0%, "Seldom" to 1/6 %"…"Very Often" to "5/6 %" until "Always" to 100%. The results are shown in Figure 16.

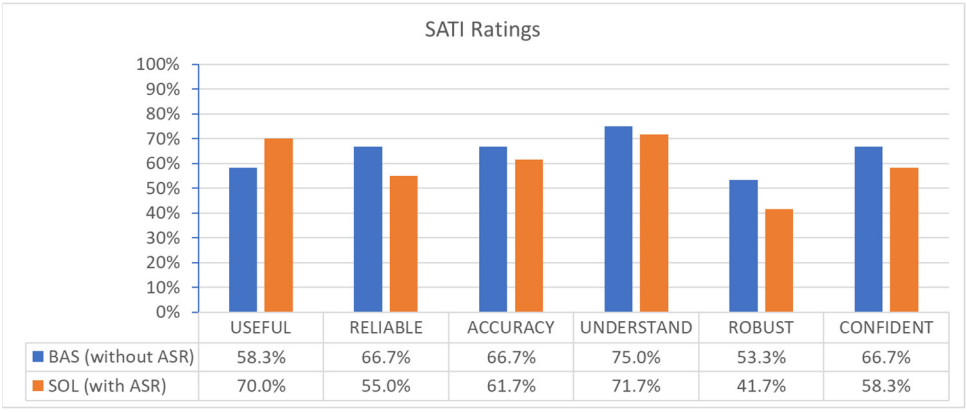


Figure 16. Subjective ATCo ratings on SATI questionnaire.

ABSR support reduced trust in automation due to SATI (α = 0.25). However, the usefulness of the system (*USEFUL* in Figure 16) was rated much better for SOL than for BAS (α = 0.05). The other five mean values are better for BAS than for the SOL condition. It is noteworthy that the four statements *RELIABLE*, *ACCURACY*, *UNDERSTAND*, and *ROBUST* from ON ATCos have better ratings for SOL than for BAS condition on average. The ambivalence of results will be discussed in Section 4.

3.4.3. Perceived Situational Awareness (SASHA ATCo) (Post-Run)

The post-run questionnaires contained the six statements of the SASHA ATCo, as listed in Appendix C.3. The seven-item answer scale ranged from "Never, Seldom, Sometimes, Often, More Often, Very Often, and Always." To present the results in a bar diagram, "Never" is translated to 0%, "Seldom" to 1/6 %"…"Very Often" to "5/6 %" until "Always" to 100%. The results are shown in Figure 17.

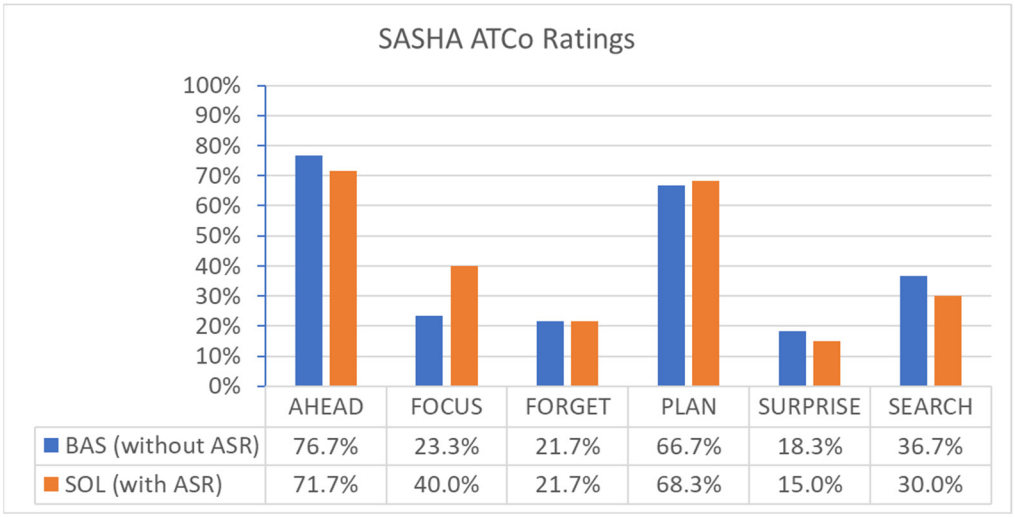


Figure 17. Subjective ATCo ratings on SASHA ATCo questionnaire.

ABSR support reduced the situation awareness of ATCos due to SASHA (α = 0.33). However, "searching for information" was less needed in the SOL condition (α = 0.15). The mean values of the first two items, *AHEAD* and *FOCUS*, are better for BAS than for SOL conditions. The mean values of the last four items, *FORGET*, *PLAN*, *SURPRISE*, and *SEARCH*, are equal or better for the SOL condition compared to the BAS condition without analyzing standard deviations, as differences in mean values are rather small.

3.5. Perceived Workload (High Workload Contribution) (Post-Run)

The post-run questionnaires contained a free-text question about high workload: "Which factors/events/conditions have contributed to potentially high workload?".

The structured answers and the number of ATCos noting this after each conducted simulation run (multiple notions in one questionnaire answer possible) were as follows:

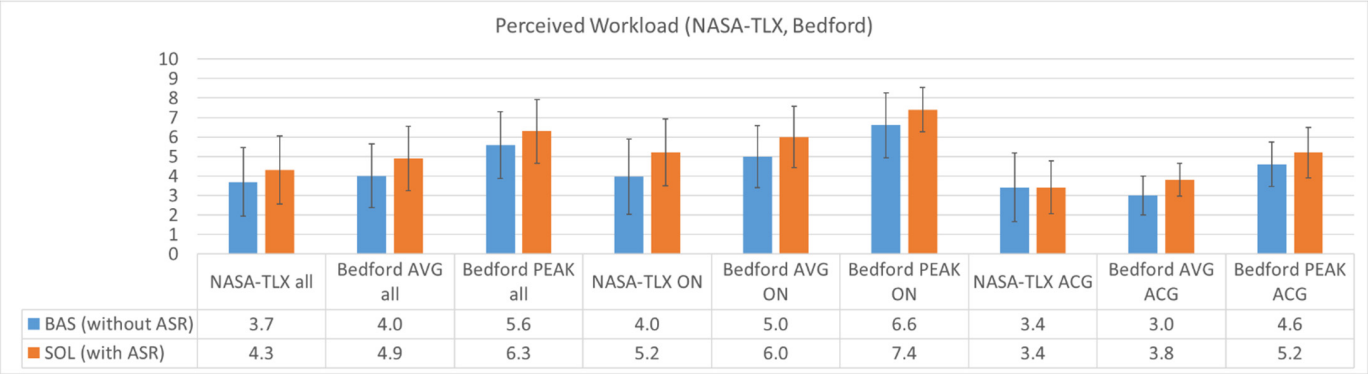
- New/unknown airspace/airport layout (especially multiple remote towers): 15 times;
- New/unknown equipment/hardware/software/electronic flight strips: 7 times;
- Checking of ABSR output (only in solution condition): 4 times;
- Unexpected/unusual air traffic situations: 3 times;
- Other: Secondary task (2 times), tower view/runway perspective (2 times), slightly different phraseology to always name the calling tower (2 times), miscommunication, system errors.

Interpreting the above results, 15 of 20 ATCo answers stated that the unknown multiple remote tower environment with unknown airport layouts induced a higher workload. Furthermore, many ATCos remarked that the flight strip handling was difficult (as some details were different from "home"). This means that the majority of workload-increasing factors can be assigned to environmental aspects that should normally not be tested in the ABSR validation trials. The above-listed checking of ABSR output, as well as unexpected situations and some further aspects, seem to have been only a minor factor for the higher workload.

3.6. Perceived Workload (NASA-TLX and Bedford Workload Scale) (Post-Run)

The post-run questionnaires contained the six statements of NASA-TLX (National Aeronautics and Space Administration—Task Load Index) as listed in Appendix C.4 and the two statements of the Bedford workload scale to rate the average workload (AVG) and peak workload (PEAK) on a scale from 1 to 10 with 10 being the highest workload. In addition, the 15 pair-wise comparisons of workload contributing factors (as the other part of the weighted NASA-TLX questionnaire) were assessed with ATCos once.

The results of the weighted NASA-TLX and the Bedford workload scale are shown in Figure 18. Figure A1 in Appendix C shows the weight per each of the six dimensions for NASA-TLX, which is almost equally distributed except for more weight for mental workload than for physical workload. The overall weighted workload (OW) due to NASA-TLX was higher for the solution than for the baseline condition: 43.1 and 38.9 ($\alpha = 0.02$), respectively, with huge standard deviations around 17.5. However, the general difference between baseline and solution was only induced by the ON ATCo ratings, as the OW for ACG remained the same in baseline and solution.



	NASA-TLX all	Bedford AVG all	Bedford PEAK all	NASA-TLX ON	Bedford AVG ON	Bedford PEAK ON	NASA-TLX ACG	Bedford AVG ACG	Bedford PEAK ACG
BAS (without ASR)	3.7	4.0	5.6	4.0	5.0	6.6	3.4	3.0	4.6
SOL (with ASR)	4.3	4.9	6.3	5.2	6.0	7.4	3.4	3.8	5.2

Figure 18. Subjective ATCo ratings on NASA-TLX (Weighted Overall Workload).

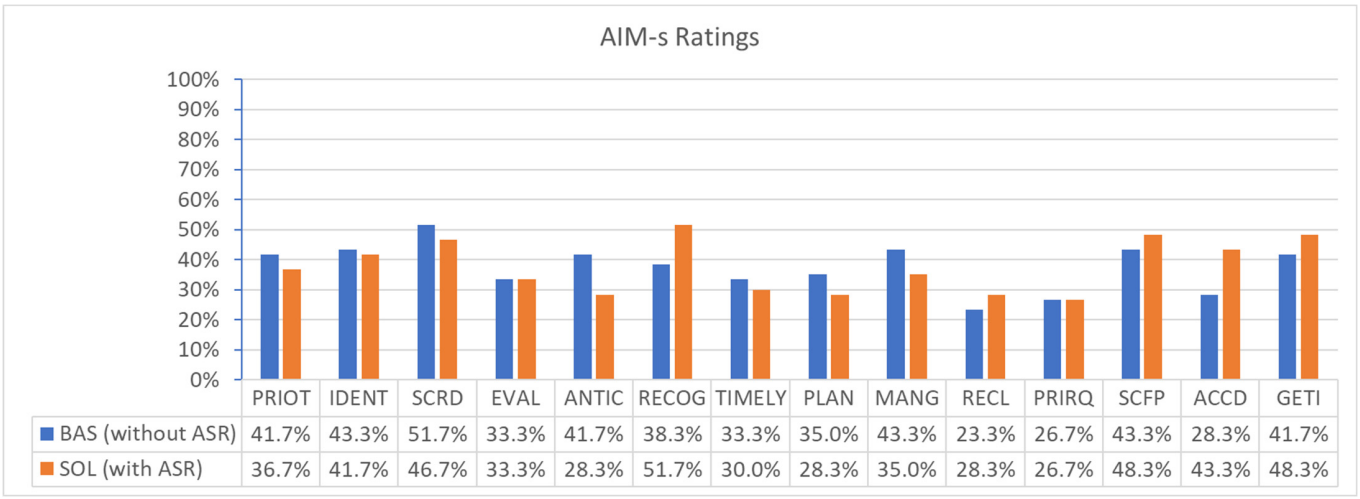
Furthermore, a clear learning effect during the validation day in terms of NASA-TLX OW can be seen. Those five ATCos who started with a baseline, rated the baseline (their first run) with an OW of 41.9; those five ATCos who started with a solution, rated the

baseline (their second run) with an OW of 32. Those five ATCos who started with the solution, rated the solution (their first run) with an OW of 48.9; those five ATCos who started with baseline, rated the solution (their second run) with an OW of 37.2.

The average and peak Bedford workload were 0.9 and 0.7 points higher, respectively, in the solution condition with ABSR support compared to the baseline condition ($\alpha = 0.001$). The peak workload was roughly 1.5 points higher than the average workload. The workload level, in general, was roughly two points lower for ACG than for ON ATCos.

3.7. Perceived Workload through Automation Impact (AIM-s) (Post-Run)

The post-run questionnaires contained the sixteen statements of AIM-s as listed in Appendix C.5. The seven-item answer scale ranged from "None, Very Little, Little, Some, Much, Very Much, Extreme." To present the results in a bar diagram, "None" is translated to 0%, "Very Little" to 1/6 %"…"Very Much" to "5/6 %" until "Extreme" to 100%. The statements SHARE and TMN are not analyzed further as there were no team members during the simulation runs (fourteen statements remain). Figure 19 shows the results.



	PRIOT	IDENT	SCRD	EVAL	ANTIC	RECOG	TIMELY	PLAN	MANG	RECL	PRIRQ	SCFP	ACCD	GETI
BAS (without ASR)	41.7%	43.3%	51.7%	33.3%	41.7%	38.3%	33.3%	35.0%	43.3%	23.3%	26.7%	43.3%	28.3%	41.7%
SOL (with ASR)	36.7%	41.7%	46.7%	33.3%	28.3%	51.7%	30.0%	28.3%	35.0%	28.3%	26.7%	48.3%	43.3%	48.3%

Figure 19. Subjective ATCo ratings on AIM-s questionnaire.

After compensating sequence effects, the overall perceived workload due to AIM-s is not statistically better with or without ABSR support. We measured an α of 0.49, which is not better than throwing a coin. However, the anticipation of the future air traffic situation was much better for SOL than for BAS ($\alpha = 0.02$). Nine of the fourteen statements have been rated better on average (less) for the SOL condition than for the BAS condition. Only the five statements related to information *RECOG*, *RECL*, *SCFP*, *ACCD*, and *GETI* have been rated worse for SOL condition compared to BAS condition.

3.8. Perceived Workload (Instantaneous Self-Assessment of Workload (ISA)) (Within-Run)

During each simulation run, ATCos needed to rate their workload of the recent five minutes on a scale from 1 (bored) to 5 (almost overloaded). The results are shown in Figure 20. The average ISA workload was 0.1 points less, i.e., better, in solution condition with ASR support compared to baseline condition with $\alpha = 0.15$ (2.1 and 2.0 points, respectively).

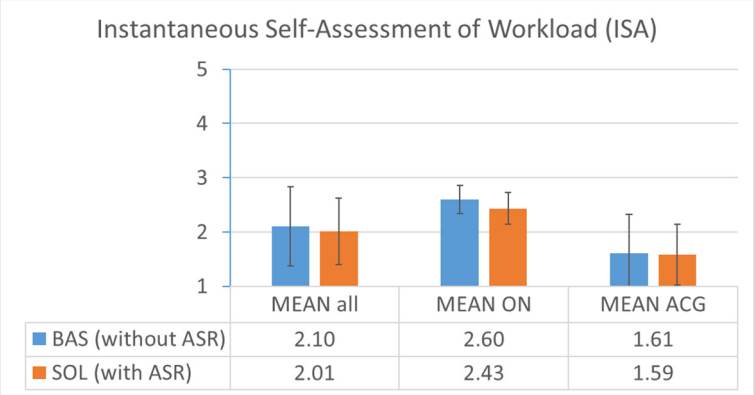


Figure 20. Subjective ATCo workload self-assessment (ISA).

The ISA of ON ATCos was on a higher level with 2.6 and 2.4, respectively, and had a much lower standard deviation of below 0.3. The ISA score of ACG ATCos was around 1.6, with a standard deviation more than twice as much as of ON ATCos.

3.9. Objectively Measured Workload with Secondary Task (Card Sorting) (Within-Run)

The ATCos always needed to make sure that their primary task of doing ATC remains safe and efficient. However, if they had time for a secondary task, i.e., free mental capacity, they should sort cards. This method has already been used in earlier ASR projects to generate a more objective measure of mental workload than just via self-ratings.

ATCos needed to sort 48 cards of a German Doppelkopf deck into six decks (Aces, Kings, Queens, Jacks, Tens, and Nines). In the beginning, all 48 cards are on one stack, with the picture side of the cards looking downwards. Each card needed to be turned around in a single move with just one hand to put it onto the correct of the six decks. After sorting, ATCos should name one to four randomly missing cards that the supervisor took out of the 48 cards deck prior to starting sorting. If there was an error in naming the missing cards, e.g., not all missing cards are named, ATCos must try again until all missing cards are named correctly. The time measurement in seconds started when the deck of 48 cards was put next to the electronic flight strip display. The time measurement ended when all missing cards were named correctly. Sorting cards were trained once in each of the thirty minutes training runs. Card sorting in the baseline and solution runs started after 10 min (for at least 15 min or at least three rounds) and again after 40 min (for at least 13 min or at least three rounds). Those time frames comprised higher traffic density to measure any difference in workload through ASR support.

The results are shown in Figure 21. ATCos finished their secondary task 8% slower in baseline runs when not being supported by ASR (395 s vs. 364 s with a standard deviation of 305 s and 262 s). This difference was 9% for ON and 7% for ACG ATCos. When compensating sequence effects with the SECT technique, ATCos were even 9% slower in baseline runs compared to solution runs. After compensating sequence effects, the α was 0.24 to reject the hypothesis that ABSR support does not reduce the workload of ATCos.

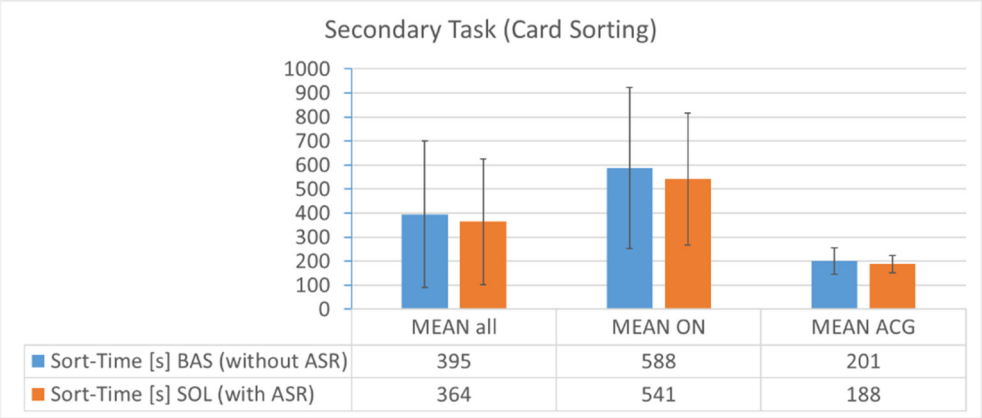


Figure 21. ATCo performance in the secondary task (card sorting).

When translating the timing results into workload, again, ON ATCos experienced a higher workload level (around 9 min sorting average) than ACG ATCos (around 3 min sorting average with more task repetitions than ON ATCos), but workload in solution condition seems to be lower than in baseline regarding the secondary task of card sorting. Additionally, the secondary task showed a great learning curve, i.e., ATCos were almost 19% slower in sorting the cards in their first simulation run compared to their second simulation run (baseline and solution alternated).

3.10. System Usability (Post-Run)

The post-run questionnaire contained the ten statements of the System Usability Scale (SUS), as listed in Appendix C.6. The results are shown in Figure 22 (one ATCo did not answer one of his ten statements both in baseline (without ASR) and solution (with ASR) condition. Therefore, the scale mean "3" ((5-1)/2) was chosen as a replacement to not heavily influence the overall result). ABSR support increases the system usability due to SUS ratings ($\alpha = 0.16$). There were three statements rated in the expected direction with an $\alpha < 0.075$, i.e., ATCos like to use the system, they do not deem it complex, and they hardly need support to use it.

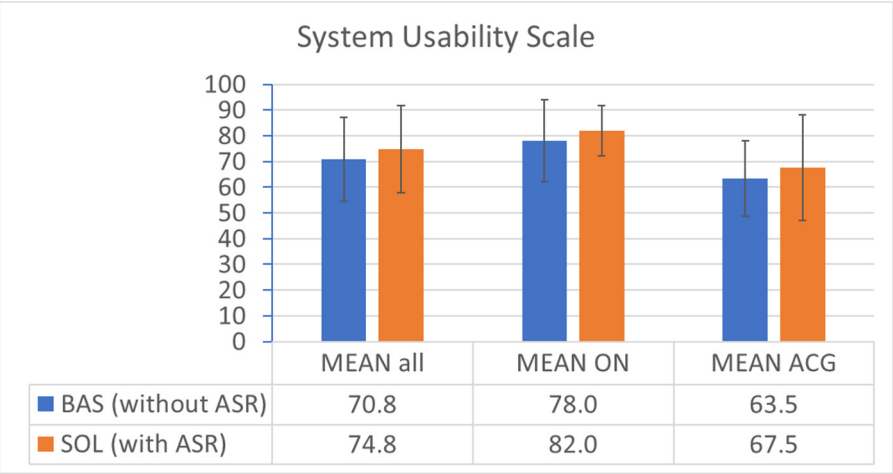


Figure 22. Subjective ATCo ratings on system usability.

Considering all ATCos, the SUS score was 4 percent absolute (5.7% relative) higher in the solution condition (SOL) with ABSR support compared to the baseline condition (BAS) without ABSR support. The difference of 4 percent remains when just analyzing the ON score or ACG score independently. However, the score itself is 14.5%, absolutely higher

for ON than for ACG. This is probably due to the fact that ON really liked the electronic flight strip display (also in the baseline version), whereas ACG ATCos needed to adapt themselves more to the strip system due to the difference in their daily-life system.

3.11. Debriefing Feedback (Post-Validation)

The debriefing was conducted as a semi-structured interview with some pre-defined questions and some options for further thoughts and inputs. The feedback of ATCos is semantically reported per category in the following subsections—the most important feedback relevant for future usage of ABSR is listed after arrow symbol bullets. However, also the remaining feedback helps to improve future simulation planning, i.e., to know which aspects that are not the core part of the study do influence the subject's experience and study results. For example, the prototypic flight strip system induced a row of effects on how the ABSR output is perceived. The last question outlines further research or usage of ABSR systems.

3.11.1. Study Preparation and Conduction

- Briefing slides via e-mail two weeks before the trials and briefing at DLR was very good;
- All ATCos felt well-trained for the purpose of the validation after one hour of training;
- Simulation pilots performed well;
- Air traffic scenarios were rated to be fine for the study purpose;
- On the one hand, the baseline condition (manual work) was similar to everyday work, so performance might be better, therefore (2 ATCos);

➢ On the other hand, ASR in solution condition was good because it supported using a flight strip system that ATCos were not used to.

3.11.2. ABSR Functionality (also Related to Electronic Flight Strip Display)

➢ ABSR concept and implementation were found to be good by many ATCos;
➢ Checking ABSR output in the flight strip display slows some ATCos because, in the baseline mode, ATCos tick while speaking;
➢ Some ATCos judged the speed of ABSR output while speaking as sufficient; two ATCos wanted to have faster output;
➢ Non-standard situations should be covered well, i.e., better, by ASR;
➢ Speech understanding (annotation process) was good for covering errors in speech recognition (transcription process);
➢ Highlighting of callsigns and status icons (in green) and the 10s-highlighting mechanism in electronic flight strips were fine for all ATCos;
➢ When ASR worked fine, a tendency to over-rely on automatism existed;
➢ In case of non-recognition, a double effort to manually recognize the error and correct it compared to pen input (2 ATCos);

- ABSR output in outside view (complete transcription and annotation in solution condition) was just checked for curiosity by all ATCos.

3.11.3. Feedback to Colleagues Not having participated

When I am home in Lithuania/Austria, I tell my colleagues that working with DLR's speech recognition was:

➢ Interesting (said by all ON ATCos);
➢ Worked pretty well (2 ATCos);
➢ Positively surprising (even when speaking fast);
➢ Very good even if not being an early adaptor of new technologies and being very safety critical.

3.11.4. Usefulness of ASR

If you would use it tomorrow in your tower controller working position (not multiple remote towers), would ASR help?

- ➢ Yes, that would be great (3);
- ➢ Nothing to be changed to be used tomorrow (1);
- ➢ Great support is possible if some/many aspects are improved (4).

3.11.5. Used Phraseology in Baseline and Solution Runs

Did you think you have spoken differently in baseline and solution conditions?

- ➢ In baseline less carefully spoken because only simulation pilots needed to understand (3 ATCos);
- ➢ Spoken closer to phraseology in solution as being better supported (2 ATCos);
- ➢ Some stated that there was no difference in speaking;
- ➢ "ATCos automatically become more phraseology conform: That is one of the greatest advantages of such a technology."

3.11.6. Flight Strip System (More Related to 'Multiple Remote Tower" than the Core Study Purpose 'ABSR Support')

- Runway bay handling needs to be improved (sorting, highlighting, timing, etc.);
- Drag-and-drop functionality over the borders of flight strip bays for individual planning purposes was needed;
- Handling training flights (touch-and-go/low approach) that do not switch from an arrival flight strip to a departure flight strip were slightly difficult;
- Strip handling for aircraft crossing the control zone is difficult with status options;
- Visual flagging of strips (left/right) would be beneficial;
- Hide some non-frequent status icons;
- "Takeoff" status should include "lineup"-status (if not given explicitly);
- A combination of the selection of taxi status and taxiway would be easier;
- Suggestions for colors, e.g., ground vehicles, consistency with other systems;
- One ATCo loved the flight strip system; the majority of ATCos were ok with it;
- Many ATCos liked the fade-away functionality of flight strips;
- The portion of gazes at the three areas 'flight strip display,' 'outside view,' and 'radar view': too much on flight strips and too few on outside view where one can hardly identify small objects.

3.11.7. Further Applications/Ideas/Things to Be Changed?

- ➢ Callsign highlighting in flight strip display from pilot utterance would help to identify the communication partner;
- ➢ Speech log for pilot utterances (especially in emergency situations) anywhere on the controller screen;
- Connect ABSR output with:
  - a. Radar information for automatic setting of landed/departed status;
  - b. Lighting system to turn off stop bar lights in case of lineup clearance;
  - c. Follow the greens for correct lighting;
  - d. Airport phone conversation to automatically extract and include stand numbers given by the airport;
  - e. Safety net functionality for dedicated aspects in case of good error rates, e.g., readback error detection;
  - f. Transcription for incident analysis and searching for callsigns; other analysis on transcribed data;
  - g. Great technology for on-the-job training.

4. Discussion on Major Study Results

The results on mental workload, situation awareness, satisfaction, acceptance, trust, and usability are ambivalent. The subjective post-run ratings on NASA-TLX, Bedford workload scale, and AIM-s, when interpreted as a whole, indicate a worse performance in solution runs with ABSR support compared to baseline runs without ABSR support.

However, the subjective post-validation rating on ABSR support for workload, the self-assessed workload ratings during the simulation runs by ISA, and the performance measurement of the objective secondary task indicate that ABSR support positively influences ATCo workload.

There might also be an influence through the usage of standardized and tailor-made questionnaires. The general low to medium workload level, as rated with roughly two on average on the five-point instantaneous self-assessment of workload scale, causes that it is hard to unambiguously measure a workload effect. Hence, the necessity for controller support functionalities might also be low in such a multiple remote tower environment.

The complexity of the task came with supervising three airports remotely at the same time with a working position the ATCos had not seen before. This could be the reason why especially the callsign highlighting was well-acknowledged by ATCos in order to reduce search times at the different displays. A workload reduction, especially in low workload conditions, is not always beneficial. Hence, it is also a success if the mental workload of ATCos is balanced at a medium level without peaks and boredom.

Similarly, the post-run rating on situation awareness (SASHA) indicates a negative influence, whereas the two rated post-validation statements on situation awareness at an acceptable level with ABSR support have answer values in the most positive scale third. Very similar effects were also seen for satisfaction, acceptance, and trust when comparing post-run ratings with overall post-validation answers.

The usability ratings (post-run and post-validation) seem to all indicate favor for ABSR support. The score of the system usability scale was four points better for the solution (with ABSR support) compared to the baseline (without ABSR support). A total of 80% of ATCos (with 8/10 or more points on the questionnaire scale) stated that they would accept such an ABSR system in their usual working position and that they could apply operating methods in a timely manner. Though, a row of adjustments were encouraged by ATCos, i.e., to make ABSR also reliable under non-nominal conditions where the pressure on ATCos is already high. The need for changes was rated very inhomogeneous by the different ATCos, i.e., some had already seen good support with the prototype's current technology readiness level, and others wanted to increase the number of covered situations and examples.

However, the comparison of a further objective measure with a subjective measurement again shows the ambivalence of some ATCo ratings: While ACG ATCos rated the perceived callsign recognition quality with 1.8 points higher than ON ATCos on a 10-point scale and the perceived command recognition quality with 1.6 points lower than ON ATCos such an effect cannot be seen in the online recognition rates where the callsign recognition rate and the command recognition rate in solution runs of ON ATCos was 2% and 10% (consistently both) better than of ACG ATCos, respectively.

Our study results based on text-to-concept analysis also revealed a potential safety issue for multiple remote towers: Even if ATCos were asked to utter the name of their current transmission station in each radio transmission, the station name, e.g., *vilnius tower*, was missing in every fifth utterance. This might confuse listening to cockpit crews being on or flying to one of the other two airports.

The subjective feedback through questionnaires etc., and the results from objective measurements at least are not consistent or even contradictory. This is a hint that ABSR's performance does not match with ATCos expectations. Objectively a word error rate of 10% with a command recognition rate of 80% might be sufficient to already have positive effects on workload. The ATCos are then, however, not trusting the system, which will be a showstopper. Objective improvements are not enough. ATCos also need to be convinced by their subjective feelings. Previous validation trials for Frankfurt airport to support apron

controllers by ABSR to reduce workload for pre-filling electronic flight strips [12] and for Vienna approach controllers [41] indicate that a command recognition rate greater than 90% is needed.

5. Conclusions and Outlook

5.1. Conclusions on ABSR Study in Multiple Remote Tower Environments

Human-in-the-loop trials were conducted with five Austrian and five Lithuanian air traffic controllers (ATCos) to validate whether an assistant-based speech recognition (ABSR) system can support air traffic controllers in a multiple remote tower environment. In baseline runs, controllers needed to manually maintain electronic flight strips without ABSR support, whereas in solution runs, they were supported by ABSR through callsign highlighting and automatically inputting recognized commands from ATCo utterances into electronic flight strips.

This study recorded a huge amount of data with results analyses that are shared with other researchers by this article. The chosen "within-subject design" [46] assessed the dependent variables mental workload, situation awareness, satisfaction, acceptance, trust, and usability with the independent variable "availability of ABSR support". Further qualitative feedback was gathered on ABSR accuracy, technical functionality, and operating methods. Although a very small number of training data of 3.6 and 0.9 h, respectively, was available for the adaption of the ABSR models to Lithuanian and Austrian tower phraseology, some results show statistical significance and are in line with findings of earlier ABSR projects from an approach environment [8]. The text-to-concept accuracy of the speech understanding module performed well, i.e., correcting wrong word recognition by context information. A callsign recognition rate of 94.2% and a command recognition rate of 82.9% were achieved, although each 10th word was wrongly recognized due to the observed word error rate of 9.8%. Given an independent distribution of word errors and an average callsign length of five words, a word error rate of 10% would result in a callsign recognition rate of below 60%, i.e., $(1–0.1)^5$. For an average command length of six words, including values, qualifiers, and conditions plus the five words for the callsign, the expected command recognition rate would be below 35%, i.e., $(1–0.1)^{11}$. These theoretical values were outperformed by our speech understanding module (command recognition) by using context information.

The study results on human factors comprised subjective ratings on mental workload, situation awareness, satisfaction, acceptance, trust, and usability via standardized and tailor-made questionnaires, the self-assessed workload during simulation runs, and an objective method to assess workload based on a secondary task.

The analysis results on the dependent variables were ambivalent. The reasons are the small number of study subjects, the prototype of a non-operational user interface, and the low workload resulting from low to medium traffic in the multiple remote tower environment of the chosen airports. A positive influence on workload was found with the self-assessed workload ratings during the simulation runs and the performance in the secondary task as a more objective measurement during simulation runs. Future validation trials involving ATCos should focus more on objective or live measurements than on retrospective ratings.

Our study results with ATCos reporting on benefits and drawbacks raise detailed awareness and give recommendations on which aspects of automatic speech recognition and understanding for a multiple remote tower environment are already solved and which aspects require deeper research to go beyond the now achieved technology readiness level four.

The speech-to-text performance is a prerequisite to enable good text-to-concept performance. An error analysis after the validation trials revealed processor overload as a factor in decreasing our speech-to-text performance. When applying our command extraction on offline speech-to-text analysis results having a word error rate of 4.4%, we achieve a command recognition rate of 91.8% and a callsign recognition rate of 98.2%. The data

analysis showed that ABSR support has a statistically significant positive effect on the usage of ICAO phraseology: The above-reported solution runs have higher command recognition rates than baseline runs because ATCos obtain better support if recognition rates are higher. If ATCos are sticking closer to ICAO phraseology just by the pure presence of an ABSR system, that will already be a safety feature. Some ATCos, i.e., the human operators that would use the operating system later on, highlighted that such an ABSR system would be a great support in their working position.

5.2. Outlook on Future Work

The amount of training data must be further increased, given representative samples. Furthermore, a large amount of data must be recorded from operations rooms (not from labs) because this is the operational environment. The European-wide agreed ontology for the annotation of ATC utterances was successfully used and enhanced in this study and should be further exploited or standardized. The continuous mutual enhancements of the ontology for en-route/oceanic, approach, tower, and apron traffic within the ASR projects HAAWAII (Highly Automated Air Traffic Controller Workstations with Artificial Intelligence Integration (HAAWAII), Homepage: https://www.haawaii.de (accessed on 4 April 2023)) (as the successor of MALORCA (Machine Learning of Speech Recognition Models for Controller Assistance (MALORCA), Homepage: https://www.malorca-project.de (accessed on 4 April 2023)), and STARFiSH (Safety and Artificial Intelligence Speech Recognition (STARFiSH), Homepage: https://www.dlr.de/fl/desktopdefault.aspx/tabid-1149/1737_read-74905/ (accessed on 4 April 2023)) tremendously build a base for interoperability of systems. Hence, following ASR activities can build on strong shoulders and reuse the achieved good results and methods of such ABSR projects.

For the specific case of electronic flight strips, eye tracking technology could be of further help to make sure that ATCos checked the ABSR output [47]. This technology could also be used to assess the time to recognize and correct an ABSR error (Times to correct ABSR errors in an ATM environment have been investigated in "Automatic Speech Recognition and Understanding for Radar Label Maintenance Support Increases Safety and Reduces Air Traffic Controllers' Workload" of Helmke et al. presented at the 15th USA/Europe Air Traffic Management Research and Development Seminar (ATM2023), Savannah, GA, USA, 5–9 June 2023). Furthermore, the support through callsign highlighting when recognized from pilot utterances should be investigated and potentially feed attention guidance systems at the controller working position. To summarize, the validation trials have shown the potential of using the output of an ABSR system in the multiple remote tower environment and revealed aspects to be considered when moving forward to higher technology readiness levels.

Author Contributions: Conceptualization, O.O.; Methodology, H.H., S.S. (Shruthi Shetty), M.K. and H.E.; Software, O.O., S.S. (Shruthi Shetty), M.K., H.E., S.S. (Saeed Sarfjoo) and P.M.; Validation, O.O., H.H. and S.S.-M.; Formal analysis, O.O. and M.M.; Investigation, O.O.; Resources, Š.M., T.P., H.U. and A.Č.; Data curation, O.O.; Writing—original draft, O.O.; Writing—review and editing, H.H., S.S. (Shruthi Shetty), M.K., H.E., S.S.-M., S.S. (Saeed Sarfjoo), P.M., Š.M., T.P., H.U., M.M. and A.Č.; Supervision, O.O. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the SESAR Joint Undertaking under the European Union's Horizon 2020 research and innovation program under grant agreement No 874470.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: We like to thank the ten air traffic controllers of ACG and ON who participated in the validation study in spring 2022 under a strict hygienic protocol due to COVID-19 as well as roughly one dozen of air traffic controllers from ON, ACG, and the Polish Air Navigation Services Agency (PANSA) who contributed to the recording of training data (surveillance data and speech

utterances) in earlier multiple remote tower trials at DLR Braunschweig. Further, we thank the air traffic management simulation department and the simulation pilots at DLR's Institute of Flight Guidance as well as our colleague Lennard Nöhren for supporting software preparation of simulation environment and conduction of a row of multiple remote tower human-in-the-loop validation studies in the course of our automatic speech recognition activities.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of this study; in the collection, analysis, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

ABSR	Assistant Based Speech Recognition
ACG	Austro Control
AIM-s	Assessing the Impact on Mental Workload
ASR	Automatic Speech Recognition
ATC	Air Traffic Control
ATCo	Air Traffic Controller
ATIS	Automatic Terminal Information Service
ATM	Air Traffic Management
BAS	Baseline Runs
CARS	Controller Acceptance Rating Scale
CoCoLoToCoCo	Controller Command Logging Tool for Context Comparison
CPU	Central Processing Unit
CWP	Controller Working Position
Del	Deletions
DLR	German Aerospace Center
DTT	Digital Tower Technologies
EASA	European Union Aviation Safety Agency
EFS	Electronic Flight Strip System
EUROCAE	European Organization for Civil Aviation Equipment
HMI	Human Machine Interface
ICAO	International Civil Aviation Organization
Ins	Insertions
ISA	Instantaneous Self-Assessment
LevenDist	Levenshtein Distance
NASA-TLX	National Aeronautics and Space Administration Task Load Index
Off	Offline (analysis of audio files after the simulation runs)
ON	Oro Navigacija
Onl	Online (analysis as experienced by ATCos during simulation runs)
OW	Overall Weighted Workload
SASHA	Situation Awareness for SHAPE
SATI	SHAPE Automation Trust Index
SD	Standard Deviation
SECT	Sequence Effect Compensation Technique
SHAPE	Solutions for Human Automation Partnerships in European ATM
SOL	Solution Runs
Subs	Substitutions
SUS	System Usability Scale
TWR	Tower
WER	Word Error Rate

Appendix A. Speech-To-Text Accuracy

The following tables in this Appendix A show the speech recognition performance on the word level, i.e., the word error rates (WER). The first row must be read like this; 1,944 words were spoken. Ninety-seven errors occurred, i.e., 43 words were substituted by another word, 38 words were not recognized at all (deleted), and 16 words were

inserted, i.e., not said, but a word was recognized. This results in a word error rate of 5.1% (97/1944).

Table A1. Speech-To-Text performance for offline recognition on audio files (Off).

Sample	# Words	LevenDist	# Subs	# Del	# Ins	% WER
MEAN all	1944	97	43	38	16	5.1
MEAN ON	1966	94	38	36	20	5.0
MEAN ACG	1921	99	48	39	13	5.1
MEAN w/o outlier run	1971	90	40	34	16	4.5
MEAN BAS all	1902	104	46	43	15	5.7
MEAN BAS ON	1891	100	41	43	16	5.7
MEAN BAS ACG	1913	109	51	44	14	5.7
MEAN BAS w/o outlier run	1961	98	44	39	15	5.0
MEAN SOL all	1985	89	40	32	17	4.4
MEAN SOL ON	2041	88	36	30	23	4.3
MEAN SOL ACG	1929	90	44	34	11	4.6
MEAN SOL w/o outlier run	1980	81	36	28	17	4.1

Rows are shaded, when containing all ATCos, i.e., both from ACG and ON.

Table A2. Speech-To-Text accuracy for real-time online recognition from voice stream (Onl).

Sample	# Words	LevenDist	# Subs	# Del	# Ins	% WER
MEAN all	1936	245	46	175	24	13.6
MEAN ON	1954	199	38	140	21	11.9
MEAN ACG	1918	290	54	209	27	15.3
MEAN w/o outlier run	1967	212	41	152	19	11.0
MEAN BAS all	1891	300	54	219	27	17.4
MEAN BAS ON	1871	261	42	196	23	17.1
MEAN BAS ACG	1911	339	66	241	32	17.8
MEAN BAS w/o outlier run	1959	254	50	181	23	13.2
MEAN SOL all	1980	189	38	131	21	9.8
MEAN SOL ON	2037	136	34	83	19	6.8
MEAN SOL ACG	1924	242	42	178	22	12.8
MEAN SOL w/o outlier run	1976	171	32	123	15	8.9

Rows are shaded, when containing all ATCos, i.e., both from ACG and ON.

The following two tables show the frequency of certain words appearing in the gold transcriptions and the number of unique words needed to reach a certain portion of all words in the gold transcriptions, respectively.

Table A3. 1-grams of gold transcriptions.

Rank	Word	Count	Portion
1	one	2393	6.43%
2	zero	1479	3.97%
3	tower	1473	3.96%
4	three	1356	3.64%
5	runway	1154	3.10%
6	five	1085	2.91%
7	seven	925	2.48%
8	two	923	2.48%
9	four	898	2.41%
10	to	888	2.38%
11	cleared	808	2.17%
12	right	795	2.13%
13	vilnius	747	2.01%
14	eight	721	1.94%
15	nine	720	1.93%

Table A3. *Cont.*

Rank	Word	Count	Portion
16	via	601	1.61%
17	air	571	1.53%
18	degrees	556	1.49%
19	and	539	1.45%
20	knots	531	1.43%
21	bravo	465	1.25%
22	wind	456	1.22%
23	alfa	409	1.10%
24	taxi	408	1.10%
25	kaunas	390	1.05%
	others	15,947	42.8%
1-505	SUM	37,238	100%

Table A4. The number of different words needed to reach a certain portion of all uttered words.

Count	Portion
61	80%
101	90%
145	95%
283	99%
505	100%

Appendix B. Text-To-Concept Accuracy

The following tables lists the relative frequency of supported air traffic control command types from the gold annotations.

Table A5. Percentage of used command types in gold annotations occurring more often than 1% (7560 commands in total).

Command Type	Portion of All Commands
STATION	20.2%
INFORMATION WINDSPEED	7.5%
INFORMATION WINDDIRECTION	7.5%
TAXI TO	6.4%
GREETING	5.6%
TAXI VIA	4.8%
AFFIRM	4.0%
INFORMATION QNH	3.3%
CLEARED VIA	2.9%
STARTUP	2.9%
CLEARED TO	2.9%
CLEARED TAKEOFF	2.8%
FAREWELL	2.8%
CLEARED LANDING	2.8%
SQUAWK	2.8%
LINEUP	2.4%
REPORT	1.5%
PUSHBACK	1.4%
INFORMATION ACTIVE_RWY	1.4%
NO_CONCEPT	1.4%
REPORT_MISCELLANEOUS	1.4%
VACATE VIA	1.2%
CLEARED TOUCH_GO	1.1%
others	8.9%

The following six tables present the speech understanding performance per study subject group and per command type group for gold, offline, and online transcriptions, respectively.

Table A6. Text-to-concept quality for gold transcriptions (assumed to be 100% correct).

Gold Transcription	Command Recognition Rate	Command Error Rate	Callsign Recognition Rate	Callsign Error Rate	Unknown Classified Rate	Amount of Data
all ATCos ALL	95.9%	2.4%	99.8%	0.2%	13.3%	100.0%
ON ATCos ALL	97.1%	1.5%	99.7%	0.2%	12.5%	49.9%
ACG ATCos ALL	94.8%	3.2%	99.9%	0.1%	14.2%	50.1%
ATCos ALL w/o outlier run	95.8%	2.5%	99.8%	0.2%	13.2%	91.8%
all ATCos BAS	95.9%	2.4%	99.7%	0.3%	13.8%	49.0%
ON ATCos BAS	97.6%	1.3%	99.7%	0.3%	13.0%	24.1%
ACG ATCos BAS	94.1%	3.5%	99.8%	0.2%	14.7%	24.8%
all ATCos SOL	96.0%	2.3%	99.8%	0.1%	12.8%	51.0%
ON ATCos SOL	96.6%	1.8%	99.7%	0.2%	12.0%	25.8%
ACG ATCos SOL	95.4%	2.9%	100.0%	0.0%	13.7%	25.3%

Table A7. Text-to-concept quality for gold transcriptions (assumed to be 100% correct) per command type groups.

Command Type Group	# Command Types	Command Recognition Rate
Relevant	34	97.3%
EFS	21	97.4%
Status	18	96.7%
Outside	3	96.0%
Hypo-EFS	4	99.2%

Table A8. Text-to-concept quality for Off transcriptions (current best word error rates of automatic speech-to-text with callsign boosting on audio files).

Offline	Command Recognition Rate	Command Error Rate	Callsign Recognition Rate	Callsign Error Rate	Unknown Classified Rate	Amount of Data
all ATCos ALL	91.4%	4.5%	98.4%	0.9%	14.0%	100.0%
ON ATCos ALL	92.7%	3.9%	98.6%	0.6%	12.8%	49.9%
ACG ATCos ALL	90.1%	5.1%	98.2%	1.2%	15.2%	50.1%
ATCos ALL w/o outlier run	91.7%	4.4%	98.7%	0.9%	13.9%	91.8%
all ATCos BAS	91.0%	4.6%	98.6%	0.8%	14.5%	49.0%
ON ATCos BAS	92.8%	3.6%	99.0%	0.3%	13.2%	24.1%
ACG ATCos BAS	89.3%	5.5%	98.1%	1.2%	15.8%	24.8%
all ATCos SOL	91.8%	4.5%	98.2%	1.1%	13.6%	51.0%
ON ATCos SOL	92.7%	4.1%	98.1%	0.9%	12.6%	25.8%
ACG ATCos SOL	90.9%	4.8%	98.3%	1.2%	14.6%	25.3%

Table A9. Text-to-concept quality for Off transcriptions (current best word error rates of automatic speech-to-text with callsign boosting on audio files) per command type groups.

Command Type Group	# Command Types	Command Recognition Rate
Relevant	31	92.4%
EFS	21	93.4%
Status	18	92.7%
Outside	3	90.5%
Hypo-EFS	4	96.3%

Table A10. Text-to-concept quality for Onl transcriptions (automatic speech-to-text with callsign boosting from continuous stream).

Online	Command Recognition Rate	Command Error Rate	Callsign Recognition Rate	Callsign Error Rate	Unknown Classified Rate	Amount of Data
all ATCos ALL	79.4%	7.0%	91.7%	3.1%	15.4%	100.0%
ON ATCos ALL	84.2%	5.5%	92.1%	2.4%	13.8%	49.9%
ACG ATCos ALL	74.6%	8.6%	91.3%	3.9%	17.0%	50.1%
ATCos ALL w/o outlier run	81.2%	6.6%	94.0%	2.5%	14.9%	91.8%
all ATCos BAS	75.7%	7.5%	89.1%	3.8%	16.2%	49.0%
ON ATCos BAS	80.1%	5.6%	88.9%	2.8%	14.6%	24.1%
ACG ATCos BAS	71.4%	9.3%	89.3%	4.8%	17.9%	24.8%
all ATCos SOL	82.9%	6.6%	94.2%	2.4%	14.5%	51.0%
ON ATCos SOL	88.0%	5.4%	95.2%	2.0%	13.2%	25.8%
ACG ATCos SOL	77.7%	7.9%	93.2%	2.9%	16.1%	25.3%

Table A11. Text-to-concept quality for Onl transcriptions (automatic speech-to-text with callsign boosting from continuous stream) per command type groups.

Command Type Group	# Command Types	Command Recognition Rate
Relevant	31	80.7%
EFS	21	79.2%
Status	18	80.0%
Outside	3	81.0%
Hypo-EFS	4	87.2%

Appendix C. Questions and Statements of Questionnaires

The following full-text questions and statements were contained within the listed post-run questionnaires:

Appendix C.1. Statement and Answer Scale from CARS

The color coding shows worse answers in red and good answers in green.

"Please read the descriptors and score your overall level of user acceptance experienced during the run. Please check the appropriate number."

▪ Improvement mandatory. Safe operation could not be maintained.
▪ Major Deficiencies. Safety not compromised, but system is barely controllable and only with extreme controller compensation.
▪ Major Deficiencies. Safety not compromised but system is marginally controllable. Considerable compensation is needed by the controller.
▪ Major Deficiencies. System is controllable. Some compensation is needed to maintain safe operations.
▪ Very Objectionable Deficiencies. Maintaining adequate performance requires extensive controller compensation.
▪ Moderately Objectionable Deficiencies. Considerable controller compensation to achieve adequate performance.
▪ Minor but Annoying Deficiencies. Desired performance requires moderate controller compensation.
▪ Mildly unpleasant Deficiencies. System is acceptable and minimal compensation is needed to meet desired performance.
▪ Negligible Deficiencies. System is acceptable and compensation is not a factor to achieve desired performance.
▪ Deficiencies are rare. System is acceptable and controller does not have to compensate to achieve desired performance.

Appendix C.2. Statements from SATI Questionnaire

1. In the previous working period, I felt that the system was useful. [USEFUL]
2. In the previous working period, I felt that the system was reliable. [RELIABLE]
3. In the previous working period, I felt that the system worked accurately. [ACCURACY]
4. In the previous working period, I felt that the system was understandable. [UNDERSTAND]
5. In the previous working period, I felt that the system worked robustly (in difficult situations, with invalid inputs, etc.). [ROBUST]
6. In the previous working period, I felt that I was confident when working with the system. [CONFIDENT]

Appendix C.3. Statements from SASHA Questionnaire

1. In the previous run, I was ahead of the traffic. [AHEAD]
2. In the previous run, I started to focus on a single problem or a specific aircraft. [FOCUS]
3. In the previous run, there was a risk of forgetting something important (such as inputting the spoken command values into the labels). [FORGET]
4. In the previous run I was able to plan and organize my work as wanted. [PLAN]
5. In the previous run I was surprised by an event I did not expect (such as an aircraft call). [SURPRISE]
6. In the previous run I had to search for an item of information. [SEARCH]

Appendix C.4. Questions from NASA-TLX Questionnaire

1. How mentally demanding was the task? [Mental Demand, MD]
2. How physically demanding was the task? [Physical Demand, PD]
3. How hurried or rushed was the pace of the task? [Temporal Demand, TD]
4. How successful were you in accomplishing what you were asked to do? [Operational Performance, OP]
5. How hard did you have to work to accomplish your level of performance? [Effort, EF]
6. How insecure, discouraged, irritated, stressed, and annoyed were you? [Frustration, FR]

Furthermore, the 15 pairwise comparisons of workload contributing factors have been analyzed. When looking at the subscores for all six NASA-TLX dimensions, half of them (three) were rated equal or better in SOL compared to BAS (PD, EF, FR), and the other half

was rated vice versa (MD, TD, OP). In general, physical demand (PD, 3.3%) was rated as being a less important contributor to workload, and mental demand (MD, 23.3%) was the most important contributor to workload. The other four dimensions were rather equally important contributors to the overall workload (TD 22%, OP 18%, EF 16.7%, FR 16.7%). The horizontal axis in Figure A1 shows the weight; the area shows the contribution of this very dimension to the OW of BAS and SOL conditions, respectively.

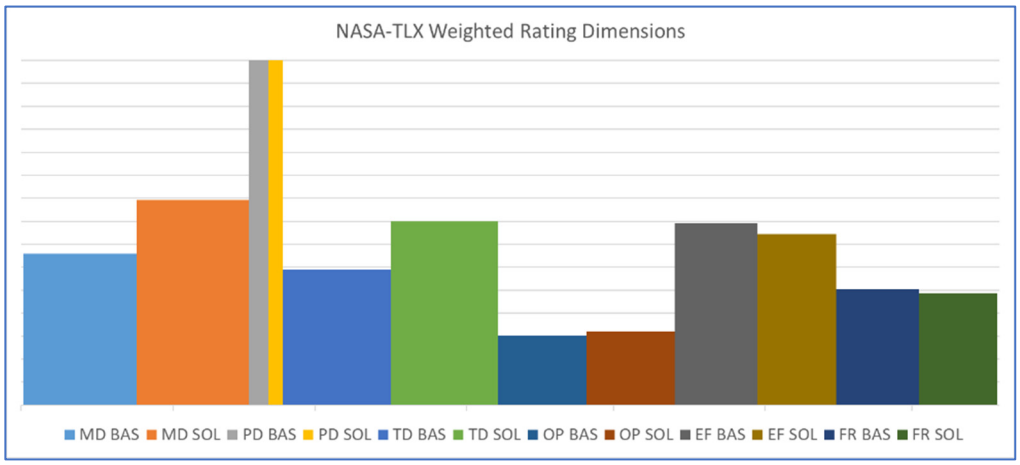


Figure A1. ATCo ratings on NASA-TLX (Weighted Workload Factors).

Appendix C.5. Questions from AIM-s Questionnaire

1. In the previous run, how much effort did it take to prioritize tasks? [PRIOT]
2. In the previous run, how much effort did it take to identify potential conflicts? [IDENT]
3. In the previous run, how much effort did it take to scan radar or any display? [SCRD]
4. In the previous run, how much effort did it take to evaluate conflict resolution options against the traffic situation and conditions? [EVAL]
5. In the previous run, how much effort did it take to anticipate the future traffic situation? [ANTIC]
6. In the previous run, how much effort did it take to recognize a mismatch of available data with the traffic picture? [RECOG]
7. In the previous run, how much effort did it take to issue timely commands? [TIMELY]
8. In the previous run, how much effort did it take to evaluate the consequences of a plan? [PLAN]
9. In the previous run, how much effort did it take to manage flight data information? [MANG]
10. In the previous run, how much effort did it take to share information with team members? [SHARE]
11. In the previous run, how much effort did it take to recall necessary information? [RECL]
12. In the previous run, how much effort did it take to anticipate team members' needs? [TMN]
13. In the previous run, how much effort did it take to prioritize requests? [PRIRQ]
14. In the previous run, how much effort did it take to scan flight progress data? [SCFP]
15. In the previous run, how much effort did it take to access relevant aircraft or flight information? [ACCD]
16. In the previous run, how much effort did it take to gather and interpret information? [GETI]

Appendix C.6. Statements from SUS Questionnaire

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.

4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Appendix D. Validation Setup Details

The left and right sides of the outside view areas presented current meteorological data as relevant for aircraft takeoff and landing (see Figure A2), i.e., wind speed in knots (here 10) and wind direction with an additional red arrow (here 070°) according to the runway orientation (grey rectangle), the active runway name (here 05), the airport International Civil Aviation Organization (ICAO) code (EYKA), the QNH (here 1001), the visibility conditions (here 9999, i.e., no visibility restrictions), and cloud information (in green circles).

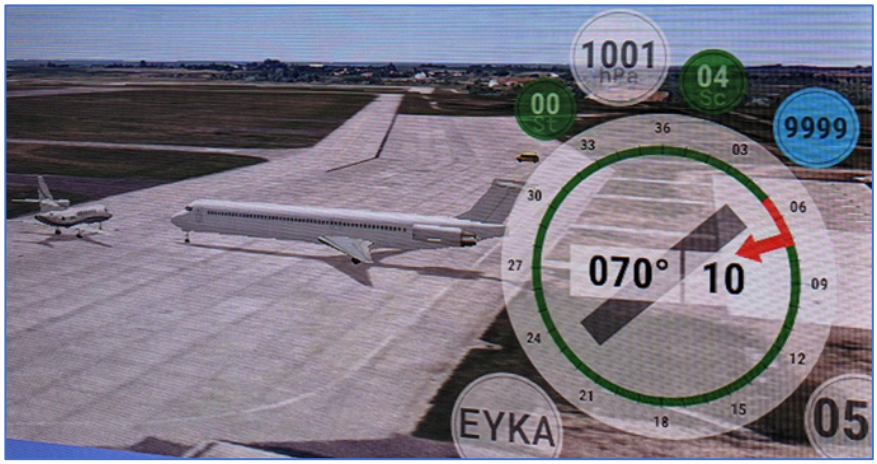


Figure A2. Remote tower outside view with a small aircraft passing a parking aircraft on the apron and meteorological information in and around the compass rose on the right.

An adjacent laboratory room accommodated three simulation pilot workstations. Each workstation consisted of a monitor to visualize the simulation pilot interface (see Figure A3) for one of the three simulated airports, a keyboard, and a mouse.

Figure A3. Simulation pilot interface for a simulated airport with time, pseudo flight strips for arrival and departure traffic, and radar views for airport surface and surrounding.

References

1. Lin, Y. Spoken Instruction Understanding in Air Traffic Control: Challenge, Technique, and Application. *Aerospace* **2021**, *8*, 65. [CrossRef]
2. Schäfer, D. Context-Sensitive Speech Recognition in the Air Traffic Control Simulation. Ph.D. Thesis, The University of Armed Forces, Munich, Germany, 2001.
3. Updegrove, J.A.; Jafer, S. Optimization of Air Traffic Control Training at the Federal Aviation Administration Academy. *Aerospace* **2017**, *4*, 50. [CrossRef]
4. Cordero, J.M.; Rodriguez, N.; de Pablo, J.M.; Dorado, M. Automated speech recognition in controller communications applied to workload measurement. In Proceedings of the 3rd SESAR Innovation Days, Stockholm, Sweden, 26–28 November 2013.
5. Cordero, J.M.; Dorado, M.; de Pablo, J.M. Automated speech recognition in ATC environment. In Proceedings of the 2nd International Conference on Application and Theory of Automation in Command and Control Systems, London, UK, 29–31 May 2012; pp. 46–53.
6. Kleinert, M.; Helmke, H.; Shetty, S.; Ohneiser, O.; Ehr, H.; Prasad, A.; Motlicek, P.; Harfmann, J. Automated Interpretation of Air Traffic Control Communication: The Journey from Spoken Words to a Deeper Understanding of the Meaning. In Proceedings of the IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), Virtual, 3–7 October 2021. [CrossRef]
7. Helmke, H.; Rataj, J.; Mühlhausen, T.; Ohneiser, O.; Ehr, H.; Kleinert, M.; Oualil, Y.; Schulder, M. Assistant-Based Speech Recognition for ATM Applications. In Proceedings of the 11th USA/Europe Air Traffic Management Research and Development Seminar (ATM2015), Lisbon, Portugal, 23–26 June 2015.
8. Helmke, H.; Ohneiser, O.; Mühlhausen, T.; Wies, M. Reducing Controller Workload with Automatic Speech Recognition. In Proceedings of the 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 25–29 September 2016.
9. Helmke, H.; Slotty, M.; Poiger, M.; Herrer, D.F.; Ohneiser, O.; Vink, N.; Cerna, A.; Hartikainen, P.; Josefsson, B.; Langr, D.; et al. Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ.16-04. In Proceedings of the IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), London, UK, 23–27 September 2018. [CrossRef]
10. Helmke, H.; Ondřej, K.; Shetty, S.; Arilíusson, H.; Simiganoschi, T.S.; Kleinert, M.; Ohneiser, O.; Ehr, H.; Zuluaga-Gomez, J.-P.; Smrz, P. Readback Error Detection by Automatic Speech Recognition and Understanding—Results of HAAWAII project for Isavia's Enroute Airspace. In Proceedings of the 12th SESAR Innovation Days, Budapest, Hungary, 5–8 December 2022.
11. Chen, S.; Kopald, H.D.; Elessawy, A.; Levonian, Z.; Tarakan, R.M. Speech inputs to surface safety logic systems. In Proceedings of the IEEE/AIAA 34th Digital Avionics Systems Conference (DASC), Prague, Czech Republic, 13–17 September 2015. [CrossRef]
12. Kleinert, M.; Shetty, S.; Helmke, H.; Ohneiser, O.; Wiese, H.; Maier, M.; Schacht, S.; Nigmatulina, I.; Sarfjoo, S.S.; Motlicek, P. Apron Controller Support by Integration of Automatic Speech Recognition with an Advanced Surface Movement Guidance and Control System. In Proceedings of the 12th SESAR Innovation Days, Budapest, Hungary, 5–8 December 2022.
13. Ohneiser, O.; Helmke, H.; Kleinert, M.; Siol, G.; Ehr, H.; Hobein, S.; Predescu, A.-V.; Bauer, J. Tower Controller Command Prediction for Future Speech Recognition Applications. In Proceedings of the 9th SESAR Innovation Days, Athens, Greece, 2–5 December 2019.
14. Ohneiser, O.; Helmke, H.; Shetty, S.; Kleinert, M.; Ehr, H.; Murauskas, Š.; Pagirys, T. Prediction and extraction of tower controller commands for speech recognition applications. *J. Air Transp. Manag.* **2021**, *95*, 102089. [CrossRef]
15. Badrinath, S.; Balakrishnan, H. Automatic Speech Recognition for Air Traffic Control Communications. *Transp. Res. Rec.* **2021**, *2676*, 798–810. [CrossRef]
16. Pellegrini, T.; Farinas, J.; Delpech, E.; Lancelot, F. The Airbus Air Traffic Control Speech Recognition 2018 Challenge: Towards ATC Automatic Transcription and Call Sign Detection. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019. [CrossRef]
17. García, R.; Albarrán, J.; Fabio, A.; Celorrio, F.; Pinto de Oliveira, C.; Bárcena, C. Automatic Flight Callsign Identification on a Controller Working Position: Real-Time Simulation and Analysis of Operational Recordings. *Aerospace* **2023**, *10*, 433. [CrossRef]
18. Helmke, H.; Ohneiser, O.; Buxbaum, J.; Kern, C. Increasing ATM Efficiency with Assistant Based Speech Recognition. In Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 26–30 June 2017.
19. Kleinert, M.; Helmke, H.; Moos, S.; Hlousek, P.; Windisch, C.; Ohneiser, O.; Ehr, H.; Labreuil, A. Reducing Controller Workload by Automatic Speech Recognition Assisted Radar Label Maintenance. In Proceedings of the 9th SESAR Innovation Days, Athens, Greece, 2–5 December 2019.
20. Chen, S.; Kopald, H.D.; Chong, R.; Wei, Y.; Levonian, Z. Read back error detection using automatic speech recognition. In Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 26–30 June 2017.
21. Fürstenau, N.; Jakobi, J.; Papenfuss, A. Introduction: Basics, History, and Overview. In *Virtual and Remote Control Tower Research Topics in Aerospace*; Fürstenau, N., Ed.; Springer: Cham, Switzerland, 2022; pp. 3–22. [CrossRef]
22. Möhlenbrink, C.; Papenfuß, A. Eye-data metrics to characterize tower controllers' visual attention in a multiple remote tower exercise. In Proceedings of the 6th International Conference on Research in Air Transportation (ICRAT2014), Istanbul, Turkey, 26–30 May 2014.
23. Papenfuss, A.; Friedrich, M. Head Up Only—A design concept to enable multiple remote tower operations. In Proceedings of the IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 25–29 September 2016.

24. Fürstenau, N.; Papenfuss, A. Model Based Analysis of Subjective Mental Workload During Multiple Remote Tower Human-In-The-Loop Simulations. In *Virtual and Remote Control Tower. Research Topics in Aerospace*; Fürstenau, N., Ed.; Springer: Cham, Switzerland, 2022; pp. 293–342. [CrossRef]
25. Hamann, A.; Jakobi, J. Changing of the Guards: The Impact of Handover Procedures on Human Performance in Multiple Remote Tower Operations. In *Virtual and Remote Control Tower. Research Topics in Aerospace*; Fürstenau, N., Ed.; Springer: Cham, Switzerland, 2022; pp. 343–363. [CrossRef]
26. Friedrich, M.; Timmermann, F.; Jakobi, J. Active supervision in a Remote Tower Center: Rethinking of a new position in the ATC Domain. In Proceedings of the 19th International Conference on Engineering Psychology and Cognitive Ergonomics, EPCE 2022 as part of the 24th HCI International Conference, HCII 2022, Virtual, 26 June—1 July 2022; Springer: Cham, Switzerland, 2022; pp. 265–278. [CrossRef]
27. Li, W.-C.; Kearney, P.; Braithwaite, G. The Certification Processes of Multiple Remote Tower Operations for Single European Sky. In *Virtual and Remote Control Tower. Research Topics in Aerospace*; Fürstenau, N., Ed.; Springer: Cham, Switzerland, 2022; pp. 511–541. [CrossRef]
28. Schier, S.; Rambau, T.; Timmermann, F.; Metz, I.; Stelkens-Kobsch, T.H. Designing the Tower Control Research Environment of the Future. Deutscher Luft- und Raumfahrtkongress, DLRK2013. In Proceedings of the English: German Aerospace Congress, Stuttgart, Germany, 10–12 September 2013.
29. Schier, S.; Manske, P. *VisiTop II—Briefing-Unterlagen*; Section 4.2. DLR-internal report; DLR Institute of Flight Guidance: Braunschweig, Germany, 2015.
30. Ohneiser, O.; Sarfjoo, S.; Helmke, H.; Shetty, S.; Motlicek, P.; Kleinert, M.; Ehr, H.; Murauskas, Š. Robust Command Recognition for Lithuanian Air Traffic Control Tower Utterances. In Proceedings of the InterSpeech, Brno, Czech Republic, 30 August–3 September 2021. [CrossRef]
31. Shetty, S.; Helmke, H.; Kleinert, M.; Ohneiser, O. Early Callsign Highlighting using Automatic Speech Recognition to Reduce Air Traffic Controller Workload. In *Human Factors in Transportation, Proceedings of the International Conference on Applied Human Factors and Ergonomics (AHFE2022), New York, NY, USA, 24–28 July 2022*; Plant, K., Praetorius, G., Eds.; AHFE International: New York, NY, USA, 2022; Volume 60. [CrossRef]
32. Jordan, C.S.; Brennen, S.D. *Instantaneous Self-Assessment of Workload Technique (ISA)*; Defence Research Agency: Portsmouth, UK, 1992.
33. Bongo, M.F.; Seva, R.R. Evaluating the Performance-Shaping Factors of Air Traffic Controllers Using Fuzzy DEMATEL and Fuzzy BWM Approach. *Aerospace* **2023**, *10*, 252. [CrossRef]
34. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*; Hancock, P.A., Meshkati, N., Eds.; North Holland Press: Amsterdam, The Netherlands, 1988; p. 198. [CrossRef]
35. Hart, S.G. NASA-Task Load Index (NASA-TLX); 20 years later. In Proceedings of the Human Factors and Ergonomics Society, San Francisco, CA, USA, 16–20 October 2006; Volume 50, pp. 904–908. [CrossRef]
36. Roscoe, A.H. Assessing Pilot Workload in Flight. In Proceedings of the AGARD Conference Proceedings Flight Test Techniques, Lisbon, Portugal, 2–5 April 1984.
37. Dehn, D.M. Assessing the Impact of Automation on the Air Traffic Controller: The SHAPE Questionnaires. *Air Traffic Control Q.* **2008**, *16*, 127–146. [CrossRef]
38. Lee, K.K.; Kerns, K.; Bone, R.; Nickelson, M. The Development and Validation of the Controller Acceptance Rating Scale (CARS): Results of Empirical Research. In Proceedings of the 4th USA/Europe Air Traffic Management R&D Seminar, Santa Fe, NM, USA, 3–7 December 2001.
39. Brooke, J. SUS—A quick and dirty usability scale. In *Usability Evaluation in Industry*; Jordan, P.W., Thomas, B., McClelland, I.L., Weerdmeester, B.A., Eds.; Taylor and Francis: London, UK, 1996; pp. 189–194.
40. Bangor, A.; Kortum, P.T.; Miller, J.T. An empirical evaluation of the system usability scale. *Intl. J. Hum.-Comput. Interact.* **2008**, *24*, 574–594. [CrossRef]
41. Helmke, H.; Kleinert, M.; Ahrenhold, N.; Ehr, H.; Mühlhausen, T.; Ohneiser, O.; Motlicek, P.; Prasad, A.; Zuluaga-Gomez, J. Automatic Speech Recognition and Understanding for Radar Label Maintenance Support Increases Safety and Reduces Air Traffic Controllers' Workload. In Proceedings of the 15th USA/Europe Air Traffic Management Research and Development Seminar (ATM2023), Savannah, GA, USA, 5–9 June 2023.
42. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710.
43. Ohneiser, O.; Helmke, H.; Shetty, S.; Kleinert, M.; Ehr, H.; Murauskas, Š.; Pagirys, T.; Balogh, G.; Tønnesen, A.; Kis-Pál, G.; et al. Understanding Tower Controller Communication for Support in Air Traffic Control Displays. In Proceedings of the 12th SESAR Innovation Days, Budapest, Hungary, 5–8 December 2022.
44. ICAO. *ATM (Air Traffic Management): Procedures for Air Navigation Services*; DOC 4444 ATM/501; International Civil Aviation Organization (ICAO): Montréal, QC, Canada, 2007.
45. Helmke, H.; Shetty, S.; Kleinert, M.; Ohneiser, O.; Ehr, H.; Prasad, A.; Motlicek, P.; Cerna, A.; Windisch, C. Measuring Speech Recognition and Understanding Performance in Air Traffic Control Domain Beyond Word Error Rates. In Proceedings of the 11th SESAR Innovation Days, Virtual, 7–9 December 2021.

46. Charness, G.; Gneezy, U.; Kuhn, M.A. Experimental methods: Between-subject and within-subject design. *J. Econ. Behav. Organ.* **2012**, *81*, 1–8. [CrossRef]
47. Ohneiser, O.; Adamala, J.; Salomea, I.-T. Integrating Eye- and Mouse-Tracking with Assistant Based Speech Recognition for Interaction at Controller Working Positions. *Aerospace* **2021**, *8*, 245. [CrossRef]

Text-To-Speech Application for Training of Aviation Radio Telephony Communication Operators

OLIVER OHNEISER
German Aerospace Center (DLR), Institute of Flight Guidance, Lilienthalplatz 7, 38108 Braunschweig, Germany

UMAIR AHMED
Clausthal University of Technology, Institute for Informatics, Albrecht-von-Groddeck-Straße 7, 38678 Clausthal-Zellerfeld, Germany

***Abstract*— Air Traffic Control (ATC) and its dedicated radio telephony communication are critical components of safe and efficient air traffic. After the COVID-19 pandemic, the aviation industry faced a shortage of air traffic controllers (ATCos) and pilots, highlighting a significant problem: managing resources for training new ATCos and pilots.**

This paper explores using a text-to-speech application (TTS app) to simulate aviation radio telephony communication. The app utilizes open-source pre-trained TTS models fine-tuned using publicly available ATC communication-specific datasets. It synthesizes textual ATC utterances to simulate ATCo instructions and pilot responses, creating a realistic two-way communication scenario. It includes twenty ATCo and eight pilot voice models developed using two distinct fine-tuning approaches: (1) an end-to-end TTS method leveraging deep learning techniques and (2) a voice cloning method supporting multi-lingual speech generation.

The app was evaluated in an online study by 20 international study subjects, comprising 14 ATCos, 4 pilots, and 2 individuals from other aviation backgrounds. The performance of the voice models varied across different aspects of audio quality such as overall experience, clarity, pronunciation, intonation, naturalness, and speed due to more than 4100 subjective rating values. The voice cloning models were rated significantly better overall than the end-to-end models. The female voice cloning models were rated significantly better overall than the male voice cloning models – both fine-tuned with ATCo data. The majority of voice cloning models especially for ATCo utterances received average overall ratings between 4 and 4.5 out of the highest score of 5. More than 83% of ratings classified the audio articulation speed as optimal. While some issues on pauses, pronunciation, and volume consistency were noted, the general feedback, especially on realism, showed the feasibility of TTS for ATC communication training. The possibility to synthesize speech faster than real-time and initial explorations of Large Language Models for TTS show that developing operational downstream applications is on the horizon.

***Index Terms*— Air Traffic Control (ATC), Aviation Radio Telephony Communication Simulation, Phraseology, Text-To-Speech (TTS), Voice Cloning**

I. INTRODUCTION

The aviation sector is grappling with a shortage of air traffic controllers (ATCo) and pilots after COVID-19 pandemic. A significant proportion of European ATCos are approaching retirement age or are attracted to other regions such as Middle East, exacerbating the shortage [1]. The pilot shortage is expected to worsen with projections indicating a global shortage of 50,000 pilots by 2025 due to an increase in the size of the aircraft fleet worldwide and a large proportion of current pilots reaching retirement age [2].

The training of this amount of new aviation operators including ATCos and pilots requires a lot of resources. Thus, any potential reduction of human working hours for training support directly decreases costs. Effective communication between ATCos and pilots – being a basic element in initial training and all kind of ATC simulations – is a cornerstone of air traffic control (ATC) operations next to skills such as strategic planning, quick decision-making, or managing unexpected situations. Aviation operators are learning the standard International Civil Aviation Organization (ICAO) phraseology to quickly and efficiently convey information and instructions via radio telephony [3, Chapter 12]. This includes the mechanisms of pilots listening to verbal ATCo instructions, pilots reading back the content of the heard instructions, and ATCos hearing back the pilot's readback in order to check the instruction understanding and potentially correct readback errors for safety reasons.

An example conversation with an utterance pair of an ATCo instruction and a pilot readback following the ICAO phraseology could look like this:
ATCo: "lufthansa one eight two lineup runway two six"
Pilot: "lining up runway two six lufthansa one eight two"

Such a simple example should reflect how aviation operators communicate in real-life – and especially in their early operational life.

A. Problem

However, in later operations, aviation operators often deviate from the ICAO phraseology to different extents and speak with a diversity of strong accents and audio speeds, which is currently less reflected in training.

A second example of an ATCo utterance omitting verbalizing the prescribed units "degrees" and "knots" for

Manuscript received XXXXX 00, 0000; revised XXXXX 00, 0000; accepted XXXXX 00, 0000. *(Corresponding author: O. Ohneiser)*. O. Ohneiser was the primary author.

Oliver Ohneiser is with the German Aerospace Center (DLR), Institute of Flight Guidance, Braunschweig, Germany (e-mail: oliver.ohneiser@dlr.de). He and his master student Umair Ahmed (e-mail: umair.ahmed@tu-clausthal.de) are with Clausthal University of Technology, Institute for Informatics, Clausthal-Zellerfeld, Germany.

All human individuals volunteered to participate in the text-to-speech rating study after reading the informed consent.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

the wind direction value “180°” and wind speed value “6 kt” may look like this:
“air france three eight victor wind one eight zero six runway two five cleared for takeoff”
A third example of a pilot utterance may look like this:
“contacting two three eight bye”
This example contains a sloppy abbreviated readback of the frequency “123.800” not repeating (1) the first digit always being one, (2) the decimal, (3) the trailing zeros, and (4) the aircraft callsign in direct response of a prior ATCo-pilot conversation chain.

The second and third example have the potential to mix up numbers for unintended ATC command types or command values. In current, fully automated training procedures, the degree of tolerance of such phraseology deviations is very low, while it is much higher facing operational pressure.

A fourth example of a pilot response contains a readback error with the digit “3” of the above frequency being accidentally replaced by “2” that the ATCo should spot when hearing back:
“contacting one two two decimal eight ...”

Traditional training methods are resource-intensive and may not be scalable given current and projected demand. Furthermore, they may not be accessible to all potential trainees, particularly in regions with limited access to conventional training facilities. Therefore, there is a need for a more accessible, cost-effective, and scalable approach to training, including communication training, in the aviation sector. Potential improvements to current training procedures and simulation technologies have been identified, such as reducing dependence on instructors during simulation training, utilizing web-based training methods, and updating current simulator systems to include recording and playback features [4],[5].

B. Solution

The digitization and innovative automation of ATC communication simulation without the need for costly human simulation operators can help to lightweight train new generations of aviation operators in various aspects of standard and real-life ATC communication.

This is where the potential of artificial intelligence (AI) technologies comes into play. One such AI-driven solution is the use of text-to-speech (TTS) technology for aviation radio telephony communication training. A TTS application, tailored for this specific context, could provide a realistic and interactive training environment that is accessible and affordable. It could simulate a variety of scenarios and accents, enhancing the adaptability and preparedness of trainees. Furthermore, the use of open-source datasets and voice models could reduce the dependence on extensive human resources such as simulation pilots, making the training approach more flexible.

Our developed text-to-speech application (TTS app) uses voice models fine-tuned on open-source datasets, such as the Air Traffic Control Simulation Speech (ATCOSIM[1]) corpus from Graz University of Technology [6] and the English TTS corpus of pilot speech with various accents hosted at LINDAT/CLARIAH-CZ[2] [7],[8]. The TTS app employs open-source Coqui-TTS[3] scripts to fine-tune two existing models: (1) the end-to-end encoder-vocoder LJSPEECH-VITS[4] model [9] and (2) the LJSPEECH-XTTS[5] model, which incorporates voice cloning technology to synthesize text to speech in 16 languages.

The app simulates aviation radio telephony communication by synthesizing an ATCo and a pilot speech audio file based on textual ATC utterances. More specifically, textual ATCo utterances are analysed by regular expressions to generate appropriate pilot responses to ATC commands, after which audio files are synthesized considering ATCo and pilot voice models for the two ATCo and pilot text inputs.

The TTS app introduces a suite of 28 unique voice models, each representing a distinct character with variations in origin, native language, linguistic background, accent, and specialization in the ATC sector, some embodying ATCos and others pilots. Amongst other mechanisms, it introduces random readback errors for the pilot text input and connects airline origin with speaker accent to enhance realism. Hosted locally via a Python Flask app, it is also web-accessible, allowing users to interact and provide feedback on each voice model’s quality, thereby contributing to the refinement of the models for a more realistic simulation. Looking even more into the future, the TTS technology can as well be used for operational downstream applications to replace or support the articulation of utterances in aviation operators’ communication.

C. Paper Structure

Section II presents a comprehensive review of the literature, setting the foundation for the methodology explained in Section III. The implementation details of the TTS app are presented in Section IV, while Section V details evaluating of the app and results. A critical discussion in Section VI provides insight and implications of the findings. Finally, Section VII, concludes and makes suggestions for future work.

II. LITERATURE REVIEW

This section presents the state-of-the-art on TTS activities relevant for the aviation environment from early studies to commercial products, dedicated speech corpora, and its combination with speech-to-text.

[1] https://www.spsc.tugraz.at/databases-and-tools/atcosim-air-traffic-control-simulation-speech-corpus.html
[2] http://hdl.handle.net/11234/1-1587 [1588/1462/1461]
[3] More background on Coqui-TTS will be given in Section III
[4] https://docs.coqui.ai/en/latest/models/vits.html
[5] https://docs.coqui.ai/en/stable/models/xtts.html

A. Early Text-To-Speech Studies

The use of TTS technology in general telephony has already been discussed decades ago [10]. Initial ideas of TTS in aviation highlighted the potential to alleviate pilot workload by reducing reliance on visual information. Several factors such as safety standards and validation processes to ensure accuracy and intelligibility of TTS generated speech might be reasons for slow introduction of this innovation into different aviation domains. The evolution of TTS technology, along with advancements in AI and natural language processing, has now reached a level of sophistication where its application in critical environments such as aviation becomes feasible.

Early experiments with pilots listening to ATC utterances found that intelligibility of synthetic voice is at least as good as human radio voice in average [11]. Furthermore, response times to synthetic voice messages were found to be comparable to human voice message responses [12],[13]. The level of persuasion and perception in another experiment were also rated very similar for human and synthetic voice, while the level of trust was marginally better for the human voice [14].

Conventional TTS systems often lack prosodic variation, e.g., resulting from emotion, so that the artificial nature of a voice is easily identified even while having good intelligibility [15]. Therefore, the use of synthetic voice communication has been suggested to make different accents of ATC communication partners easier to understand in order to reduce misunderstandings between ATCos and pilots [16].

B. Commercial Text-To-Speech Products

There exists a row of commercial TTS products in the aviation domain. Announcements for automatic terminal information service in the tower domain (D-ATIS) and en-route domain (D-VOLMET) are powered by TTS[6]. Voice warning systems in aircraft cockpits for imminent safety-critical situations exist for more than half a century [17],[18].

More than 100 artificial voices were created with a deep learning approach on 3000 hours of ATC audio data for Microsoft Flight Simulator's communication system to address the earlier existing limitation of the robotic nature of ATC voices[7]. The combination of voice recognition for verbal ATCo utterances and TTS technology for pilot responses[8] has been integrated with the NEWSIM platform of the German air navigation service provider DFS Deutsche Flugsicherung GmbH in Munich center to train multiple hundred ATCos[9] and was enrolled to further centers, training facilities, and ATC domains [19].

[6]http://www.speechtech.com/
[7]https://azure.microsoft.com/en-us/solutions/ai/dev-resources
[8]https://www.ufainc.com/atvoice
[9]https://www.airport-technology.com/news/newsmunich-center-uses-ufas-atvoice-system-to-train-air-traffic-controllers-5769131/

C. Combining Text-To-Speech and Speech-To-Text

Speech-To-Text technologies as part of automatic speech recognition and understanding systems have been successfully explored to support ATCos [20]. A European-wide agreed ontology has been defined to annotate the semantic meaning of ATCo and pilot utterances [21]. This ontology has as well been used to compare succeeding ATCo and pilot utterances to automatically detect readback errors [22]. However, high detection rates with low false alarm rates remain a challenge for automated systems [23]. Thus, it is important that human aviation operators are well-trained in detecting readback errors by themselves.

Recently, the design of an ATC-simulation chat bot providing students with additional training opportunities on their own has been proposed [24]. Furthermore, the feasibility of automating ATC within simulation environments for both training and experimentation has been assessed [25]. An early TTS functionality for ATCo commands and pilot readbacks has been developed and integrated into aircraft radar labels of a prototypical air situation display [26]. Furthermore, a simple TTS-based ATC communication training environment has been proposed and prototypically implemented [27]. This environment utilizes Google's Cloud Speech-To-Text voice input for three defined air traffic scenarios to manipulate simulated airborne aircraft movements. However, the speech synthesis of this prototype was claimed to not deliver realistic audio output [27].

ATC human-in-the-loop simulation trials usually require human simulation pilots to react on the ATCo utterances [28]. The number of simulation pilots often exceeds the number of ATCos in those trials, which comes with significant costs. The idea of replacing simulation pilots through the help of TTS systems arose already many years ago and has been implemented on a low technology readiness level [29]. A virtual simulation pilot engine using up-to-date advanced AI tools was suggested very recently [30]. This engine is capable of transcribing and understanding spoken instructions from ATCo trainees. Further, it can generate spoken pilot prompts following ICAO phraseology [31].

The same mechanism of automatically generating pilot repetitions in an autonomous pilot agent with speech recognition and understanding of ATCo utterances and speech synthesis for artificial pilot responses have been implemented in a deep learning-based framework [32]. Despite these innovative approaches, the study acknowledged limitations, notably the absence of qualitative analysis on the TTS-produced speech and the lack of exploration in fine-tuning the TTS module with ATC-specific audio data [33].

An Automatic Training Tool for ATC communication training has been designed leveraging cloud-based speech recognition and TTS technologies [34]. The integrated noise module could simulate, e.g., background noise, beeps, and signal volume changes [35]. An autonomous

pseudo pilot software that could communicate with human ATCos was as well discussed for unmanned aerial vehicles [36]. The above-mentioned activities show the potential use cases of a TTS app or options to integrate it, respectively.

D. Speech Corpora for Training and Fine-Tuning

To fine-tune TTS models, audio datasets along with corresponding text transcriptions are required. The public domain Linda Johnson Speech (LJSPEECH) based TTS models [37] fulfill this requirement, comprising more than 13,000 brief audio recordings. These recordings feature a single speaker who reads excerpts from seven different nonfiction books [38]. It is widely used in the TTS community due to its high quality and the variety of contained speech.

For a realistic audio experience, there is the further requirement for audio data specifically from the aviation domain, particularly radio telephony communication. This leads to two valuable speech corpora: (1) The ATCOSIM corpus from TU Graz [6] and (2) the English TTS speech corpus of air traffic (pilot) messages (with German, Czech, Serbian, and Taiwanese accents) [7].

Further corpora deal with specific aspects or lack transcriptions, e.g., Mandarin Chinese and English audio ATC utterances in the ATCSpeech corpus [39], a French-accented corpus [40], ATC utterances with differently added noise conditions in HIWIRE[10] [41], and around 140 hours of recorded English ATC communication across ground, tower, approach, and area control [42].

E. Summary of Contributions of Related Work

A few specialized speech corpora to enhance TTS and speech-to-text system performance have been set up, to facilitate the improvement of intelligibility and accuracy in ATC communication. AI-driven simulation engines have been deployed, offering nuanced, interactive training experiences for ATCos, focusing on phraseology, speaking, and comprehension. The adoption of deep learning TTS models significantly improved voice quality and naturalness. However, there is a notable absence of comparative performance analysis between systems developed with ATCC-specific corpora and general-purpose speech corpora.

III. METHODOLOGY

Our paper focuses on the development of a TTS app designed to simulate aviation radio telephony communication and the evaluation of its voice models. The primary objective is to create an interface that allows users to input ATC commands and synthesize these commands into speech using selected ATCo and pilot voice models. To accomplish this, the following steps are required:

1) Selection of a proper TTS framework, preferably open-source, to serve as the back-end of the app.
2) Identifying pre-trained open-source English language TTS models that can be fine-tuned with aviation-specific datasets.
3) Searching for publicly available aviation-specific speech corpora, preferably of various genders and accents, for fine-tuning pre-trained TTS models.
4) Fine-tuning of TTS models and monitoring the training and evaluation process.
5) Development of scripts to download, clean, and rearrange the data from speech corpora as per the requirements of the fine-tuning process.
6) Acquiring hardware resources, specifically a powerful graphical processing unit (GPU) required to run the fine-tuning process, as well as the synthesis of ATC commands to speech using the fine-tuned models.
7) Conceptualize and implement front-end and back-end, as well as an application programming interface (API), preferably a web-based interface so that the app is platform-independent and allows users to input ATC commands in text form and convert them into speech including clear instructions and feedback on the conversion process.
8) Coming up with a hosting mechanism for the app, preferably accessible from internet through a user-friendly domain name.
9) Setting up and maintaining the TTS app during the evaluation phase, creating of textual utterances, and analysing the results and feedback of study subjects.

The subsequent subsections will detail the methodology, covering each of the steps above in detail.

A. Selection of proper Text-To-Speech Framework

There are several free, open-source or commercial TTS tools, apps, and models available online. For this work, three aspects are important. Firstly, open-source tools are cost-effective as they are free to use and modify. Second, they promote reproducibility as the source code is publicly available, allowing others to replicate and verify the results. Third, open-source tools usually have a community of developers who contribute to their improvement and troubleshooting, which can be a valuable resource. Therefore, the open-source Coqui-TTS models were chosen for this work.

Coqui-TTS is an advanced library designed for high-quality TTS generation[11]. It offers several notable features and advantages over other TTS toolkits[12]:

- Pre-trained models in more than 1100 languages, supporting a wide linguistic range.

[10] https://catalog.elra.info/en-us/repository/browse/ELRA-S0293/

[11] https://docs.coqui.ai/en/latest/

[12] https://ilikeai.ai/coqui-ai/

- Tools for training and fine-tuning models in any language, enhancing flexibility and customization.
- Utilities for dataset analysis and curation, crucial for TTS applications.
- High-performance deep learning models for TTS tasks with various acoustic and vocoder models.
- Efficient model training, multi-speaker support, and a modular code base enabling innovation.
- Voice cloning with minimal audio, allowing for a highly customizable user experience.
- An advanced editor for detailed audio customization, including pitch and loudness adjustments.

Specifically, our work makes use of the Conditional Variational Autoencoder with Adversarial Learning (VITS) and XTTS models.

VITS is a TTS model that is end-to-end (encoder → vocoder together) and leverages cutting-edge deep learning techniques such as Generative Adversarial Networks (GANs) [43], Variational Autoencoders (VAEs) [44], and Normalizing Flows. It learns the text-to-audio alignment using Monotonic Alignment Search (MAS) without the need for external alignment annotations. The model's architecture is a fusion of the GlowTTS [45] encoder and HiFiGAN vocoder [46]. It is a feed-forward model with real-time capabilities on a GPU. The VITS model can also learn a new language or voice with approximately a one minute long audio clip, making it a potent tool for training TTS models in languages with limited resources.

XTTS is a generative TTS foundation model under the Coqui Public Model License (CPML). It is designed to clone voices in different languages using just a few seconds-long sample of the original voice. This technology is particularly useful for TTS apps with realistic and captivating voice model sounds.

B. Fine-Tuning with Speech Data

Fine-tuning a model in general helps to improve its performance for a specific dataset or task. Before the fine-tuning process, the fine-tuning data had to be prepared and cleaned. This involved arranging the downloaded audio files (wav) and their transcriptions in a specific structure as required by LJSPEECH. Subsequently, the fine-tuning scripts are modified to specify the location of the dataset, the metadata files, and other relevant parameters[13].

After data preparation and loading, the fine-tuning process needed to be monitored using tools such as Tensorboard. The monitoring of parameters such as training loss values and validation loss values helps to determine whether the model is learning and improving over time or if it is over-fitting or under-fitting. The fine-tuning process is set to end by itself after completing a certain defined number of epochs, at which point the fine-tuned models can be downloaded and used in the TTS app. However, the process also allows for manual termination, e.g., if the model performance deteriorates over time.

[13] https://docs.coqui.ai/en/latest/formatting_your_dataset.html#formatting-your-dataset

C. Characteristics of Speech Corpora and Fine-Tuned Voice Models

The ATCOSIM corpus comprises audio data and transcriptions from a diverse set of ATCo speakers of various ATC sectors, each with a unique profile in terms of native language, sex, mean age, and ATC experience [6]. It encompasses ten hours of audio data in English language comprising 10,078 utterances. All recordings are of 32 kilo Hertz (kHz) sample rate and were captured in real-time ATC simulations via a close-talk headset microphone. The corpus features speakers from Söllingen speaking German accent (four male with 1167, 1848, 808, and 1162 utterances, respectively), Zürich with Swiss German accent (three female with 1716, 1739, and 638 utterances, respectively), and Geneva with Swiss French accent (one female and two male with 238, 384, and 378 utterances, respectively). It includes a total of ten speakers [47], with a gender distribution of four women and six men, a mean age of approximately 31 years, and an average ATC experience of about eight years, with sector-specific experience around six years.

To facilitate user understanding, the fine-tuned ATCo TTS models based on ATCOSIM input data were renamed post hoc to reflect the speaker's gender, native tongue, sector, and sequence number. For instance, the "sm1" model became "male-german-söllingen-1", and "zf1" was renamed to "female-german-zürich-1", etc. for all models that emerged as a result of being fine-tuned using the respective dataset.

The same approach was followed for the English TTS speech corpus of air traffic (pilot) messages [7]. This corpus comprises 7377 recorded utterances of speakers, native in German (female with 1685 utterances), Czech (male with 1692 utterances), Serbian (male with 3000 utterances), and Taiwanese (male with 1000 utterances), communicating in English. The sentences recited by the speakers are derived from the field of ATC, specifically the messages utilized by aircraft pilots during standard flight operations. The text within the corpus is sourced from transcripts of actual recordings, a portion of which has been made publicly available in LINDAT/CLARIN.

The pilot TTS models fine-tuned using this dataset were renamed according to the accent of the speaker. Since the corpus contained recordings of pilot messages from speakers of four different accents, i.e., German, Czech, Serbian, and Taiwanese, the resulting TTS models were named "german-1", "czech-1", "serbian-1", and "taiwanese-1", respectively.

All voice models for ATCos and pilots – trained using XTTS – were named following this scheme. The same number of voice models was as well trained using VITS with the numbering scheme incrementing the XTTS model name numbers. This means, that e.g., "male-german-söllingen-5" to "male-german-söllingen-8" equal

"male-german-söllingen-1" to "male-german-söllingen-4" except that they are fine-tuned using VITS instead of XTTS.

D. Layout of Application to process ATC Text Commands and Synthesize Speech

The ATC communication is simulated through a Python web application based on the Flask framework. The app has a user-oriented graphical user interface (GUI) and contains two phases to run through.

In the first phase of the app, there is a fixed set of pre-defined textual ATCo utterances that the user must synthesize one by one. In this first phase, a fixed set of ATCo and pilot voice models is used to enable comparability of the following audio output ratings of the utterances for both ATCo and pilot. There are different styles how to note down textual ATCo utterances to enable user-friendliness, i.e., abbreviated ATC concepts or full words can be used. The two textual ATCo utterance examples "DLH123 climb and maintain FL240" and "lufthansa one two three climb and maintain flight level two four zero" should both be converted to the same textual pilot response "climbing and maintaining flight level two four zero lufthansa one two three".

The conversion of the textual ATCo utterances into textual pilot responses is sketched below:

1) **Extract Callsign and Command**: The code uses text processing functions and regular expressions to extract the callsign and the rest of the command from the received string.
2) **Convert Airline Designator to Full Form**: The code uses a JavaScript Object Notation (JSON) file that maps airline ICAO codes to their radio telephony designator, e.g., DLH to "lufthansa", BAW to "speed bird".
3) **Convert Numbers to Words and Aviation Terms to Full Forms**: The code uses a number-to-words mapping to convert numbers and letters in the text to their word forms (e.g., 0, 1, 2, 3, A, B to zero, one, two, three, alfa, bravo). It also converts aviation terms such as FL and RW to their full forms "Flight Level" and "Runway".
4) **Generate Pilot Response to the textual ATCo utterance**: Pre-defined verbs are converted into its gerund form (e.g., "climb and maintain" to "climbing and maintaining"). The callsign is placed at the end of the textual pilot response.

The duration of the speech synthesis process depends on the selected voice model, the length of the textual inputs, and the processing power of the machine (memory, graphical/central processing unit (CPU)). On a GPU it is usually a one-digit number of seconds for each of the two audio files – ATCo and pilot – with more precise values in Section VI.J.

E. User Interaction with Rating Metrics in the Text Synthesis Application

Users can play and listen to the resulting ATCo and pilot audio files and can also download them optionally. Users need to rate both generated audios before converting another textual utterance to speech. The rating of the synthesized speech is done via commonly used metrics in the TTS domain [48]. The judgement utilizes a Likert scale [49] visualized with stars. The stars' values range from 1 to 5, representing the lowest to the highest rating, respectively. User ratings, collected through a back-end system, are stored in a JSON file. For each audio output, there is a star scale for each of the following metrics:

- **Overall Experience:** This is an overall score that can be influenced by the other categories as well.
- **Clarity/Understandability:** This relates to how well the synthesized speech can be understood, which is a fundamental goal of TTS.
- **Pronunciation:** This is crucial for the intelligibility of single words.
- **Intonation/Melody:** This contributes to the melodic sound of the speech and can convey additional information like the mood of the speaker or the type of sentence (e.g., question vs. statement).
- **Naturalness/Realism:** To stick close to the usual ATC radio telephony sound is a key goal of TTS to be applied in the ATC environment.
- **Audio Speed:** "How was the audio speed?" (This is a drop-down menu with the options: *Optimal*, *Slow*, *Fast*). The articulation speed of the audio can affect its understandability.
- **Readback Error:** "Did you spot a readback error?" (This is a drop-down menu with the options: *Yes*, *No*). Introducing readback errors on purpose can further enhance realism and can be used to analyse the auditory attention of study subjects. *This question is only applicable for pilot audios.*
- **Comments:** Users can also give optional comments about ATCo and pilot audios.

In the second phase of the app, users need to select the ATCo and pilot voice model from two respective lists of ATCo and pilot voices. Then, the description of the selected voice model appears for information, i.e., the gender, language accent of the ATCo/pilot, and the ATC sector to which the ATCo belongs to if applicable. Subsequently, users have to enter the textual ATCo utterances into a text box. The rest of the process including synthesizing and rating is the same as in the first phase of the app.

IV. IMPLEMENTATION

This section describes the data preparation for fine-tuning, the fine-tuning process for the TTS models, the back-end, and the front-end development including the functionality of the TTS app.

A. Corpus Data Retrieval, Cleaning, and Preparation for Fine-Tuning

In the first step, audio files (wav) and transcription files (txt), which are required for the later fine-tuning process, needed to be downloaded efficiently from the online corpora. Custom bash commands were used to retrieve the portion of data as a bulk if this was not possible via the provided GUI. In the second step, a Python script creates the file "metadata.csv" including key-value pairs of audio file names and the corresponding transcription without non-standard characters in the required file format and depending on the corpus data characteristics. In the third step, another Python script cut, reformatted, and pasted selected portions of the "metadata.csv" into csv-files for the datasets with "validation_data" and "test_data", respectively, so that only the training data for fine-tuning remains in "metadata.csv".

B. Fine-Tuning Process

Coqui-TTS provides Python scripts (recipes[14]) to fine-tune its pre-trained models[15]. Some important training parameters to be adjusted via recipes or bash command are *learning rate*, *number of training epochs*, *test sentences for validation of fine-tuned model after each epoch*, *graphic card used for training*, *location of dataset used for training*, and the *name of the training run*.

The *learning rate* is a hyperparameter that determines the step size at each iteration while moving toward a minimum of a loss function. It decides the size of the model change given the estimated error. The term *number of training epochs* refers to the total count of iterations over the entire dataset that the learning algorithm will perform. The *location of the dataset used for training* is the file path where the training data is stored. *Test sentences for validation of the fine-tuned model after each epoch* are specific sentences that the model will attempt to generate after each training epoch. The fine-tuning process of the Coqui-TTS LJPEECH-VITS and LJPEECH-XTTS models, especially the average loader time and loss metrics, were monitored using Tensorboard, a visualization toolkit for machine learning.

C. Fine-Tuned ATCo and Pilot Models

The fine-tuning process resulted in a set of models for both the Coqui-XTTS and Coqui-VITS methodologies. The ATCo voice models are as follows:

- male-german-söllingen-1 to -4 (XTTS) and -5 to -8 (VITS) based on sm1 to sm4 datasets.
- female-german-zürich-1 to -3 (XTTS) and -4 to -6 (VITS) based on zf1 to zf3 datasets.
- female-french-geneva-1 (XTTS) and -2 (VITS) based on gf1 dataset.
- male-french-geneva-1 to -2 (XTTS) and -3 to -4 (VITS) based on gm1 and gm2 datasets.

The pilot voice models are as follows:

- pilot-german-1 (XTTS) and -2 (VITS)
- pilot-czech-1 (XTTS) and -2 (VITS)
- pilot-serbian-1 (XTTS) and -2 (VITS)
- pilot-taiwanese-1 (XTTS) and -2 (VITS)

D. Usage of Voice Models and Quality Hypotheses

The TTS app offered the usage of 20 ATCo models, 8 pilot models, and the *Untrained Voice* model. This is a pre-trained model which is not fine-tuned with any aviation-specific dataset. This was used on purpose to find out differences in quality metrics and perception of those models fine-tuned with and without aviation-specific datasets. All of the voice models could be selected in phase 2 of the app.

After initial exploration of the voice models and studying the literature, we formulated two hypotheses:

1) Voice models fine-tuned with XTTS receive better overall ratings in average than voice models fine-tuned with VITS.
2) Female voice models fine-tuned with XTTS on ATCo data receive better overall ratings in average than male voice models fine-tuned with XTTS on ATCo data.

The 16 fixed utterances in phase 1 of the app just used 13 ATCo models, 8 pilot models, and the *Untrained Voice* model, i.e., 7 of the ATCo VITS models were not included for the 16 fixed utterances due to their bad quality. For the result analysis, we only considered voice models that received at least three ratings because some voluntarily chosen models received less.

E. Back-End Development

The back-end of the TTS app was designed to receive textual ATCo utterances from the front-end, process them, convert them into speech, generate textual pilot responses, convert these responses into speech, logging the inference time of converting ATCo text and pilot text into speech, send both synthesized speeches to the front-end, receive user feedback on audio quality from front-end, and save the user ratings in a proper file format for further processing. The back-end was developed using the Python Flask framework, leveraging various libraries such as *torch*, *torchaudio*, and *transformers* for natural language processing and speech synthesis tasks.

The Flask application was designed with four routes. The first route was responsible for loading the main app page. The second route was dedicated to receiving

[14]VITS fine-tuning script: https://github.com/coqui-ai/TTS/blob/dev/recipes/ljspeech/vits_tts/train_vits.py and XTTS fine-tuning script: https://github.com/coqui-ai/TTS/blob/dev/recipes/ljspeech/xtts_v2/train_gpt_xtts.py

[15]Description of fine-tuning: https://docs.coqui.ai/en/latest/finetuning.html

textual ATCo utterances from the front-end, processing it, converting it into speech, and sending the resulting audio data as a byte stream to the front-end's audio player. The third route was responsible for receiving the textual ATCo utterance from the front-end, processing it, and generating a pilot response using various text processing functions and regular expressions. It would then synthesize the generated pilot response from text into speech and send the resulting speech as a byte stream to the front-end's audio player. The fourth route was handling the rating mechanism by receiving the user rating payload from the front-end and saving it to a persistent JSON file.

F. Automatic Pilot Model Selection Based on Airline Callsign

For enhancing the realism of the simulation a logic was implemented regarding the voice model selection. This is achieved by contextually selecting the pilot voice model based on the airline callsign mentioned in the textual ATCo utterance. A dictionary mapping of airline names and callsigns to their corresponding pilot voice model, configuration, and reference audio file path was created. For instance, if a French airline such as *Air France* is mentioned in a textual utterance, a pilot voice model with a French accent is selected. If no match is found between the textual utterance and the dictionary mapping, the system defaults to using the model specified in the form submitted by the user. This ensures that a model is always selected, providing a robust and user-friendly interface for the simulation.

G. Introducing Readback Errors on Purpose

For further enhancing the realism of the simulation the behavior of the aviation operator speakers includes common human errors. More precisely, readback errors were integrated on purpose in some of the pilot responses. This is done to mimic real-world scenarios where pilots might mishear or misinterpret ATC commands.

The fixed ATC utterance examples in phase 1 of the app contain two readback errors on purpose as it is an important task of ATCos to detect them [50]. The back-end replaces, e.g., "two" with "three" and "eight" with "seven" in the flight level, thereby a flight level of "two five zero" would be read back as "three five zero" by the pilot. Furthermore, at one of the 16 fixed utterances, the pilot response is modified by just uttering "say again," followed by the airline callsign, simulating a scenario where the pilot requests the ATCo to repeat the commands. This approach ensures a dynamic and realistic interaction between pilot and ATCos, closely resembling real-world aviation communication.

H. Front-End Development

The front-end of the app (see excerpt in Fig. 1) is designed to facilitate an intuitive user experience while providing robust functionality for simulating and evaluating aviation radio communications.

The app utilizes contemporary web technologies such as HyperText Markup Language version 5 (HTML5) to structure the site's content, while Cascading Style Sheets version 3 (CSS3) provides styling and responsive design, allowing the interface to adapt to various devices and screen sizes. JavaScript and Asynchronous JavaScript and XML (eXtensible Markup Language) (AJAX) are employed for dynamic content manipulation and asynchronous communication with the back-end, enabling a fluid interaction without the need for page reloads.

Before the main page loads, to avoid overloads and for security reasons, users were requested to enter a password in order to access the site[16]. The main page of the app has an introduction section which describes the purpose of the TTS app in detail. Users could then scroll down to the "Demographics Section" where they were asked to voluntarily enter some demographic information about themselves, i.e., nationality, native tongue, gender, and profession.

The main feature of the app is the "Speech Generation Section" as shown in Fig. 1. Textual ATCo utterances can be typed in (see white box with "lufthansa ..." in the upper middle part of Fig. 1) to be turned into realistic aviation operator speech. The synthetic speech can be rated using a five-star rating system (see middle part of Fig. 1) with further options for audio speed, comments, and readback error detection (see lower part of Fig. 1).

Finally, clicking the "Submit & Clear" button would send the ratings to the back-end of the app, clear the rating section and take the users back to the speech generation section where they could continue synthesizing further commands and repeat the rating process.

I. Implementing Fixed Commands and Models for App-Phase 1

In the first phase of the app, users were required to go through a set of 16 fixed ATC utterances, to be synthesized using a fixed set of ATCo and pilot voice models. This ensures comparability of user ratings who judge the different TTS parameters on audio files. As the number of ATCo voice models is greater than the number of pilot voice models, some ATCo voice models have as well been utilized to synthesize pilot responses due to the very similar audio characteristics. Five of the 16 fixed textual ATCo utterances are listed below[17]:

- "lufthansa eight one whiskey turn right heading three one zero" with ATCo voice model *female-german-zürich-1* and pilot voice model *pilot-german-1*
- "wizz air eight triple nine climb and maintain flight level two nine zero" with ATCo voice model

[16]The TTS app was hosted at: https://aviationtts.loca.lt/
[17]Some synthesized audio examples of ATCo and pilot voice utterances can be found at http://s.dlr.de/0xnK7

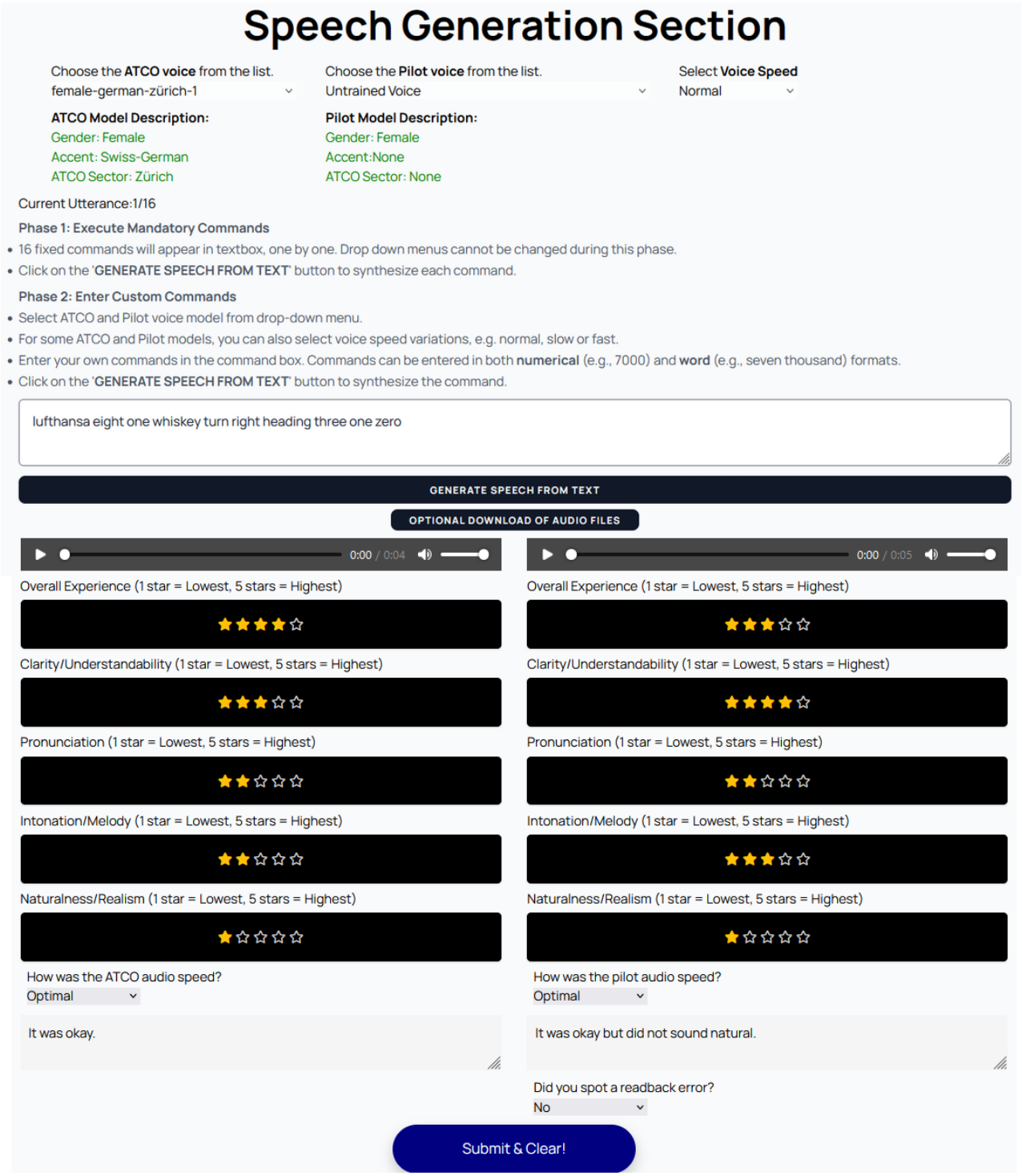


Fig. 1. Speech Generation and Voice Model Rating Section of the TTS Application

female-german-zürich-2 and pilot voice model *pilot-serbian-1*
- "oscar alfa november kilo foxtrot vacate runway to the left taxi via bravo one and alfa" with ATCo voice model *male-french-geneva-2* and pilot voice model *pilot-czech-2*
- "ryan air four four five startup and pushback is approved and you are cleared to destination echo delta delta fox flight planned route via erlos one delta departure climb seven thousand feet initially squawk one four zero one QNH one zero zero one" with ATCo voice model *male-german-söllingen-4* and pilot voice model *pilot-serbian-2*
- "AFR1435 descend flight level eight zero after passing bikmu" with ATCo voice model *male-german-söllingen-2* and pilot voice model *male-french-geneva-1* and a readback error on purpose.

All the 16 fixed utterances from the en-route, approach, tower, or ground domain are analogue to utterances of ATCos recorded in the real world or in a lab – including ATCos' deviations from prescribed ICAO phraseology like in practice. This supports the realism of the TTS app and evaluated utterances. Upon completion of the first phase of the app, users could freely enter utterances in the text-box and select voice models on their own in the second phase of the app.

V. EVALUATION RESULTS

This section presents the evaluation methodology and results of the study subject ratings on the TTS app. The interpretation and discussion of results will follow in Section VI.

A. Study Subjects

The app was evaluated by 20 international individuals of diverse ethnic and linguistic backgrounds. The nationalities of these study subjects were as follows: nine were German, five were Austrian, two were French, two were Pakistani, one was Dutch, and one was Egyptian. In terms of linguistic backgrounds, fourteen study subjects had German as their mother tongue, two had French, two had Sindhi, one had Dutch, and one had Arabic.

Study subjects were invited through emails, in which the link to the app, the password, and details of the relevant contact person were shared. This enabled them to use the app without requiring presence of any observer, e.g., they could use it at any time of the day. Some of the subjects used a laptop on-site and performed the study as a side-activity of another ATC prototype validation. It took the subjects roughly between 25 and 45 minutes to go through the intended usage of the TTS app.

B. Average Ratings and Standard Deviations for ATCo and Pilot Model Quality Metrics

Table I shows the average ratings for each of the five rating metrics – overall experience, clarity, pronunciation, intonation, and naturalness per voice model. Column "Method" indicates if the model was fine-tuned with *XTTS (X)* or *VITS (V)*. Column "Operator" indicates if the model was used for utterances of *ATCo (A)* or *Pilot (P)*. The number (#) of ratings differ as some models were used twice for the fixed utterances, needed to be rearranged slightly in the beginning of the study due to hardware memory issues, or were used in the second phase of the app as well. One rating means that a subject rated all five metrics for one voice model related to one utterance. The listed voice models "female-french-geneva-2" and "female-german-zürich-6" were the only VITS ATCo models with at least three ratings. The four listed pilot voice models "pilot-accent-2" were the only VITS models for pilots with at least twenty ratings or more. The mean score for XTTS models is 4.0 (with standard deviation of 1.05), and for VITS models, it is 2.5 (with standard deviation of 1.36). The mean score for female XTTS models fine-tuned on ATCo data is 4.3 (with standard deviation of 0.89), and for male XTTS models fine-tuned on ATCo data, it is 3.9 (with standard deviation of 1.05). The XTTS-ratings are left skewed, the VITS ratings are potentially symmetrical.

We performed a one-tailed two sample t-test with significance level 0.05 to reject the counter hypotheses of the two hypotheses formulated in Section IV.D. The calculated values for counter hypotheses 1 are : t=12.2845, degrees of freedom=670, p-value$<$1E-10. The calculated values for counter hypotheses 2 are : t=4.4233, degrees of freedom=475, p-value$<$0.000006031. Hence, both counter hypotheses can be rejected.

C. Data Volume for Fine Tuning vs. Voice Model Overall Rating

Fig. 2 visualizes the relationship between the volume of fine-tuning data and its impact on the average voice model overall ratings. The short names of the voice models include the following convention: First letter A/P for ATCo/Pilot fine-tuning data, second letter f/m for female/male voice, third letter c/f/g/s/t for Czech/French/German/Serbian/Taiwanese, fourth letter for ATCo sector if applicable g/s/z for Geneva/Söllingen/Zürich, and a digit to iterate the models.

D. Detection of Readback Errors

Table II lists the portion of readback errors that study subjects correctly detected. The two wrong pilot responses on purpose were the erroneous readback of "230 knots" instead of "220 knots" and "flight level 70" instead of "flight level 80".

TABLE I
Average Scores for each Metric of ATCo and Pilot Voice Models

Voice Model	Overall	Clarity	Pronunciation	Intonation	Naturalness	# Ratings	Method	Operator
female-french-geneva-1	4.2	4.1	4.0	4.2	4.2	76	X	A
female-french-geneva-2	1.3	1.5	1.7	1.5	1.3	6	V	A
female-german-zürich-1	4.3	4.2	3.9	4.1	4.2	42	X	A
female-german-zürich-2	4.1	4.0	3.9	4.0	4.3	42	X	A
female-german-zürich-3	4.2	4.5	4.4	4.3	4.3	23	X	A
female-german-zürich-6	2.7	3.0	3.0	2.7	2.3	3	V	A
male-french-geneva-1	3.4	3.2	3.0	3.5	3.6	18	X	A
male-french-geneva-2	3.8	3.5	3.6	3.6	3.7	19	X	A
male-german-söllingen-1	3.9	3.8	3.7	3.5	3.8	22	X	A
male-german-söllingen-2	4.0	3.9	4.1	4.0	4.4	41	X	A
male-german-söllingen-3	3.6	4.0	3.5	4.0	4.0	21	X	A
male-german-söllingen-4	3.7	3.7	3.3	3.6	3.6	36	X	A
female-french-geneva-1	4.4	4.3	4.1	4.2	4.4	28	X	P
female-german-zürich-1	4.2	4.3	3.9	4.1	4.1	25	X	P
male-french-geneva-1	4.1	3.9	4.1	4.5	4.4	19	X	P
male-french-geneva-2	4.0	4.1	4.0	4.0	4.0	45	X	P
male-german-söllingen-4	3.9	4.0	4.1	3.6	3.8	20	X	P
pilot-czech-1	3.4	4.0	3.6	3.3	3.3	21	X	P
pilot-czech-2	2.6	2.7	2.6	2.5	2.6	28	V	P
pilot-german-1	2.9	3.9	3.6	2.0	2.4	21	X	P
pilot-german-2	2.6	3.6	3.2	2.3	2.0	20	V	P
pilot-serbian-1	4.0	4.5	4.2	3.6	3.7	19	X	P
pilot-serbian-2	2.9	3.3	3.3	2.7	2.6	34	V	P
pilot-taiwanese-1	3.5	4.3	3.7	3.0	3.2	22	X	P
pilot-taiwanese-2	2.2	2.9	2.8	2.0	2.0	21	V	P
Untrained Voice	3.5	4.1	3.7	3.3	2.9	22	-	P
ATCos	4.0	3.9	3.8	3.9	4.0	349	X/V	A
Pilots	3.5	3.8	3.6	3.3	3.3	345	X/V	P
Female (XTTS, ATCo Data)	4.3	4.2	4.0	4.1	4.2	236	X	A/P
Male (XTTS, ATCo Data)	3.9	3.8	3.8	3.8	3.9	241	X	A/P
XTTS	4.0	4.0	3.9	3.8	4.0	560	X	A/P
VITS	2.5	3.0	2.9	2.4	2.3	112	V	A/P

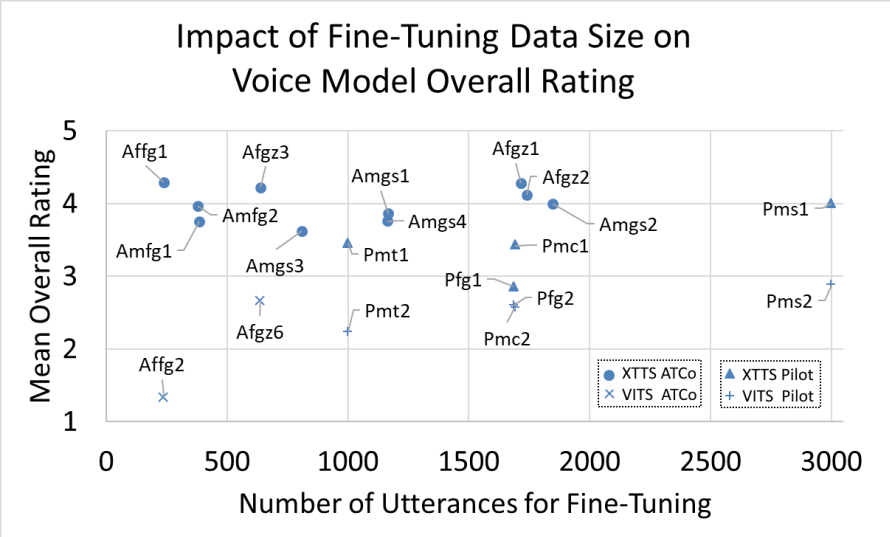


Fig. 2. Fine-Tuning Data vs. Overall Rating for ATCo and Pilot Voice Models

TABLE II

Number of Detected Readback Errors by Study Subjects

Readback Error	Occurrences	Detections
220 kt→230 kt	20	19
80 FL→70 FL	19	13

E. Correlation Coefficients for Rating Metrics

Table III illustrates the correlation coefficients between the five rating metrics of the ATCo voice models (values in lower left triangle) and pilot voice models (values in upper right triangle).

TABLE III

Correlation Coefficients of ATCo and Pilot Model Metric Ratings

Metric	Over.	Clar.	Pron.	Into.	Natu.
Overall	-	0.77	0.79	0.86	0.87
Clarity	0.78	-	0.81	0.65	0.63
Pronunciation	0.72	0.78	-	0.71	0.69
Intonation	0.71	0.67	0.66	-	0.88
Naturalness	0.77	0.70	0.68	0.75	-

F. Ratings on the Articulation Speed of Generated Audio

Table IV presents the distribution of subjective speed ratings (optimal/fast/slow) for ATCo and pilot audio articulation.

TABLE IV

Ratings on Audio Articulation Speed

Voice Model	Slow	Optimal	Fast
ATCo	3.1%	86.4%	10.5%
Pilot	18.8%	80.1%	1.1%

G. Comments from Study Subjects on ATCo and Pilot Voice Models

An extensive examination of study subject feedback with specific comments on ATCo and pilot voice models' performances including overall experience, clarity, pronunciation, intonation, naturalness, and speed revealed insights on strengths and areas for improvement. The voice models are generally well received, with some models receiving praise for their clarity and consistency. However, there are also suggestions for improvement of some models, particularly in the areas of pronunciation and modulation.

While some voice models were well appreciated, especially some VITS models expectedly showed room for improvement as those models were integrated to show the range of audio quality on purpose. Some synthesized audio examples showed difficulties in pronunciation of numerals as well as waypoint names such as *BIKMU* and had a weak modulation of the voice towards the end of the utterance due to the subjects' feedback. Furthermore, the volume of all audio files should be homogenized. The audio files should have a very slight pause after the aircraft callsign in ATCo utterances, but should not have unnecessary pauses in between values for ATC command types, which was sometimes the case. Some study subjects complained about too much accent in some voices.

H. Length of Textual Utterance (Tokens) vs. Inference Time (Seconds)

The scatter plot of Fig. 3 illustrates the relationship between the number of "tokens" in the textual utterance and the time it takes to convert this text into speech using the XTTS methodology, a process known as inference.

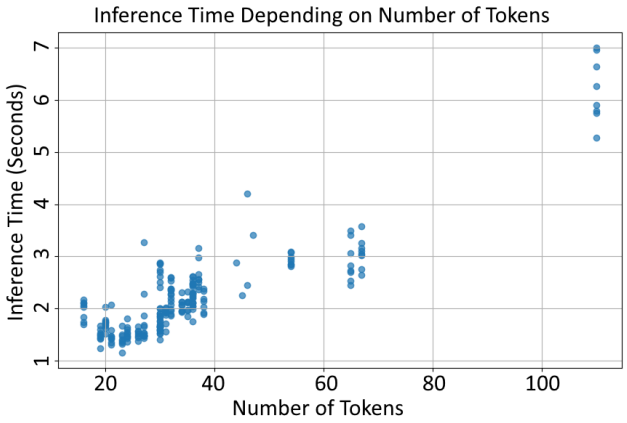


Fig. 3. Text Length in Tokens vs. Inference Time for Speech Synthesis on a Powerful GPU

In the context of TTS models, a token can be as small as a character, a common part of a word, or as large as a word. This process of breaking down text into these tokens is called tokenization. In the XTTS model, the tokenizer considers individually defined letter sequences as a token[18]. Tokens can be common English words (e.g., "the", "and", "is", "other"), parts of words (e.g., "ing", "ed", "ll"), or single letters if none of the bigger tokens can be found in a word. For example, the first utterance as listed in Section IV.I comprises of 27 tokens. The number of tokens on the 16 fixed utterances is 1.7 times greater in average than the number of syllables. The plot demonstrates how the time taken for synthesis on a powerful GPU varies with the length of the text (measured in tokens).

VI. DISCUSSION

The quantitative and qualitative feedback on various ATCo and pilot voice models in the TTS app received from the study subjects will be discussed below.

[18]The tokens in various languages are listed in this file: https://huggingface.co/coqui/XTTS-v2/resolve/main/vocab.json

A. Analysis of Speech Quality Metrics for ATCo and Pilot Models

The performance of a TTS model in simulating aviation radio telephony communication can be influenced by several factors, including the quality and amount of fine-tuning data, the specific techniques used for fine-tuning the model, and the individual preferences of the listeners. Therefore, monitoring and improvement of these models based on the feedback and scores are essential to ensure their reliability and effectiveness.

Out of a total score of 5 (see Table I), there are five voice models trained on ATCo data that achieved an overall score of 4.0 and above: "female-german-zürich-1" (4.3), "female-german-zürich-2" (4.1), "female-german-zürich-3" (4.2), "male-german-söllingen-2" (4.0), and "female-french-geneva-1" (4.2 on ATCo utterances; even 4.4 on pilot responses). In general, the "female-german-zürich-3" model can be seen as the best-rated due to its top scores in clarity (4.5), pronunciation (4.4), intonation (4.3), and naturalness (4.3). This model seems to be best in using pitch variation to convey meaning and sound human-like. Only the "male-german-söllingen-2" model has a slightly higher naturalness rating of 4.4. The "male-french-geneva-1" model had the worst ratings in all categories, but still all ratings were 3.0 or more. All ten XTTS ATCo voice models received 5 as the highest score at least once in each of the five metrics. When analysing the lowest scores for all ATCo models, "female-german-zürich-2" has never received a rating below 3 for the overall score in all 42 ratings. The VITS model "female-french-geneva-2", which was trained on the same data than the superior "female-french-geneva-1" model just received average ratings between 1.3 and 1.7 for the five metrics. This shows the dominance of XTTS voice models regarding their perceived audio quality.

When looking at the scores for voice models trained on pilot data, the "pilot-serbian-1" model has the top scores in all five metrics with an overall score of 4.0, clarity score of 4.5, pronunciation score of 4.2, intonation score of 3.5, and a naturalness score of 3.7. The "pilot-czech-1" and "pilot-taiwanese-1" models usually achieved scores between 3 and 4 on the five metrics. The "pilot-german-1" model just achieved an overall score of 2.9 with an intonation score of 2.1, showing a very ATC-unlike way of uttering responses. 19 of 20 metric values for the four XTTS pilot models were better or much better than for the VITS pilot models trained on the same data.

The voice models that were used for pilot responses, but were trained on ATCo data achieved comparable results as if they were used for ATCo utterances. However, in some metrics there were differences of up to one point between usage as ATCo voice or pilot voice, indicating that utterance contents also play an important role.

All 14 voice models used for pilot responses – independent of XTTS or VITS fine-tuning – received 5 as the highest score at least once in each of the five metrics. This shows that some subjects already liked the voice models even if they had worse audio quality. This also shows that the study just provides subjective results. When analysing the lowest scores for those models, "female-french-geneva-1" has never received a rating below 3 for the overall score in all its 28 ratings.

The *Untrained Voice* model has a high clarity score of 4.1, but the lowest intonation score of 3.3 and naturalness score of 2.9. This suggests that while the model is good at speaking clearly, it needs to be fine-tuned with aviation-specific data to accurately mimic the aviation radio telephony communication partners.

It is observed that the top-4 voice models are female voices. These scores indicate that the models were particularly effective in delivering clear and understandable speech. The female XTTS models fine-tuned with ATCo utterances were rated statistically significantly better overall than the male XTTS models fine-tuned with ATCo utterances ($p<0.0001$), despite both being fine-tuned with on average around 1000 utterances per voice model. There could be several reasons why female voices received higher ratings even if this observation may not apply to every individual model or listener:

- **Pitch and Tone:** Female voices typically have a higher pitch, which can make the speech sound clearer and easier to understand, especially in noisy environments such as aviation communication [51].
- **Speech Patterns:** Some studies suggest that female voices tend to articulate words more clearly and use a wider range of pitch and intonation, which can make the speech sound more expressive and natural [52].
- **Listener Perception:** There could also be a bias in the listener's perception. Some research has shown that listeners often perceive female voices to be warmer and more trustworthy [53]. Also speech-to-text systems seem to "prefer" female voices regarding the recognition performance [54].

B. Impact of Dataset and Fine-Tuning Method on Model Quality

The ATCo and pilot voice models can be compared regarding the use of XTTS methodology based on 560 ratings or VITS methodology based on 112 ratings for fine-tuning, respectively. This comparison is neglecting the *Untrained Voice* and voice models with less than three ratings. There is a notable difference in overall ratings between these two fine-tuning methods. The average overall rating for XTTS models is 4.0 and for VITS models it is just 2.5. The XTTS models received a statistically significantly better overall rating than the VITS models ($p<0.0001$). Also when comparing the four XTTS models fine-tuned on pilot data against the four VITS models fine-tuned on the same data, the average overall ratings have a huge difference with mean overall ratings of 3.4 and 2.6, respectively. Hence, for both ATCo and pilot models, the XTTS-trained models consistently outperform the VITS-trained models.

Upon examining the volume of data used for fine-tuning (see Fig. 2), it is observed that the overall ratings of pilot's XTTS and VITS models on the same data are closer together than of ATCo's XTTS and VITS models fine-tuned on the same data each. The four relevant pilot models have a larger dataset, with 1000 to 3000 utterances, compared to the two relevant ATCo models, which had less than 650 utterances per speaker. These findings underscore the importance of the choice of fine-tuning method in the development of TTS applications for specialized domains such as ATC communication. The XTTS model appears to be more efficient at generating speech that aligns with the expectations and requirements of ATCC professionals. Therefore, future efforts in TTS model training and fine-tuning for specialized communication domains should prioritize not only the choice of fine-tuning method but also the adequacy of datasets.

C. Impact of Data Characteristics on Model Quality

A common assumption is that a larger dataset leads to better model performance, as was observed with VITS-trained pilot models. However, the higher quantity of fine-tuning utterances does not necessarily lead to a better overall rating. The results on XTTS models show that a few multiple hundred representative utterances are already sufficient to fine-tune a well-performing voice model. Models fine-tuned with less data have, in some cases, outperformed those fine-tuned with more as shown in Fig. 2. This phenomenon is attributed to the quality of the dataset. For example the XTTS model "female-french-geneva-1" achieved one of the best overall scores despite being fine-tuned on only 238 utterances. The XTTS model "pilot-taiwanese-1", which was fine-tuned with a smaller set of utterances compared to the models "pilot-czech-1" and "pilot-german-1", received a higher overall score. This higher score is linked to the Taiwanese dataset's realistic speech patterns, which lacked the unnatural pauses between words of one utterance present in the Czech and German speech datasets. Such pauses can affect the perceived naturalness of the synthesized voice. The findings illustrate that high-quality fine-tuning data, characterized by a close approximation to natural speech patterns, is instrumental for the development of effective TTS models. Even if not systematically evaluated, fine-tuning datasets seem to not only require a wide range of speech samples but also closely mimic the rhythm and prosody inherent in ATC communication.

D. Analysis of Study Subjects' Comments on Models

The user experience with the synthesized voices of both ATCo and pilot models was extensively documented through their written textual comments. These comments provide crucial insight into the perceived quality and realism of the TTS app.

Feedback on ATCo models highlights a spectrum of issues ranging from pronunciation difficulties to the pace of speech delivery. A recurring theme in several comments is the need for clarity in numerical pronunciation and instruction delivery. Feedback suggests that deviations from expected intonation and pace can significantly impact the intelligibility of communications. Moreover, comments like "died at the end" and "voice fades at the end" point to technical issues with volume consistency, which can hinder communication in real-world aviation scenarios.

Similarly, the pilot models received feedback emphasizing the importance of naturalness in speech. The *Untrained Voice* model was praised for its lack of accent and clarity, suggesting that a neutral and clear voice is preferred for aviation communication. However, unnecessary pauses and robotic intonations were common criticisms for several models, including the "pilot-czech-1" and "pilot-german-1" models.

The "male-german-söllingen-2" model was criticized by some subjects for being challenging to understand, with specific references to the unclear pronunciation of *BIKMU* and the necessity for slower speech to accommodate pilots who are non-native English speakers. The study subjects also noted hesitations and unnatural intonations that could potentially confuse the pilots.

It is important to note that the synthesis quality of some terms is influenced by the fact that the input data is predominantly from en-route communications, with less data from approach and tower communications. This imbalance in the data can affect the performance of the models in different ATC scenarios. Furthermore, words that were not seen in the fine-tuning data may sound strange when synthesized. This is particularly relevant for artificial waypoint names such as *BIKMU*, which may not have been present in the fine-tuning data and therefore could be pronounced in an unexpected way.

Considering the issues with pauses, volume consistency, and pronunciation of some words, the model training could be improved. As the quantity and quality of input data for model training has a huge effect on the quality of the output, one should carefully select training data and acquire more real-life training data if possible.

The pause issue is usually connected with the speech rhythm in the training data, i.e., slowly speaking speakers in audio data from simulations might reduce the degree of realism. If there are more audio data from operational environments with corresponding transcriptions available, audio data from simulations could be omitted. To cover the volume consistency issue, a pre-processing step could harmonize the volume of input audio data.

The pronunciation issue is often connected with artificial waypoint names and other specific entity names that have not been part of the audio training data. There are three ways to improve the pronunciation: (1) providing phonetic spelling for specific entity names, (2) include audio training data that encompass the later application environment with those entity names, and (3) leverage a speech-to-text system on the TTS audio output to automatically compare the original TTS input text with

the speech-to-text output text and improve the identified differences. The latter approach with a human evaluator to analyze TTS input and its speech-to-text output can of course be used for all kind of unexpected deviations – not only for waypoints and entity names.

Feedback underscores the importance of a balance between clear pronunciation, appropriate pacing, and realistic intonation. Moreover, depending on the ATC use case of TTS, technical aspects such as consistent volume and absence of background noise can be critical to ensure effective training or operational scenarios. This feedback will be instrumental in guiding the next iterations of model fine-tuning and development.

E. Detection of Readback Errors in Fixed Utterances

The majority of study subjects was able to spot intentionally induced readback errors. The number of correct or missing readback error detections was an objective value (see Table II). However, subjects could replay each audio file as often as they wanted. In fact, out of the 20 subjects just one pilot subject did not spot the speed readback error in the fifth utterance. This indicates a high level of attentiveness at the beginning of the TTS app evaluation. However, six out of 19 subjects failed to spot the flight level readback error in the fifteenth utterance. On the one hand, in this utterance there was also a *DIRECT_TO* command, which might have caused lower attention to the flight level value readback in one of the last utterances. On the other hand, just two of those six failing subjects were ATCos who are usually drilled to detect such readback errors. Only one out of the four pilots detected the second readback error as pilots are usually just trained to listen and repeat an instruction, but not to hear back. It has to be noted that the subjects claimed 43 further readback errors on a variety of utterances next to the 32 detected ones for the intended errors. This highlights huge differences in the perception of what amount of assumed or actual deviation from the ICAO phraseology is acceptable for one or the other individual, because all textual ATCo utterances were taken from communication data of ATCos during operation or human-in-the-loop simulation.

For example, one subject complained about the callsign being at the end of the pilot readback, which is common practice today. Others were hesitating if they would accept the terminology “directing towards a waypoint” in operational life. These findings emphasize that realistic ATC communication simulation environments in training with natural utterances can help to foster a better mutual understanding of ATCos and pilots in operation. If simulations are conducted to analyse non-communication-specific aspects, clarity of the voice should have higher priority than naturalness.

F. Analysis of Correlation Coefficients for Models

The correlation coefficients in Table III show how different audio characteristics impact the overall metric of synthesized voices in both the ATCo and pilot models. All correlation coefficients were in the positive range between 0.63 and 0.88 indicating moderate of high correlation. The theoretically possible range was from -1.0 to 1.0. Strong positive correlations in the magnitude between 0.7 and 0.9 are observed across many main characteristics, indicating that improvements in clarity, pronunciation, intonation, and naturalness generally improve overall metric ratings.

For ATCo models, a particularly notable correlation between clarity and the two metrics overall rating as well as pronunciation can be seen (correlation coefficients of 0.78 each). This implies that clear articulation in communication is paramount for the effectiveness of a TTS system in ATC. Furthermore, naturalness also presents a strong link with the overall metric (correlation coefficient of 0.77), suggesting its importance in the perceived quality of TTS voices. The intonation has a strong link with naturalness (correlation coefficient of 0.75), but weaker links with clarity and pronunciation (correlation coefficients of 0.67 and 0.66, respectively).

In the case of pilot models, the three metrics overall rating, naturalness, and intonation have a very strong link with correlation coefficients of 0.88, 0.87, and 0.86, respectively. These values exceed the correlation coefficients for ATCos by at least 0.1. Similar to the ATCo model rating, there is a strong link between clarity and the two metrics overall rating as well as pronunciation (correlation coefficients of 0.77 and 0.81, respectively). Interestingly, the correlation coefficient between clarity and naturalness (0.63) as well as clarity and intonation (0.65) are the weakest among all. This could reflect real-world scenarios where cockpit engine noise interferes with the clarity of communication or pilots might mumble, yet the speech still needs to sound natural.

G. Analysis of Ratings on Articulation Speed of Generated Speech

Table IV, representing the speed ratings for the ATCo and pilot utterance articulation, conveys user preferences about the pace of synthesized speech. Most study subjects rated the speed of both ATCo and pilot audio output as optimal, with 86.4% for ATCo and 80.1% for pilot models, indicating satisfaction with pacing.

For the pilot models, only a small fraction of audio outputs were considered too fast (1.1%), however, a relevant portion were found to be too slow (18.8%). For example, a Taiwanese pilot utterance with just around two syllables per second was rated too slow by more than two thirds of subjects. The other way round, only a small fraction of audio outputs are considered too slow (3.1%) for the ATCo models, however, some were found to be too fast (10.5%). These findings go well along with the perceived speed of audio files in the fine-tuning data.

It is important to note that “Optimal” was the default selected option in the audio speed rating drop-down menu. This could potentially introduce a “default bias”, where a majority of subjects may not have actively

changed the selection, leading to an over-representation of "Optimal" ratings. To mitigate this bias and obtain a more accurate representation of user preferences, future surveys could consider not pre-selecting any option in the rating drop-down. This would require users to actively make a selection, providing a more reliable reflection of their true preferences. This consideration is crucial for the interpretation of the rating results and the subsequent tuning of the voice models.

H. Limitation of Study Subjects' Ratings

Even if the professional aviation background of study subjects was a benefit, the number of subjects was rather small with 20 individuals. Their subjective ratings will usually only lead to statistically significantly results if bigger groups of ratings are analysed together. However, the amount of individuals means that there were more than 4100 ratings on synthesized audio files including a few additional ratings on individually, freely entered textual ATCo utterances and pilot responses. The rating results might depend on the concrete combination of textual ATCo utterances and the selected voice models, i.e., if voice models would have been exercised on other textual utterances, the rating might have differed. Furthermore, there was a mix of ATCo and pilot models as well as of "good" and "less good" voice models on purpose to see the effect of quality differences. The number of ratings per voice model also varied – some received more than 70 ratings, typically there were around 20 ratings per voice model. Seven VITS models out of the 28 generated models were excluded on purpose from the fixed utterances in the first phase of the app due to their low audio quality and hardly received any rating in the second phase of the app. The 16 fixed utterances were always presented in the same order – with some slight exceptions due to hardware memory issues in the first phase of the study – leading to potential order effects. The study subjects did not have the possibility to rate models against each other. They listened to an ATCo and a pilot audio and then rated the two voice models. In addition, the judgements on voice models might have been different if audio outputs would have been used in an actual human-in-the-loop ATC simulation with actively controlling ATCos or flying pilots.

I. Analysis of Textual ATCo Utterance Length vs. Synthesis Time

The different hardware setups must be considered for interpreting the inference time results. Initially, the TTS app was hosted on a standard laptop with the following specifications:

- CPU: Intel(R) Core(TM) i7-10850H CPU @ 2.7 GHz, 6 cores, 12 threads
- GPU: Nvidia Quadro T1000 4 GB PCIe
- Memory: 31 GB RAM, 2 GB Swap
- Architecture: x86_64, 32-bit & 64-bit modes

Due to memory constraints on this laptop, each inference required the TTS models to be reloaded to the working memory, leading to longer and more variable inference times for the first study subjects. This is reflected in the broader dispersion of inference times for a given number of tokens. The majority of utterances took 10 to 20 seconds to be synthesized. However, the longest utterance with more than 100 tokens needed 50 seconds[19]. Subsequently, the TTS app was migrated to a high-powered computation server with the following specifications:

- CPU: AMD EPYC 7502 32-Core Processor, 32 cores per socket, 2 sockets, 128 threads
- GPU: Nvidia A100 80 GB PCIe & Tesla T4 16 GB
- Memory: 1 TB RAM, 3.8 GB Swap
- Architecture: x86_64, 32-bit & 64-bit modes

With the enhanced resources, the models usually remained preloaded, streamlining the inference process significantly. This transition is evident in the concentration of data points towards the lower end of the inference time spectrum (see Fig. 3). The majority of utterances took 1 to 3 seconds to be synthesized with the longest utterance requiring up to 7 seconds. This illustrates the substantial impact that dedicated hardware resources can have on the operational efficiency of the app, especially as the complexity and length of the input text increase.

J. Analysis of Real-Time Factor

The Real-Time Factor (RTF) is a crucial metric in the field of speech processing. It provides a measure of the speed of a speech processing system relative to real-time. In the context of speech processing, "real-time" refers to the ability of the system to process speech data at the same rate as it is produced. The RTF is calculated as the ratio of the total processing time to the total speech duration. Both the processing time and speech duration are typically measured in seconds. An RTF of less than one indicates that the system is processing speech faster than real-time, while an RTF greater than one indicates that the system is slower than real-time.

As explained above, the performance of a speech processing system varies significantly depending on the computational resources of the machine it runs on. This is confirmed by the following analysis. The RTF on the standard laptop using CPU for speech synthesis was almost 3 while it was just around 0.5 on the computation server using its GPU - a factor of six based on computational resources. This should be considered for potential operational usage of further TTS developments for ATCC as one important aspect for future applications is the real-time operational demand. For simplicity, an average length ATC utterance duration of 5 seconds is

[19]The inference times on the standard laptop are not shown in Fig. 3 as this setup is not recommended for any training or operational use due to its bad performance.

assumed. This is even more than the 4.8 seconds reported for en-route ATCo's initial calls [55], more than the 4.5 seconds reported for radio utterances in approach, and more than the 4.3 seconds reported for radio utterances in the tower environment [56]. Given the calculated real-time factor of the TTS app of 0.5, an average ATC utterance would require 2.5 seconds to be synthesized from text to speech. In a US en-route environment, a pilot readback delay of 3.3 seconds was found, i.e., the time between the end of the ATCo utterance and the start of the pilot utterance [55].

When thinking of an application that synthesizes a pilot response based on an ATCo instruction, it would realistically feel like real-time capability if the speech synthesis takes 2.5 seconds in average and the resulting audio would be played after further 0.8 seconds buffer time. Of course, there are certain limitations to this calculation: (1) readback delay times in en-route might be longer than in tower/approach where pilots expect more frequent ATCo utterances [55], (2) average times do not "consider" very long utterances or very short readback delays, and (3) there might be further processing time needed, e.g., for speech-to-text processing of the prior ATCo utterance as the input for the text-to-speech synthesis.

To also tackle the "non-average duration" utterances, there exist an easy method to improve the TTS speed. The majority of word compounds that are usually found in ATC radio utterances can be prepared, i.e., the speech synthesis can start after aircraft information is available via surveillance data. Then, the callsign of this aircraft could already be synthesized and saved as an audio file for later playback as part of a complete TTS message. The same method is applicable for ATC command types and values, e.g., to pre-synthesize "turn right heading three six zero", "turn right heading three five zero", ... "descending flight level eight zero", "descending flight level seven zero", ... "direct bikmu", "direct domux", ... "contact one one eight decimal three", or even "lufthansa seven hotel victor contact one one eight decimal three", etc.

The concatenation of existing word compounds would accelerate the generation of the complete speech especially if the computer-generated texts stick more to the ICAO phraseology than to individual phraseology of humans. The pre-synthesis technique is probably not feasible for single words. This means that pre-synthesizing of "lufthansa" to be concatenated with pre-synthesized letters and numbers to a callsign might sound too much different from the usual ATC tone with quickly following syllables dependent of the predecessor and successor syllable. However, word compounds are usually as well grouped melodically in human radio ATC utterances so that the pre-synthesizing technique could be a reasonable speed improvement technique without a loss of realism. First non-representative tests in an ATC simulator support the assumed feasibility of pre-synthesized utterances.

VII. CONCLUSION AND FUTURE WORK

We presented the development and evaluation of an AI-based text-to-speech (TTS) application aiming to simulate aviation radio telephony communication partners. To the best of the authors' knowledge, this is one of the first attempts to construct such a close-to-real-life communication simulation in the domain of air traffic control not only including ICAO standard phrases, but also utterances and speech quality as in actual operations with intensive ratings on the synthesized speech quality. Compared to the literature, our work is open-source and reproducible. It also received a human evaluation on the synthesized cutting-edge technology speech from operational experts and the artificial voices supported not just one but both communication partners in air traffic control – ATCo and pilot. The app utilized open-source TTS models, fine-tuned using publicly available datasets of ATCo-pilot voice recordings.

Analysis of subjective data from 20 study subjects on 21 ATCo and pilot voice models, i.e., textual comments and more than 4100 rating values, indicate the potential of the app, while also suggesting some areas for further refinement.

Analysis of the speech quality metrics overall experience, clarity, pronunciation, intonation, and naturalness mainly for the 16 fixed ATCo utterances and pilot responses synthesized with those voice models provided insights into their performance. Four voice cloning models for ATCos based on female Swiss speaker data with German and French accent achieved the highest overall score on ATCo utterances in the range of 4.1 to 4.4 out of a total score of 5.

Also a male voice model trained on ATCo data achieved an intonation score on a synthesized pilot response of 4.5 and a naturalness score of 4.4, respectively. The best model trained on pilot speaker data with Serbian accent received an overall score of 4 and a clarity score of 4.5. For both, ATCo and pilot models, the ATC typical intonation was the major factor for a high overall score. Interestingly, the *Untrained Voice* model, despite receiving a high clarity score of 4.1, scored much lower in other metrics as indicated with a naturalness score of 2.9. This demonstrates the need and significance of fine-tuning voice models with aviation-specific data. The average overall rating for voice cloning models (XTTS) was significantly better than for end-to-end models (VITS). It could be explored if a hybrid approach that combines voice cloning and end-to-end models improve the average TTS performance even if it is not expected due to the big performance difference of the two approaches. Female XTTS models fine-tuned on ATCo data were rated significantly better overall than male XTTS models.

Study subjects were mainly confident with the articulation speed of synthesized speech. However, it can also be beneficial to use slow rated voice models for initial aviation operator training phases and the faster rated voice models at a more advanced stage.

Before targeting higher technology readiness levels, the TTS app should undergo further testing and evaluation based on a broader set of utterances. Future ATCC corpora should be more inclusive to reflect the global nature of aviation communication, covering a wider range of accents, dialects, languages, scenarios, and ATC environments. The latter means to make sure that training data is not only from simulations and focusing on en-route environment, but is covering the complete ATC domain from ground, via tower, to approach, en-route, and even oceanic traffic control as well as comprise operational recordings with utterances for aircraft or airspace situations even in non-nominal conditions if possible. Further studies could evaluate the long-term effectiveness and adaptability of TTS systems in diverse aviation environments. Also ethical implications of synthetic voice communication in ATC, focusing on trust, reliability, and accountability in critical scenarios might need deeper analysis.

Furthermore, the reported differences in performance of accurately detecting readback errors highlights the potential of using an automated TTS app to train precisely hearing back in aviation communication.

With enough computation power, the real-time factor reached 0.5 with average-length utterances being synthesized faster than real-time in roughly two seconds. This shows that the technology can not only be valuable for use in training, but also for operational downstream applications. Those applications could include speech generation for ATCo utterances, simulation pilot/ATCo utterances, or pilot responses to support the human aviation operators in certain use cases. An example use case could be issuing specific ATC command types, reducing the number of human simulation pilots if supported by automatic readbacks through a combination of speech-to-text and text-to-speech, or generating the initial pilot call on the next ATC frequency. Such use cases could lead to a reduction of aviation operator workload as they do not need to verbalize these utterances themselves anymore.

While the current TTS app generates pilot responses from textual ATCo utterances using regular expressions, its static nature poses challenges in adapting to the variability inherent in ATC communications. Further enhancements should not only use word-by-word transcriptions of ATCo utterances, but their machine-readable, extracted semantic meaning. This would ease deriving the pilot response text from semantic annotations within a defined ontology for ATC utterances. More precisely, this would help that "corrections", wind, or greeting/farewells from the ATCo utterance would not be repeated by the pilot response. If pilots are asked to report some flight characteristic, this could be automatically looked up and integrated into the pilot response. In case the ATCo is targeting multiple pilots ("break break" commands), there could me multiple readbacks with different pilot voice models.

In addition, employing advanced AI models for text and speech processing, such as open-source Large Language models (LLM) like LLaMA (Large Language Model Meta AI) [57] could advance the generation of pilot responses and ATCo utterances in general. These models perform well in instruction-following and language understanding and could be fine-tuned with ATC-specific data to act as a pilot, responding to a wide array of ATC commands in a dynamic and contextual manner. Such open-source models are offering a cost-effective alternative to paid models like GPT-4 [58] that has already been explored for testing some ATC utterance examples with the described approach. Incorporating real-world noise into pilot voice simulations could as well enhance the user experience.

REFERENCES

[1] M. Eccles, "Europe's air traffic controllers are falling off the radar." http://s.dlr.de/s6aC9, 2023.

[2] M. Placek, "The pilot shortage - statistics & facts." http://s.dlr.de/QTdOH, 2024.

[3] ICAO, *Doc 4444 – Procedures for Air Navigation Services – Air Traffic Management*. Montréal, QC, Canada: International Civil Aviation Organization, 16 ed., 2016.

[4] J. A. Updegrove and S. Jafer, "Optimization of Air Traffic Control Training at the Federal Aviation Administration Academy," *Aerospace*, vol. 4, no. 4, 2017. https://doi.org/10.3390/aerospace4040050.

[5] J. Updegrove and S. Jafer, "Recommendations for next generation air traffic control training," in *IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*, DASC 2017, (St. Petersburg, FL, USA, 17-21 Sep), 2017. https://doi.org/10.1109/DASC.2017.8102129.

[6] K. Hofbauer, S. Petrik, and H. Hering, "The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, and D. Tapias, eds.), LREC 2008, (Marrakech, Morocco, 28-30 May, 2008), European Language Resources Association (ELRA), 2008. http://s.dlr.de/mKsCN.

[7] J. Matoušek and D. Tihelka, "English TTS speech corpus of air traffic (pilot) messages - Czech/German/Serbian/Taiwanese accent," 2014. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-1587 [1588/1462/1461].

[8] M. Jůzová and D. Tihelka, "Minimum Text Corpus Selection for Limited Domain Speech Synthesis," in *Text, Speech and Dialogue* (P. Sojka, A. Horák, I. Kopeček, and K. Pala, eds.), (Cham, Switzerland), pp. 398–407, Springer International Publishing, 2014. https://doi.org/10.1007/978-3-319-10816-2_48.

[9] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, (Virtual Conference, 18–24 Jul), pp. 5530–5540, PMLR, 2021. http://s.dlr.de/cwo2z.

[10] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, vol. 3. Luxembourg: Springer Science & Business Media, 1997.

[11] E. Manaker, "Pilot ability to understand synthetic voice and radio voice when received simultaneously," Tech. Rep. ACT-R-82-O1, Grumman Aerospace Corporation, New York, NY, USA, 1982. http://s.dlr.de/LABw3.

[12] J. L. Wheale, "The Speed of Response to Synthesized Voice Messages," *British Journal of Audiology*, vol. 15, no. 3, pp. 205–212, 1981. https://doi.org/10.3109/03005368109081439.

[13] C. A. Simpson and D. H. Williams, "Response Time Effects of Alerting Tone and Semantic Context for Synthesized Voice Cockpit Warnings," *Human factors*, vol. 22, no. 3, pp. 319–330, 1980. https://doi.org/10.1177/001872088002200306.

[14] S. E. Stern, J. W. Mullennix, C. lynn Dyson, and S. J. Wilson, "The Persuasiveness of Synthetic Speech versus Human Speech," *Human Factors*, vol. 41, no. 4, pp. 588–595, 1999. https://doi.org/10.1518/001872099779656680.

[15] I. R. Murray, J. L. Arnott, and E. A. Rohwer, "Emotional stress in synthetic speech: Progress and future directions," *Speech Communication*, vol. 20, no. 1, pp. 85–91, 1996. https://doi.org/10.1016/S0167-6393(96)00046-5.

[16] L. Dhavala, "Use of Synthetic Voice to Improve Communication between Air Traffic Controllers and Pilots," tech. rep., Emirates Aviation University, Dubai, UAE, 2014. http://s.dlr.de/4MZuC.

[17] D. E. Thorburn, "Voice Warning Systems – A Cockpit Improvement that should not be overlooked," Tech. Rep. AMRL-TR-70-138, Aerospace Medical Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, OH, USA, 1971. http://s.dlr.de/TN5Tb.

[18] K. Qureshi, "Exploration of text-to-speech applications in Aviation," Tech. Rep. Master Level Seminar Paper, Technische Universität Clausthal, Department of Computer Science, Clausthal-Zellerfeld, Germany, 2023.

[19] M. Slotty and O. Rühl, "Speech recognition finds its way into DFS - Procedure for the introduction of speech recognition in research and training applications; original German title: Spracherkennung findet Einzug in der DFS - Vorgehensweise bei der Einführung von Spracherkennung im Forschungs- und Trainingseinsatz," *TE im Fokus*, no. 2, pp. 31–37, 2012. http://s.dlr.de/8I72H.

[20] H. Helmke, O. Ohneiser, T. Mühlhausen, and M. Wies, "Reducing Controller Workload with Automatic Speech Recognition," in *IEEE/AIAA 35th Digital Avionics Systems Conference*, DASC 2016, (Sacramento, CA, USA, 25-29 Sep), 2016. https://doi.org/10.1109/DASC.2016.7778024.

[21] H. Helmke, M. Slotty, M. Poiger, D. F. Herrer, O. Ohneiser, N. Vink, A. Černá, P. Hartikainen, B. Josefsson, D. Langr, R. García Lasheras, G. Marin, O. G. Mevatne, S. Moos, M. N. Nilsson, and M. B. Pérez, "Ontology for Transcription of ATC Speech Commands of SESAR 2020 Solution PJ.16-04," in *IEEE/AIAA 37th Digital Avionics Systems Conference*, DASC 2018, (London, United Kingdom, 23-27 Sep), 2018. https://doi.org/10.1109/DASC.2018.8569238.

[22] H. Helmke, M. Kleinert, S. Shetty, O. Ohneiser, H. Ehr, H. Arilíusson, T. S. Simiganoschi, A. Prasad, P. Motlíček, K. Veselý, K. Ondřej, P. Smrz, J. Harfmann, and C. Windisch, "Readback Error Detection by Automatic Speech Recognition to Increase ATM Safety," in *14th USA/Europe Air Traffic Management Research and Development Seminar*, ATM 2021, (Virtual, 20-24 Sep), 2021. http://s.dlr.de/LtMbB.

[23] H. Helmke, K. Ondřej, S. Shetty, H. Arilíusson, T. S. Simiganoschi, M. Kleinert, O. Ohneiser, H. Ehr, and J. P. Zuluaga-Gómez, "Readback Error Detection by Automatic Speech Recognition and Understanding - Results of HAAWAII project for Isavia's Enroute Airspace," in *12th SESAR Innovation Days*, SID 2022, (Budapest, Hungary, 05-08 Dec), 2022. http://s.dlr.de/9NrZG.

[24] P. Heinrich, D. Hollerer, C. Karthaus, and M. Gellrich, ""'Don't drop the plane to fly the mic!" – Designing for Modern Radiotelephony Education in General Aviation," in *21. Fachtagung Bildungstechnologien (DELFI)*, pp. 187–192, Bonn, Germany: Gesellschaft für Informatik e.V., 2023. https://doi.org/10.18420/delfi2023-30.

[25] G. Taylor, J. Miller, and J. Maddox, "Automating Simulation-Based Air Traffic Control," in *Interservice/Industry Training, Simulation, and Education Conference*, no. 2193 in I/ITSEC 2005, (Orlando, FL, USA, 28 Nov-1 Dec), 2005. http://s.dlr.de/0iDpS.

[26] T. S. Schmeier, "Multi-modality at the controller working position - Selection and integration of text-to-speech software into a prototype human-machine interface for air traffic controllers; original German title: Multimodalität am Fluglotsenarbeitsplatz - Auswahl und Integration einer Text-To-Speech-Software in eine prototypische Mensch-Maschine-Schnittstelle für Fluglotsen," Tech. Rep. DLR-IB-FL-BS-2016-81, German Aerospace Center (DLR), Institute of Flight Guidance, Department Controller Assistance, Braunschweig, Germany, 2016.

[27] T. Auinger, "Design and implementation of a simulated aircraft for air traffic control communication training," Master's thesis, Paris Lodron Universität Salzburg, Salzburg, Austria, 2019. http://s.dlr.de/b1Etd.

[28] N. Ahrenhold, H. Helmke, T. Mühlhausen, O. Ohneiser, M. Kleinert, H. Ehr, L. Klamert, and J. P. Zuluaga-Gómez, "Validating Automatic Speech Recognition and Understanding for Pre-Filling Radar Labels - Increasing Safety While Reducing Air Traffic Controllers' Workload," *Aerospace*, vol. 10, no. 6, 2023. https://doi.org/10.3390/aerospace10060538.

[29] R. Tarakan, K. Baldwin, and N. Rozen, "An Automated Simulation Pilot Capability to Support Advanced Air Traffic Controller Training," in *The 26th Congress of ICAS and 8th AIAA ATIO*, (Anchorage, AK, USA, 14-19 Sep), 2008. https://doi.org/10.2514/6.2008-8897.

[30] A. Prasad, J. P. Zuluaga-Gómez, P. Motlicek, S. Sarfjoo, I. Nigmatulina, and K. Vesely, "Speech and Natural Language Processing Technologies for Pseudo-Pilot Simulator," in *12th SESAR Innovation Days*, SID 2022, (Budapest, Hungary, 05-08 Dec), 2022. http://s.dlr.de/28nwA.

[31] J. P. Zuluaga-Gómez, A. Prasad, I. Nigmatulina, P. Motlíček, and M. Kleinert, "A Virtual Simulation-Pilot Agent for Training of Air Traffic Controllers," *Aerospace*, vol. 10, no. 5, 2023. http://doi.org/10.3390/aerospace10050490.

[32] J. Zhang, P. Zhang, D. Guo, Y. Zhou, Y. Wu, B. Yang, and Y. Lin, "Automatic repetition instruction generation for air traffic control training using multi-task learning with an improved copy network," *Knowledge-Based Systems*, vol. 241, p. 108232, 2022. https://doi.org/10.1016/j.knosys.2022.108232.

[33] Y. Lin, Y. Wu, D. Guo, P. Zhang, C. Yin, B. Yang, and J. Zhang, "A Deep Learning Framework of Autonomous Pilot Agent for Air Traffic Controller Training," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 5, pp. 442–450, 2021. http://doi.org/10.1109/THMS.2021.3102827.

[34] P. Stanislav, L. Šmídl, and J. Švec, "An Automatic Training Tool for Air Traffic Control Training," in *17th Annual Conference of the International Speech Communication Association*, Interspeech 2016, (San Francisco, CA, USA, 08-12 Sep), pp. 782–783, ISCA, 2016. http://s.dlr.de/VgM6m.

[35] L. Šmídl, J. Švec, A. Chýlek, D. Tihelka, J. Matoušek, P. Stanislav, A. Pražák, P. Ircing, and J. Psutka, *Talking with artificial pilot : A dialogue system for training air traffic controllers*, pp. 194–200. Brno, Czechia: Tribun EU, 2019. http://s.dlr.de/Odtc1.

[36] M. Lowry, T. Pressburger, D. A. Dahl, and M. Dalal, "Towards Autonomous Piloting: Communicating with Air Traffic Control," in *AIAA Scitech 2019 Forum*, (San Diego, CA, USA, 07-11 Jan), 2019. https://doi.org/10.2514/6.2019-2207.

[37] K. Ito and L. Johnson, "The LJ Speech Dataset." http://s.dlr.de/4wadz, 2017.

[38] J. Chen, L. Ye, and Z. Ming, "MASS: Multi-task anthropomorphic speech synthesis framework," *Computer Speech & Language*, vol. 70, p. 101243, 2021. https://doi.org/10.1016/j.csl.2021.101243.

[39] B. Yang, X. Tan, Z. Chen, B. Wang, M. Ruan, D. Li, Z. Yang, X. Wu, and Y. Lin, "ATCSpeech: A Multilingual Pilot-Controller Speech Corpus from Real Air Traffic Control Environment," in *21st Annual Conference of the International Speech Communication Association*, Interspeech 2020, (Shanghai, China, 25–29 Oct), pp. 399–403, ISCA, 2020. http://dx.doi.org/10.21437/Interspeech.2020-1020.

[40] E. Delpech, M. Laignelet, C. Pimm, C. Raynal, M. Trzos, A. Arnold, and D. Pronto, "A Real-life, French-accented Corpus of Air Traffic Control Communications," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, eds.), LREC 2018, (Miyazaki, Japan, 07-12 May), European Language Resources Association (ELRA), 2018. http://s.dlr.de/WOGFx.

[41] J. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P. Breton, V. Clot, R. Gemello, M. Matassoni, and P. Maragos, "The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication," *Online*, 2007. http://s.dlr.de/qrRgn.

[42] L. Šmídl, J. Švec, D. Tihelka, J. Matoušek, J. Romportl, and P. Ircing, "Air traffic control communication (ATCC) speech corpora and their use for ASR and TTS development," *Language Resources and Evaluation*, vol. 53, no. 3, pp. 449–464, 2019. https://doi.org/10.1007/s10579-019-09449-5.

[43] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High Fidelity Speech Synthesis with Adversarial Networks," in *International Conference on Learning Representations*, ICLR, 2020. http://s.dlr.de/aDICb.

[44] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019. http://doi.org/10.1561/2200000056.

[45] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, (Red Hook, NY, USA, Vancouver/Online, BC, Canada, 06-12 Dec, 2020), Curran Associates Inc., 2020. http://s.dlr.de/chQ1B.

[46] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, (Red Hook, NY, USA, Vancouver/Online, BC, Canada, 06-12 Dec), Curran Associates Inc., 2020. http://s.dlr.de/zo4C0.

[47] K. Hofbauer and S. Petrik, "ATCOSIM Air Traffic Control Simulation Speech Corpus," Tech. Rep. TR TUG-SPSC-2007-11, Graz University of Technology, Austria, 2008. http://s.dlr.de/lEn6Y.

[48] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He, S. Zhao, T. Qin, F. Soong, and T. Liu, "NaturalSpeech: End-to-End Text-to-Speech Synthesis With Human-Level Quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 4234–4245, Jun 2024. https://doi.org/10.1109/TPAMI.2024.3356232.

[49] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, "Likert Scale: Explored and Explained," *Current Journal of Applied Science and Technology*, vol. 7, no. 4, pp. 396–403, 2015. https://doi.org/10.9734/BJAST/2015/14975.

[50] M. Ragnarsdottir, H. Waage, and E. Hvannberg, "Language Technology in Air Traffic Control," in *IEEE/AIAA 22nd Digital Avionics Systems Conference (DASC)*, vol. 1 of *DASC 2003*, (Indianapolis, IN, USA, 12-16 Oct), pp. 2.E.2–21–13, 2003. https://doi.org/10.1109/DASC.2003.1245815.

[51] S. Warhurst, C. Madill, P. McCabe, R. Heard, and E. Yiu, "The Vocal Clarity of Female Speech-Language Pathology Students: An Exploratory Study," *Journal of Voice*, vol. 26, no. 1, pp. 63–68, 2012. https://doi.org/10.1016/j.jvoice.2010.10.008.

[52] Y. Takefuta, E. G. Jancosek, and M. Brunt, *A Statistical Analysis of Melody Curves in the Intonation of American English*, pp. 1035–1039. Montréal, Canada, 22–28 Aug 1971, Berlin, Boston: A. Rigault and R. Charbonneau (Eds.), De Gruyter Mouton, 1972. https://doi.org/10.1515/9783110814750-142.

[53] A. Schirmer, M. H. Chiu, C. Lo, Y.-J. Feng, and T. B. Penney, "Angry, old, male – and trustworthy? How expressive and person voice characteristics shape listener trust," *PLOS ONE*, vol. 14, pp. 1–16, 01 2019. https://doi.org/10.1371/journal.pone.0210555.

[54] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?," in *6th Annual Conference of the International Speech Communication Association*, Interspeech 2005, (Lisbon, Portugal, 4-8 Sep), pp. 2205–2208, ISCA, 2005. https://doi.org/10.21437/Interspeech.2005-699.

[55] K. M. Cardosi and P. W. Boole, "Analysis of Pilot Response Time to Time-Critical Air Traffic Control Calls," Tech. Rep. DOT/FAA/RD-91/20 DOT-VNTSC-FAA-91-12, Department of Transportation, Federal Aviation Administration, USA, 1991. http://s.dlr.de/8nLM6.

[56] T. Pellegrini, J. Farinas, E. Delpech, and F. Lancelot, "The Airbus Air Traffic Control Speech Recognition 2018 Challenge: Towards ATC Automatic Transcription and Call Sign Detection," in *20th Annual Conference of the International Speech Communication Association*, Interspeech 2019, (Graz, Austria, 15-19 Sep), pp. 2993–2997, ISCA, 2019. https://doi.org/10.21437/Interspeech.2019-1962.

[57] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. https://arxiv.org/abs/2302.13971.

[58] OpenAI Team, "GPT-4 Technical Report," 2024. https://arxiv.org/abs/2303.08774.

Oliver Ohneiser received his bachelor's degree in information technology from Baden-Württemberg Cooperative State University Mannheim (Germany) in 2009, his master's degree in computer science as well as his doctorate degree (PhD) in Aerospace Engineering from the Technical University of Braunschweig (Germany) in 2011 and 2017, respectively.

He joined German Aerospace Center (DLR) in 2006 and is now with the department "Controller Assistance" of DLR's Institute of Flight Guidance in Braunschweig. He investigates modern interaction technologies at controller working positions and led projects concerning automatic speech recognition and understanding in air traffic management. He completed a three-month research semester at Federal Aviation Administration (FAA) William J. Hughes Technical Center in Atlantic City, NJ, USA in 2015.

Dr. Ohneiser has been a private lecturer for aeronautical informatics at Clausthal University of Technology (Germany) since 2021. He received several scholarships and best paper awards – among others three from AIAA/IEEE Digital Avionics Systems Conference (DASC) – and was ICAS-IFAR Award Third-Place Recipient in recognition of "excellent achievements" with his PhD thesis.

Umair Ahmed received his bachelor's degree from Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Karachi, Pakistan in 2015, and his master's degree from Clausthal University of Technology, Germany in 2024 – both in computer science.

He has been teaching as a lecturer of computer science at the Lasbela University of Agriculture, Water, and Marine Sciences (LUAWMS) in Pakistan since 2018. Meanwhile, he is also associated as an IT System Administrator with CELUS GmbH, a deep tech software company based in Munich, Germany, on a mission to accelerate innovations in electronic design, since 2022.

Mr. Ahmed has been awarded the overseas master's scholarship as part of the Faculty Development Program, to come to Germany and pursuing his Master's degree in computer science.

Comparison of Air Traffic Controller Display Techniques for Reaching Target Times at Significant Waypoints

Oliver Ohneiser
Institute of Flight Guidance,
Department Controller Assistance,
German Aerospace Center (DLR)
Lilienthalplatz 7,
38108 Braunschweig, Germany,
Oliver.Ohneiser@DLR.de

Vicki Ahlstrom
William J. Hughes Technical Center,
Human Factors Branch,
Federal Aviation Administration
Atlantic City International Airport,
NJ 08405, USA,
Vicki.Ahlstrom@FAA.gov

Kevin Tracy, Brett Williams
Hi-Tec Systems, Inc.
6727 Delilah Road,
Suite 100,
Egg Harbor Township,
NJ 08234, USA

***Abstract*—With the introduction of a time-based air traffic control approach, new display techniques are necessary to support controllers. This is especially true for ensuring compliance with aircraft target times at significant waypoints. In this paper, five different visual aids were examined in a small-scale, reduced complexity study: Slot Marker, Time-To-Gain/Lose, Timeline, TargetWindow, and a Baseline display. Sixteen study participants had to manage aircraft to arrive at waypoints on schedule. Simulation results showed that participants performed worst using no visual aid at all. Based on questionnaire responses, the Slot Marker display was most intuitive and easy to use. Measured performance parameters show that high time accuracy goes along with greater numbers of given commands. However, multiple commands cause inefficient flight trajectories due to speed adjustments. Therefore, a trade-off between high time accuracy and low economically reasonable command rates is necessary for visual aids in air traffic control.**

Keywords—Air Traffic Controller, Display Technique, Time-Based Air Traffic Management, Visual Aids

I. INTRODUCTION

Air traffic controllers traditionally use a distance-based approach for Air Traffic Management (ATM). In the future, this approach will transition in three steps to a time-based and later to a trajectory-based and performance-based approach, due to the implementation of NextGen (Next Generation Air Transportation System) in the US [1] and SESAR (Single European Sky ATM Research) in Europe [2], [3], [4]. For time-based and trajectory-based approaches to ATM, time differences and negotiated times have to be achieved with high accuracy [5].

Separation will still be the main technique for maintaining safety in air traffic. However, using a time-based instead of distance-based approach, in which aircraft are separated using time differences or meet assigned times at significant waypoints could increase capacity in ATM especially in headwind situations. New controller tasks, like estimating waypoint arrival times and timespans via ground speed and distance are hardly manageable with today's tools [6]. Furthermore, the trend of more passive monitoring instead of actively controlling air traffic [7] requires new support tools that do not lead to more fatigue [8]. Therefore, many visual display aids have been proposed to support controllers doing time-based ATM.

Those aids are calculated by controller support systems. With new display aids, controllers can reach defined target times by comparing actual and planned aircraft states or follow advised maneuvers. However, controllers still have the chance to deviate from support system's plan and implement their own solution if necessary. The benefits and shortcomings of those different proposed visual aids have not been thoroughly evaluated from a human factors perspective.

The study described in this paper compares five different alternatives for supporting air traffic controllers with time-based display aids. The simulation included basic characteristics of the real world application but was reduced in complexity to gather results quickly.

Section II describes different display techniques for time-based ATM. Section III outlines the design and questionnaires for the study. Results of a microworld study are presented in section IV. These results are discussed in section V, followed by conclusions in section VI.

II. RELATED WORK

Different display techniques for time-based air traffic management have been proposed over the years. All of them involve a calculation of target times or time differences. Modern controller support tools like arrival, departure, or surface managers have the ability to schedule all aircraft in a certain airspace or on ground.

Those systems can compute target times at selected waypoints due to a negotiated four-dimensional trajectory. Arrival times for threshold, Initial Approach Fixes (IAF), or merge points can be visualized as an alphanumeric time string.

A. *Time-To-Gain / Time-To-Lose*

Time-To-Gain or Time-To-Lose for significant waypoints can be derived from trajectories or a projection of current flight status. Resulting times are displayed differently in various systems. Times may represent seconds to gain/lose or Early-/Late-Indicators [9] (for an implemented version see figure 3). Values can be shown in the flight data block (FDB), on a timeline [10] with brackets indicating estimated and scheduled arrival times [11], or via a traffic light system to indicate the minutes of delay for each aircraft [12].

B. *Relative Position Indicator / Ghosts / Targets / TargetWindow / Slot Marker*

The projection of a virtual aircraft on a display is called a "ghost" [13]. Aircraft of one route (master/image reference line) are projected on another merging route (slave/target reference line) as a virtual aircraft ghost position [14], [15]. Separation between ghost and real aircraft on this route then shows relative temporal spacing between those objects. This was originally done for two arrival streams on converging runways simulating a dependent parallel approach [16].

Runway configuration changes can also be supported using relative position indicators [17]. Indicator target circles on a route visualize aimed positions for merging two arrival streams also taking into account several turns of an aircraft [18], [19]. Another approach is using "slot marker" circles to show the aircraft's expected position along its trajectory if it were conforming to the schedule [20], [21] (for an implemented version see figure 4). Similar target position indicators may also be used for certain waypoints in upper airspace, for wake vortexes [22], or for aircraft on several arrival routes to be mapped onto one centerline [23], [24]. Concepts are distinguishing between precisely hitting planned positions respectively target times (ghosts as very exact "values") and visual advisory positions or areas as targets for the controller to hit (guidance range).

A TargetWindow (for an implemented version see figure 5) is a marked interval on the centerline where it is safe for individually guided aircraft to be fed into the arrival stream by the controller [25]. Target positions in this window indicate the best positions after a turn-to-base maneuver. When aircraft are flying on a downwind, they will get a turn-to-base command to perform the base and final leg [26]. Target positions for those aircraft are indicated by a dotted semicircle with the open side facing the aircraft.

The surrounding TargetWindow symbolizes a safe area around this target position even if the aircraft does not hit its planned position exactly. Furthermore there is a buffer of half a mile, shown by a tapering of the TargetWindow at both ends. This helps ensuring that controllers do not violate separation minima from predecessors and successors.

C. *Timeline*

To visualize calculated time plans, controller assistance systems have one or multiple timelines for runway thresholds or IAFs with assigned label and runway information [12], [27], [28], [29]. These timelines may display scheduled and estimated times of arrival as well as the aircraft's sequence number within the arrival stream (for an implemented version see figure 6).

D. *Further Visual Aids and their Effects*

Credeur et al. evaluated three different final-approach spacing aid display formats and a baseline in a study with 12 controllers and a monochrome display [30]. This was the first comparing study for time-based ATC displays. A so called graphical advisory marker and a direct course error countdown showed best performance in this study. The graphical marker was preferred based on responses to the questionnaire. The centerline slot marker [31] performed worse than the two other final approach spacing aids and was most difficult to use. Nevertheless, all visual display aids reduced heading commands and increased the arrival rate by roughly 10% compared to the unaided scenario [30]. These results suggest that visual aids may also be beneficial for time-based metering tasks.

Having in mind all these visual aids, some questions succeed. It is of interest what effects visual support techniques have on timely accuracy and the number of commands to implement visualized suggestions. The time-space diagram for example is a controller support tool for speed control of continuous descent approaches to avoid many tactical speed and altitude commands [32]. Additional displays that effectively help controllers with their specific tasks could also lead to reduced workload and increased situation awareness [29], [33]. Situation awareness consists of steady maintenance of the controllers' mental picture with positions and movements of aircraft based on the mental model comprising airspace, aircraft, air traffic control (ATC) procedures, and human machine interaction [34].

III. Description of Study Setup and Design

We expect that controller assistance with visual aids improves the accuracy of meeting target times. There might also be a difference between presenting time information in spatial-graphical form (Slot Marker, Timeline, TargetWindow) or digital form only (Time-To-Gain/Lose, Baseline). Furthermore, the level of detail (e.g., displaying of seconds vs. more general time information via regions) of the visual aid may influence performance. A greater level of detail also visualizes differences between actual and target states better. This can lead to improved accuracy of times, but may negatively influence the number of commands in some display aids [20].

The aim of this study was to reveal general benefits and shortcomings of display techniques supporting time-based air traffic management. Chosen visual aids were either already operational in some air traffic control displays or prototypes. Five ways of assisting controllers to meet aircraft target times were evaluated: Baseline, Time-To-Gain/Lose, Slot Marker, TargetWindow, and Timeline.

The schedule with target times was fixed for all aircraft at the start of the scenario. The basic hypothesis was: The use of display techniques compared to the baseline, would support controllers time accuracy for metering aircraft to significant waypoints.

There were six "measures" of interest during the study: Percentage of aircraft arriving at waypoint with ±2 as well as ±5 seconds compared to target time, average difference between target time and realized time, number of losses of separation, number of speed commands, and a subjective rating on seven questionnaire items.

The five corresponding air traffic scenarios for the five displays were identical (very similar number and appearance of aircraft; only mixed order of appearance). This guarantees nearly the same conditions and allow for comparisons between display techniques. The combination of a display with its air traffic scenario is called configuration.

A. Simulation Layout and Participants

Basic layout used a dark background like in many actual controller displays (see figure 1). Three different routes (dark green lines with assigned names north (N), center (C), and south (S)) starting in the east joined in a single waypoint (merge point) in the west.

Angles between center/north and center/south, respectively, were approximately 45 degrees. All three routes had a length of 30 nautical miles from their beginning to the merge point, followed by one common five nautical mile strip on a westbound single route.

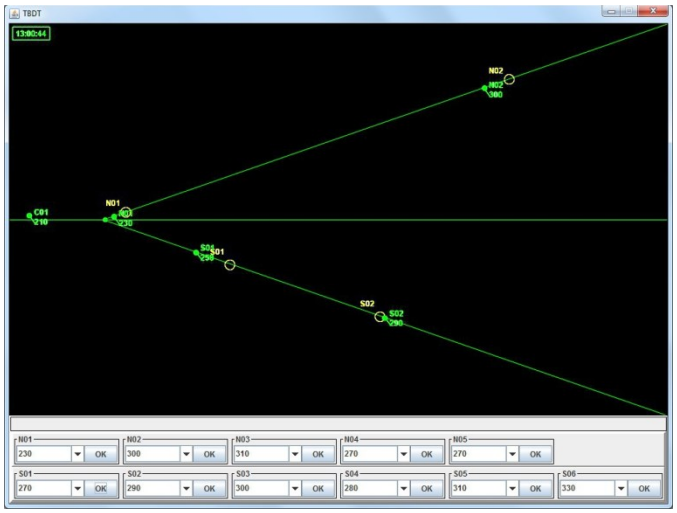


Figure 1. Microworld software in JAVA with merging routes and drop-down menus for aircraft speed commands, labels described in text

Aircraft were depicted as light green dots moving westwards on defined routes. Every aircraft had a call sign (like UAL123) in the first FDB line and actual speed in the second line. There was no differentiation between ground speed and indicated airspeed and no wind or weather influence in the scenarios. Hence, a given speed command affects aircraft such that it exactly flies the given speed value. Only aircraft on northern and southern routes could be influenced by giving speed instructions.

Participants used drop-down menus with selectable speeds ranging from 180 to 380 knots in steps of 10 knots. Communication between aircraft pilots and study participants, vectoring of aircraft with direction commands, changes in altitude, and other usual air traffic commands were not included in the study.

The task consisted of metering aircraft as if meeting the requirements of a scheduling system. Participants were told to give as few speed commands as possible to the aircraft while still meeting given target times at the merge point as precisely as possible.

Aircraft that received speed commands immediately accelerated or decelerated at a rate of one knot per second. The display also had an update rate of once per second. Therefore, participants had immediate feedback on the effect of their command. They were able to adjust the effects with a new command until aircraft reached the merge point.

Before the main task, participants performed a training task, which required them to get aircraft to the merge point at a specific target time, with a countdown showing the time remaining. Participants were able to give speed commands and see how many seconds the aircraft gained or lost on its way to the merge point, depending on how much they reduced or increased aircraft speed. Through this task, participants received visual feedback on how well they performed during training.

During the study that took place in April 2015, sixteen (n=16) subjects with an average age of 47 years participated. Two of them were female, fourteen male. One participant was a retired tower and TRACON controller, the others were researchers in the field of air traffic control.

B. Display Techniques

The baseline display had a minimal amount of time information (third line of aircraft labels in figure 2).

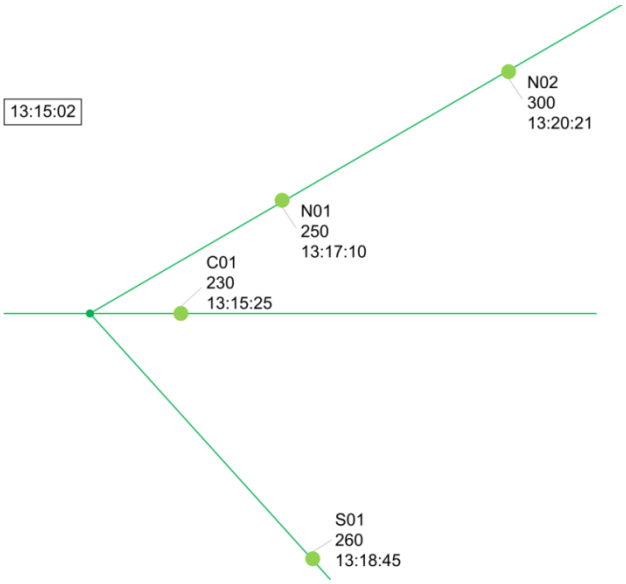


Figure 2. Baseline display

A display frame for current time with hours, minutes, and seconds was shown on the upper left side of the screen. A target time with hours, minutes, and seconds until the merge point was attached to every aircraft FDB in the third line. This target time did not change during the run.

Time-To-Gain or Time-To-Lose display showed the corresponding amount of seconds in the third FDB line (see figure 3). Times were calculated as differences between target times and projected times at merge point. The projected time took into account current speed and distance in nautical miles remaining until the merge point. The aircraft could be projected to reach the merge point too early, too late, or at the target time.

The time difference between target time and actual projected time changed as participants modified aircraft speed. As shown in figure 3, Time-To-Gain was depicted as a green alphanumeric value (e.g., +17), whereas Time-To-Lose was displayed as a red value (e.g., -10).

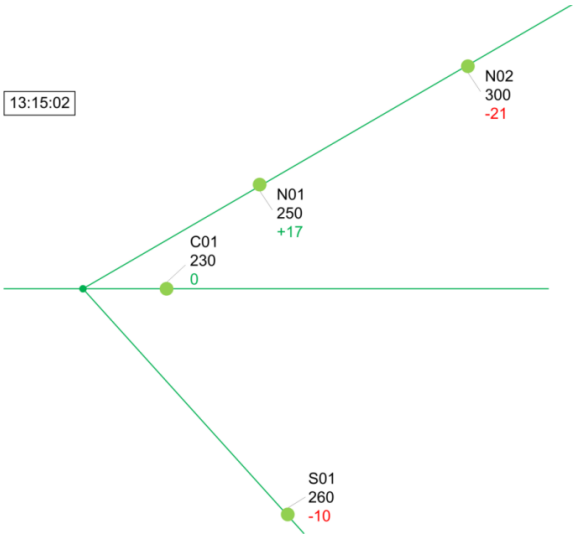


Figure 3. Time-To-Gain/Lose display

The Slot Marker configuration displayed call sign and actual speed (see figure 4).

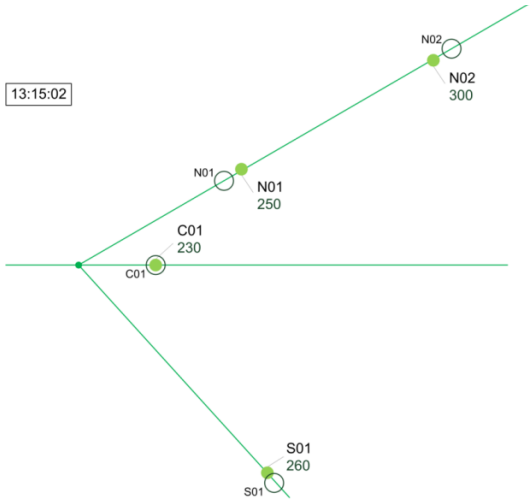


Figure 4. Slot Marker display

A thin dark yellow circle was displayed for every aircraft on northern and southern route. This Slot Marker circle represented the optimum position for the aircraft to reach the merge point exactly at the defined target time when flying at the current speed. Thus, a circle will be to the left of an aircraft symbol on screen if it has Time-To-Gain and to the right if it has Time-To-Lose.

The TargetWindow configuration displayed a shape with optimum positions on the center route (see figure 5). Thin dark yellow dashed lines highlighted safely separated areas between real aircraft on center route (in figure 5 yellow lines are shown in black). Semi circles with opening to the cardinal direction of join (top if aircraft joins from north route, bottom if aircraft joins from south route) pointed to the optimal target position regarding target time to be hit at merge point. The TargetWindow moved westward on the center route with ongoing time.

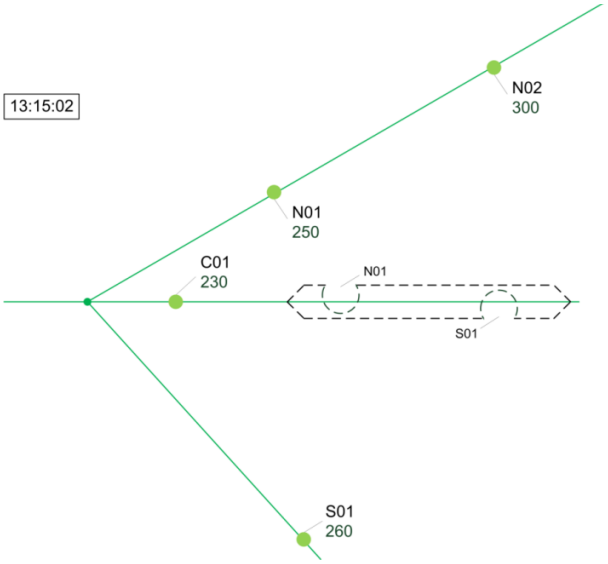


Figure 5. TargetWindow display

The Timeline configuration had a moving timeline with corresponding aircraft in the upper left corner of the display (see figure 6).

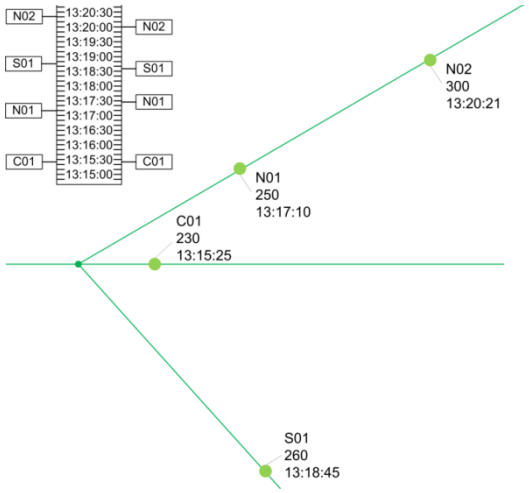


Figure 6. Timeline display

Actual time in hours, minutes, and seconds was shown at the bottom. 30 second steps were displayed from top to bottom of the timeline. On the left side, call signs of aircraft were attached with a thin line to their scheduled target time at the merge point. On the right side, call signs were attached to their projected times at the merge point. The vertical position of an aircraft's call sign in the left and right side may differ if the aircraft is predicted to be earlier or later than its scheduled time at the merge point.

Comparing displays with those from the previous study of Credeur et al. [30], Baseline or Timeline are conceptually like corresponding "manual" scenario, but with defined target times. Time-To-Gain/Lose display could be compared to the direct course error countdown with an Early-/Late-Indicator if given advisories are directly executed. The Slot Marker and the TargetWindow are to some extent similar to the Centerline Slot Marker configuration.

However, the study by Credeur et al. evaluated visual aids for final spacing with turn advisories from downwind over base leg to final. The study in this paper differs in several points. Display techniques are directly used to meet target times at a merge point, which is not necessarily placed in the last phase of approach.

Furthermore, electronic technologies and display techniques have changed during the last decades (e.g., TargetWindow is new). These changes may influence the type of display presentation that is most effective in this simulation.

In all displays, a steady comparison between actual and target state was possible in different ways, such as monitoring alphanumeric times, values of seconds, slots on same routes, target positions on a center route, or different vertical positions on a timeline.

C. Tasks of the Study Participants

Participants performed six 10-minute simulation runs. The first was the training run, followed by one run with each of the five visual aids (Baseline, Time-To-Gain/Lose, Slot Marker, TargetWindow, and Timeline). The order of these five runs was counterbalanced. Each run had a unique but similar traffic scenario.

Several aircraft approached the merge point from the north, center, and south routes in different alternations, so there was a steady flow of aircraft from all directions. Aircraft were initially separated 75 seconds from each other. Target times had a difference of at least 75 seconds. Initial speed was within a range of 230 and 330 knots.

A violation of minimum separation of three nautical miles was therefore possible. Two modifications were made to increase difficulty of the task in later levels. First, speed of aircraft increased for aircraft appearing later. Second, differences between target time and initial projected time at merge point were increased. For each participant, user inputs and resulting changes were captured in a log file.

D. Questionnaires

Participants had to fill out a questionnaire after each run. Seven statements were presented with alternating positive and negative formulations. Participant rated these statements on a five-point Likert scale [35]: Strongly disagree (0), Disagree (1), Neither agree nor disagree (2), Agree (3), Strongly agree (4).

An additional field was provided for remarks and suggestions. Questionnaire items were as follows:

- 1. The last display technique helped me meet aircraft target times at the merge point.
- 2. I liked the visualization of the last display technique in general.
- 3. The last display technique or some of the displayed elements were confusing.
- 4. The optimal use of the visual display aid to fulfill the task of meeting aircraft target time was not clear to me.
- 5. It was clear to me how much to increase or decrease speed of aircraft using the visual elements of last display technique.
- 6. Colors, shape, and font of last display technique were insufficient. So for example colors could not be perceived well, shape was not explicit enough or font was too small.
- 7. I was able to assign each displayed visual support element on screen to its corresponding aircraft (e.g., flight data block entries, markers, flight approach directions, or timeline labels).
- 8. Remarks on last display technique or suggestions for improvement.

Participant performance was evaluated based on three metrics:

- Difference between actual arrival time and target time at merge point
- Total number of commands
- Total number of separation minima violations

We were also interested in how performance changed over time.

IV. Results of Study

In this section, we present the results comprising subjective ratings, time difference, number of speed commands, number of losses of separation, and an overall badness score rating.

A. Subjective Ratings

The average of the sixteen participant ratings on the statements 1 to 7 of the questionnaire can be seen in figures 7 and 8. The scale was inverted for the negative formulated statements 3, 4, and 6, so higher values are better for all items.

Therefore bar heights representing ratings above the mean scale value $\mu_0 = 2$ show a positive response to questionnaire statements. Black lines depict positive and negative standard deviations σ for each item.

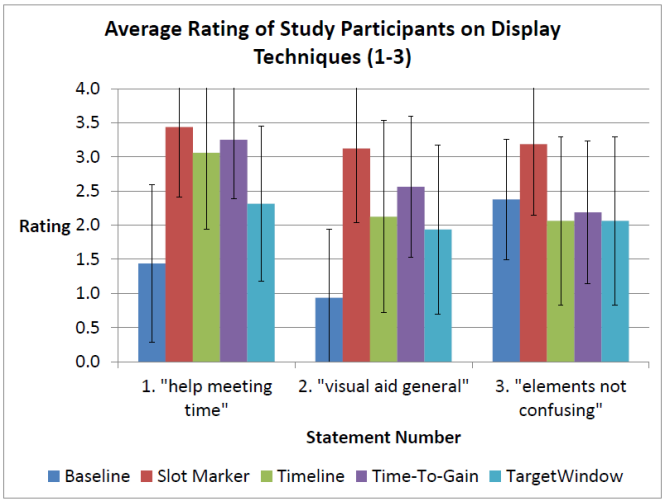


Figure 7. Average ratings on questionnaire statements 1 to 3 (with standard deviations)

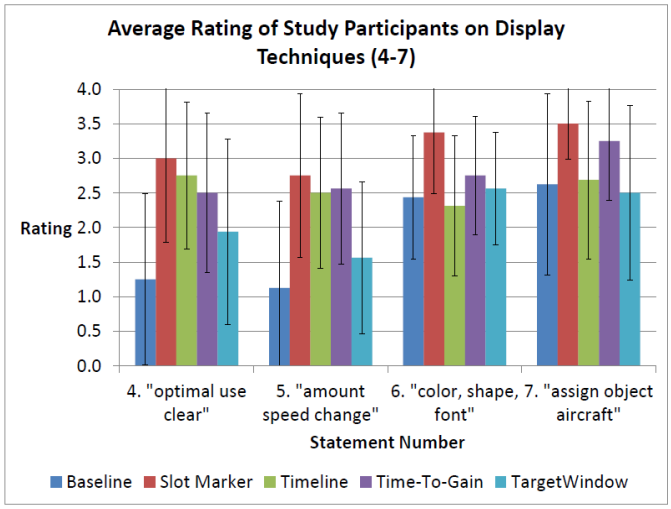


Figure 8. Average ratings on questionnaire statements 4 to 7 (with standard deviations)

The five-step Likert scale is interpreted as an ordinal scale with the sample size n. Assuming a uniform distribution of answers, we use the Wilcoxon signed-rank test as a non-parametric statistical hypothesis test (see figure 9). These results provide evidence against the null hypothesis that all visual aids have equal subject approval ($p < 0.005$ for items 1-3, $p < 0.05$ for items 4-7).

Assuming a normal distribution of answers, the arithmetic average value $\bar{X}$ is tested against μ_0 with a one sample size t-test to get a trend of subjective participant opinions. The corresponding null and alternative hypotheses are as follows:

$$H_0: \mu \leq 2 \quad (1)$$

$$H_1: \mu > 2 \quad (2)$$

The t-value is computed as:

$$t = \frac{\bar{X}-\mu_0}{\sigma}\sqrt{n} \quad (3)$$

To state significance, a 97.5% confidence interval with a significance level of $\alpha = 0.025$ and n-1 degrees of freedom was taken into account. Significant positive acceptance is given by $t > 2.131$.

Concluding the null hypothesis can be rejected, the alternative hypothesis can be accepted. The analogue way is true for a significant negative rating of $t < -2.131$ (significance of t-values and Wilcoxon signed-rank test in figure 9 is quite the same).

No.	Baseline	Slot Marker	Timeline	Time-To-Gain	TargetWindow
1	-1.952	5.578	1.098	3.782	5.839
2	-4.259	4.137	-0.202	0.355	2.183
3	1.695	4.538	0.202	0.202	0.716
4	-2.423	3.303	-0.187	2.818	1.732
5	-2.782	2.535	-1.600	1.826	2.058
6	1.962	6.214	2.764	1.232	3.503
7	1.908	11.619	1.581	2.416	5.839

Figure 9. T-values of average ratings on statements 1 to 7 (red cells show negative, green cells positive, yellow cells no significance; underline indicates corresponding significance with Wilcoxon signed-rank test)

B. *Time Difference*

The average percentage of aircraft that met their target times within ±2 and ±5 seconds is shown in figure 10 for all five displays.

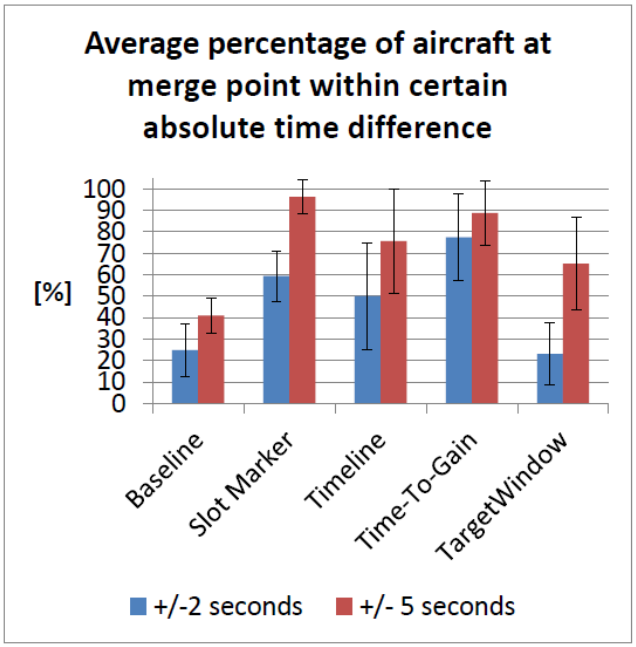


Figure 10. Average percentage of aircraft meeting target times within certain intervals (with standard deviations)

The box plot diagram in figure 11 shows the average time differences between scheduled and realized target time at the merge point ($\emptyset t_{diff}$). The lower end of the black line is the minimum, the upper end the maximum average time difference of a participant per display. The lower green border is the 25% quartile, the upper red border consists of the 75% quartile.

The median (border between red and green area) gives more detailed information about average time performance.

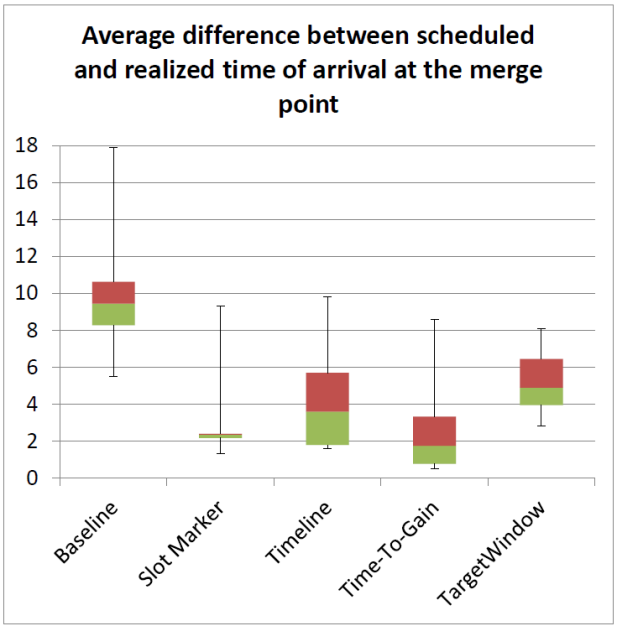


Figure 11. Average target time performance over all displays

C. Number of Speed Commands

Figure 12 is a box plot with the average number of speed commands (Cmd) of all participants per display.

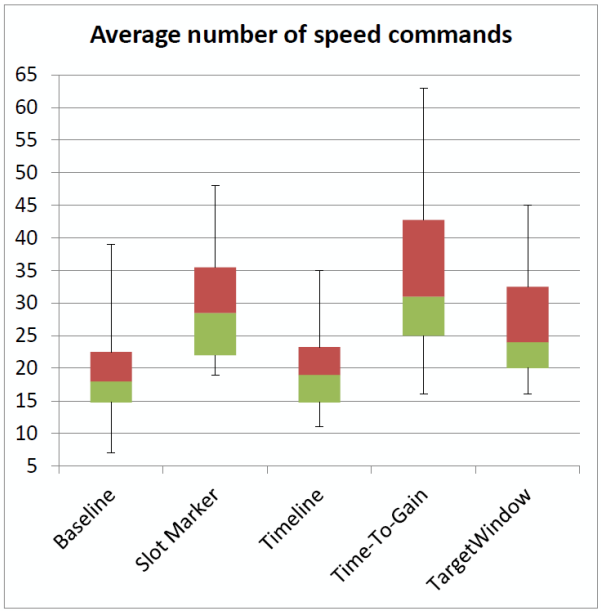


Figure 12. Average number of speed commands

D. Number of Losses of Separation

The average number of losses of separation (LoS) per display and subject was within a range of 0 to 1.2. The LoS per display was as follows: Slot Marker (0.1), TargetWindow (0.1), Time-To-Gain (0.2), Timeline (0.2), Baseline (0.9). Only four participants did not have any loss of separation in all of their five simulation runs.

E. Badness Score for Overall Rating

For comparison, we calculate an aggregate badness score per display (D) and participant (P). The higher the number of commands, the average time difference, and the number of losses of separation, the higher the badness score. The lower this score, the better the run of the participant. The bad point value V_D for each display is calculated with:

$$V_D = 2\left(\frac{\emptyset t_{diff_{D,P}}}{\emptyset t_{diff_{All\_D,P}}}\right) + \left(\frac{\#Cmd_{D,P}}{\#Cmd_{All\_D,P}}\right) + \#LoS_{D,P} \quad (4)$$

For an individual ranking we evaluated the performance of participants against each other with a bad point value V_P (the denominators need to be changed):

$$V_P = 2\left(\frac{\emptyset t_{diff_{D,P}}}{\emptyset t_{diff_{D,All\_P}}}\right) + \left(\frac{\#Cmd_{D,P}}{\#Cmd_{D,All\_P}}\right) + \#LoS_{D,P} \quad (5)$$

For ranking purposes, individual bad points for each display of every participant were calculated. Those individual bad points were between 10.9 and 28.9 across all 16 participants. The individual ranking of those ranks over all subjects is shown in figure 13.

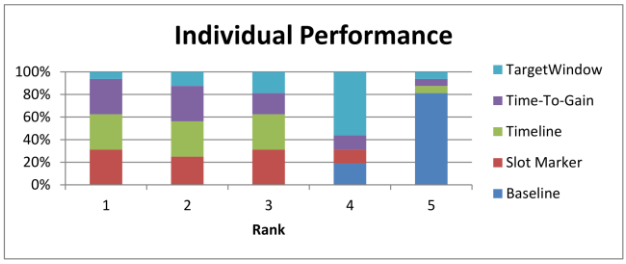


Figure 13. Ranks of displays per participant

F. Performance Change over Time

For analysis purposes, we divided the sequence of aircraft into two halves (roughly five minutes each) to investigate the learning curve of study subjects. Figure 14 shows the average percentage of aircraft that arrived at the merge point within ±2 and ±5 seconds of their scheduled time.

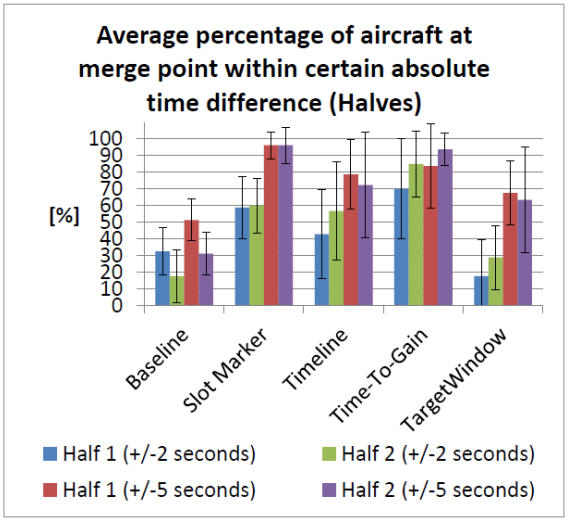


Figure 14. Average percentage of aircraft meeting target times within certain intervals in display simulation run halves (with standard deviations)

Figure 15 reveals the number of speed commands split in two halves.

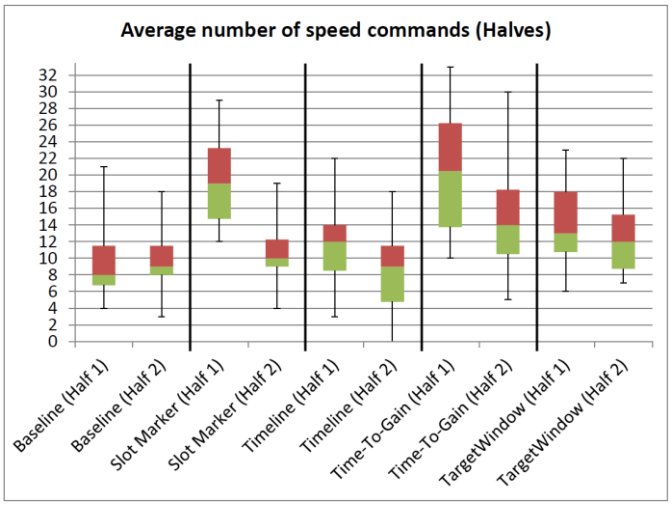


Figure 15. Average number of speed commands during display simulation run halves

V. Discussion of Results

The Slot Marker display always achieved the highest value and statistical significance in all items of subjective ratings (figures 7, 8, and 9). Rank two and three cannot be distinguished very well between Time-To-Gain and Timeline display. TargetWindow display has rank four due to the lower subjective ratings on items 1, 4, and 5. Baseline display obviously received the worst support and even had statistically significant negative responses to some questionnaire items (figure 9).

A. Performance Metrics

The average percentage of aircraft that arrived at the merge point within ±2 seconds was best in the Time-To-Gain display (figure 10). Baseline and TargetWindow had the worst results in this category. Expanding the delay time from 2 to 5 seconds, Slot Marker display was best. In the Baseline display run, the task of meeting target times was objectively very hard for the subjects to fulfill, with an average percentage of aircraft that arrived within ±5 seconds of below 50%. The median of almost 10 seconds time difference again shows bad performance in the Baseline display run (figure 11). The other display runs were much better, with a median between 1 and 5 seconds. Small red and green areas in the plot for the Slot Marker display run and low average time differences show that most participants were able to easily use and perform comparably well with this visual support.

The average number of speed commands per display (figure 12) was opposite to performance in some parts. Participants adjusted speeds very infrequently using the Baseline display, because they did not know in which direction and what amount to adjust. The low number using the Timeline display can be explained by a screen resolution that only allows for recognizing time differences of roughly 3 seconds instead of 1 second, as in the Time-To-Gain condition.

Participants made a lot of changes if they had direct and quick feedback about the delay in seconds and tried to get rid of every second of delay.

The number of losses of separation was much higher using the Baseline display compared to the other four. It was hard for the participants to supervise spacing via target times and separate aircraft at the same time.

The other displays appeared more intuitive to use with less separation due to meeting of already “separated” target times between aircraft. Evaluating the participant individual ranking of displays, Baseline was worst (figure 13). TargetWindow display was second worst for the majority of participants. The ranking was equivalent between Time-To-Gain, Timeline, and Slot Marker displays.

B. Performance Change over Time

Some effects of participants’ improvements during simulation runs can be seen in figures 14 and 15. The percentage of timely accurate aircraft decreased in the course of Baseline display run (figure 14). Subjects were not able to find any better strategy to get rid of larger time differences in the second half of the runs.

The Slot Marker display seems to be quite intuitive from the beginning. The good results already existed in the first half and did not improve during the second half. In contrast to this, the Time-To-Gain display showed better results in the second half. Furthermore, the number of commands increased during the simulation run halves regarding the Baseline display (figure 15). Timeline and TargetWindow showed slightly less, and Slot Marker and Time-To-Gain showed much fewer ATC speed commands in the second part of the simulation run.

So, after a certain learning time, giving fewer speed commands also resulted in less target time deviation of aircraft. Overall, the performance improvement of participants was quite large and should be analyzed in longer simulation runs or with more intensive training. A practice scenario for each visual aid or a switching on/off function during the training run would have been better for this reason.

C. General Discussion

Combining subjectively and objectively measured results, the Slot Marker display worked best for most of the participants, having a direct feedback with aircraft inside or outside a “hole”. Once the aircraft was in the circle and did not overshoot to the border, the task was solved.

Some interaction and screen representation factors may have influenced the results, which can be seen in the participants’ comments. In general, a resizable screen, bigger fonts and larger visual aids (e.g., Slots) were preferred by some of the study participants. Some subjects suddenly realized wrong speeds of their controlled aircraft very close to the merge point. Therefore, they learned that a very early reaction is necessary in case of great differences between scheduled and actual projected target time. The centerline aircraft sometimes were used for pacing. Nevertheless, (especially in the Baseline display run that was strongly disliked by most of the participants) the decision to accelerate or decelerate was made only 30 seconds before reaching the merge point.

An early reaction of participants was possible with the Timeline that also showed aircraft not yet on screen. The Timeline was perceived as a good strategic support for metering. However, the resolution on the screen was too small and steps of seconds respectively ten seconds to compare between scheduled and estimated time of arrival were hardly recognized. One subject suggested providing visual feedback like highlighting if two aircraft label lines on the left and right side of the timeline match horizontally.

Participants reported that the difficulty with the Time-To-Gain/Lose display was keeping in mind whether positive or negative numbers indicate being early or late. The value "minus 5" in this display meant to lose 5 seconds. But some participants interpreted it as being 5 seconds late and therefore gave a speed command in the wrong "direction". Furthermore, the red values could be mixed up with an emergency indication.

The TargetWindow display showed a slower learning curve than the other displays. The trigonometry relations were difficult and it required much "last minute tweaking" of the participants to hit the target positions.

The Slot Marker display was not perceived as clear-sighted as other visual aids. Some participants wondered if they could figure out the speed of the moving circles. But this was not the correct tactic as the Slot Marker changed "speeds" with the changing speeds of aircraft. However, many subjects liked this display and said it was the most intuitive and easiest to use. Another participant suggested a hybrid solution of Timeline, Slot Marker, and TargetWindow.

The participant who was an air traffic controller really liked the TargetWindow display. He stated that he has to look at the centerline for spacing anyway, so the TargetWindow offered a good representation of target position suggestions. He said that the Timeline display drew his attention away from the radar screen, which might cause serious situations in real life. Although only one of the 16 participants was an experienced controller, non-expert results have also shown comparable results in the past [36]. However, this is not generalizable as expert and novice perception and abilities can also differ a lot [37]. Nevertheless, the results of this study are a good basis for setting up future studies on time-based ATM support functionalities.

VI. Conclusion and Outlook

The study compared different air traffic controller display aids. With the described configuration and given participants, the best support to meet target times at significant waypoints was offered by the Slot Marker display. The results of the Baseline display were the worst of all five displays. Conditions with more precise visual support (Slot Marker, Time-To-Gain/Lose) resulted in better accuracy. But this also caused more speed adaptations.

Very similar results were found in a study comparing five different predictive displays with 50 participants for a process control task [38]. The display with least resulting alarms (negative performance/accuracy) had the greatest number of adjusting commands and vice versa.

Hence, for the ATC task, the best trade-off support between time accuracy and costly speed adaptation of aircraft has to be found due to demanded requirements. However, the comparison of visual aids for meeting target times at significant waypoints illustrated the need of feasible controller support functionalities to fulfill future air traffic control requirements

Direct application of these results to a real air controller display is difficult, as the complexity of all evaluated displays was much less than real air traffic control displays. Study participants may have used slightly different approaches and ways of working than controllers. Nevertheless, the results showed a trend of which display technique was preferable to others. Supporting controllers with any visual aid compared to the Baseline display should be recommended.

Making the stand-alone simulation software an online accessible application for controllers could enlarge the amount of data and feedback. Planned improvements for the simulated environment include changes in the interaction and the task description. Influencing aircraft's speed directly at the label and individual adjustable sizes could help. In addition, the number of possible speed commands per display simulation run could be reduced to one. This would be closer to reality and force the participants to think more intensively about their decision, but would require more training time. More generalizable results would be generated by running in a realistic controller environment using conventional simulation displays with integrated visual aids and pseudo-pilots.

For future developments, it is important to gather iterative feedback during a rapid prototyping implementation process [39], [40]. Feedback should consist of subjective opinions and objectively measured parameters. This can help to integrate benefits of different visual aids into a single solution and get rid of individual drawbacks of other support functionalities.

Acknowledgment

The work was conducted during the FAA-DLR researcher exchange program in 2015. We also like to thank our reviewers Jon Rein (Federal Aviation Administration (FAA), William J. Hughes Technical Center, Atlantic City, NJ, USA) and Norbert Fürstenau (German Aerospace Center (DLR), Institute of Flight Guidance, Braunschweig, Germany).

References

[1] Federal Aviation Administration, "NextGen - Implementation Plan," Washington, D.C., USA, 2014.

[2] SESAR, "European ATM Master Plan: The Roadmap for Sustainable Air Traffic Management," Brussels, Belgium, 2012.

[3] P. Brooker, "Air Traffic Control Separation Minima: Part 1 - The Current Stasis," in Journal of navigation (Online) 64, pp. 449–465, 2011.

[4] P. Brooker, "Air Traffic Control Separation Minima: Part 2 - Transition to a Trajectory-based System," in Journal of navigation (Online) 64, pp. 673–693, 2011.

[5] G. McDonald and J. Bronsvoort, "Concept of operations for air traffic management by managing uncertainty through multiple metering points," Proceedings of the Third International Air Transport and Operations Symposium 2012, pp. 217–230, 2012.

[6] A. Neal, J. Flach, M. Mooij, S. Lehmann, S. Stankovic, and S. Hasenbosch, "Envisaging the Future Air Traffic Management System," in The International Journal of Aviation Psychology, Volume 21, No. 1, pp. 16–34, 2011.

[7] U. Metzger and R. Parasuraman, “The Role of the Air Traffic Controller in Future Air Traffic Management: An Empirical Study of Active Control versus Passive Monitoring,” in Human Factors: The Journal of the Human Factors and Ergonomics Society 43, No. 4, 2011.

[8] M.A. Nealley and V.J. Gawron, “The Effect of Fatigue on Air Traffic Controllers,” in The International Journal of Aviation Psychology, Volume 25, No. 1, pp. 14-47, 2015.

[9] T.J. Davis, H. Erzberger, and H. Bergeron, “Design of a Final Approach Spacing Tool for TRACON Air Traffic Control,” NASA Technical Memorandum 102229, NASA Ames Research Center, Moffett Field, CA, USA, 1989.

[10] M.C. Picardi, “Controller-human interface design for the Final Approach Spacing Tool,” Proceedings of the 6th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design and Evaluation of Man-Machine Systems, pp. 27–29, 1995.

[11] T. Prevot, P. Lee, T. Callantine, N. Smith, and E. Palmer, “Trajectory-oriented time-based arrival operations: Results and recommendations,” San Jose State University, NASA Ames Research Center, Budapest, Hungary. Fourth USA/Europe Air Traffic Management Research and Development Seminar (ATM2003), 2003.

[12] V. Kapp and M. Hripane, “Improving TMA sequencing process: Innovative integration of AMAN constraints in controllers environment,” Digital Avionics Systems, 2008. DASC, 27th Conference, pp. 3.D.1–1–3.D.1–9, IEEE, 2008.

[13] A.D. Mundra, ”Display aid for air traffic controllers,” US Patent 4,890,232, Dec. 26, 1989.

[14] T.A. Becher, D.R., Barker, and A.P. Smith, “Methods for maintaining benefits for merging aircraft on terminal RNAV routes,” Digital Avionics Systems Conference, 2004. DASC, 23rd Conference, Volume 1, pp. 2.E.1–1–2.E.1–13, IEEE, 2004.

[15] C. Beers, “SOURDINE II: D6.6 Concept of operation for Schiphol airport simulations,” Consortium of NLR, AENA, Airbus F, EUROCONTROL, ISDEFE, INECO, and SICTA, 2005.

[16] A.P. Smith and T.A. Becher, “A Study of SPACR Ghost Dynamics applied to RNAV Routes in the Terminal Area,” Digital Avionics Systems, 2005. DASC, 24th Conference, pp. 2.D.2–1–2.D.2–11, IEEE, 2005.

[17] P. MacWilliams, A.P. Smith, and T.A. Becher, “RNP RNAV Arrival Route Coordination,” Digital Avionics Systems, 2006. DASC, 25th Conference, pp. 2C1–1–2C1–11, IEEE, 2006.

[18] J. Shepley, “Near-Term Terminal Area Automation for Arrival Coordination,” The MITRE Corporation, Napa, CA, USA. Eighth USA/Europe Air Traffic Management Research and Development Seminar (ATM2009), 2009.

[19] S. Atkins and B. Capozzi, “Relative Position Indicator for merging mixed RNAV and vectored arrival traffic,” Digital Avionics Systems, 2009. DASC, 28th Conference, pp. 2A4–1–2A4–13, IEEE, 2009.

[20] B. Parke, N. Bienert, E. Chevalley, F. Omar, N. Buckley, C. Brasil, H.-S. Yoo, A., Borade, C. Gabriel, P. Lee, J. Homola, and N. Smith, “Exploring management of arrival spacing using route extensions with terminal spacing tools,” San Jose State University, NASA Ames Research Center. Digital Avionics Systems, 2015. DASC, 34th Conference, pp. 3E1–1–3E1–12, IEEE, 2015.

[21] A. Al Gingihy, D. Murray, and S. Ataya, “Terminal decision support tool,” Federal Aviation Administration, 2013.

[22] K. Burnett, G. Scully, D. Davis, J. Krause, K. Cooper, R. Musclow, and P. Beasly, “Visual Aircraft Spacing Tool: Patent Application,” US2009/0287364 A1, United States Patent Application Publication, 2009.

[23] M. Uebbing-Rumke and M.-M. Temme, “Controller Aids for Integrating Negotiated Continuous Descent Approaches into Conventional Landing Traffic,” German Aerospace Center (DLR), Berlin, Germany. Ninth USA/Europe Air Traffic Management Research and Development Seminar (ATM2011), 2011.

[24] H. Oberheid, B. Weber, M.-M. Temme, and A. Kuenz, ”Visual Assistance to Support Late Merging Operations in 4D Trajectory-Based Arrival Management,” Digital Avionics Systems, 2009. DASC, 28th Conference, pp. 2.C.4–1–2.C.4–11, IEEE, 2009.

[25] O. Ohneiser, “Flight guidance support to integrate conventional equipped aircraft into a time based arrival flow” (original German title: Führungsunterstützung zur Integration konventionell ausgerüsteter Luftfahrzeuge in einen zeitbasierten Anflugstrom). in M. Grandt and S. Schmerwitz, editors, Future visualization systems for vehicle- and process guidance (original German title: Fortschrittliche Anzeigesysteme für die Fahrzeug- und Prozessführung), pp. 175–192, Bonn, Germany. German Society of Aerospace, 2012.

[26] O. Ohneiser, M.-M. Temme, and J. Rataj, “Trawl-Net Technology for Timely Precise Air Traffic Controller Turn-To-Base Commands,” German Aerospace Center (DLR), Lisbon, Portugal. Eleventh USA/Europe Air Traffic Management Research and Development Seminar (ATM2015), 2015.

[27] H. Helmke, R. Hann, M. Uebbing-Rumke, D. Müller, and D. Wittkowski, “Time-Based Arrival Management for Dual Threshold Operation and Continuous Descent Approaches,” German Aerospace Center (DLR) and Deutsche Flugsicherung GmbH, Napa, CA, USA. Eighth USA/Europe Air Traffic Management Research and Development Seminar (ATM2009), 2009.

[28] S.J. Landry, T. Farley, and T. Hoang, “Expanding the use of time-based metering: multi-center traffic management advisor,” NASA Ames Research Center, Baltimore, MD, USA. Sixth USA/Europe Air Traffic Management Research and Development Seminar (ATM2005), 2005.

[29] T.J. Davis, H. Erzberger, S.M. Green, and W. Nedell, ”Design and Evaluation of an Air Traffic Control Final Approach Spacing Tool,” in Journal of Guidance, Control, and Dynamics, 14(4), pp. 848–854, 1991.

[30] L. Credeur, W.R. Capron, G.W. Lohr, D.J. Crawford, D.A. Tang, and W.G. Rodgers Jr., “Final-Approach Spacing Aids (FASA) Evaluation for Terminal-Area, Time-Based Air Traffic Control,” Technical Paper 3399, NASA Langley Research Center, Hampton, VA, USA, 1993.

[31] Z. Chi, “An Adaptive Final Approach Spacing Advisory System: Modeling, Analysis, and Simulation,” MIT Flight Transportation Laboratory, Report R 91-3, 1991.

[32] A.M.P. de Leege, M.M. van Paassen, A.C. in’t Veld, and M. Mulder, “Time–Space Diagram as Controller Support Tool for Closed-Path Continuous-Descent Operations,” in Journal of Aircraft 50, No. 5, pp. 1394–1408, 2013.

[33] A. Vuckovic, P. Sanderson, A. Neal, S. Gaukrodger, and B.L.W. Wong, “Relative Position Vectors: An Alternative Approach to Conflict Detection in Air Traffic Control,” in Human Factors: The Journal of the Human Factors and Ergonomics Society 55, No. 5, pp. 946–964, 2013.

[34] R.H. Mogford, “Mental Models and Situation Awareness in Air Traffic Control,” in The International Journal of Aviation Psychology, Volume 7, No. 4, pp. 331–341, 1997.

[35] R.A. Likert, “Technique for the Measurement of Attitudes,” in Archives of Psychology 22, No. 140, pp. 5–55, 1932.

[36] R. Klomp, M. Mulder, M.M. van Paassen, and M.I. Roerdink, “Re-design of an Inbound Planning Interface for Air Traffic Control,” AIAA Guidance, Navigation, and Control Conference, American Institute of Aeronautics and Astronautics, 2011.

[37] W.G. Chase and H.A. Simon, “Perception in chess,” in Cognitive Psychology, 4, pp. 55–61, 1973.

[38] S. Yin, C.D. Wickens, M. Helander, and J.C. Laberge, “Predictive Displays for a Process-Control Schematic Interface,“ in Human Factors, Vol. 57, No. 1, pp. 110–124, 2015.

[39] H.J. Davison Reynolds, K. Lokhande, M. Kuffner, and S. Yenson, “Human-Systems Integration and Air Traffic Control,” in Lincoln Laboratory Journal 19, No. 1, pp. 34–49, 2012.

[40] C. König, T. Hofmann, and R. Bruder, “Application of the user-centred design process according ISO 9241-210 in air traffic control,” in Work: A Journal of Prevention, Assessment and Rehabilitation 41, No. 1, pp. 167–174, 2012.

Bad Weather Highlighting: Advanced Visualization of Severe Weather and Support in Air Traffic Control Displays

Oliver Ohneiser, Matthias Kleinert, Kathleen Muth,
Olga Gluchshenko, Heiko Ehr, Niklas Groß, Marco-Michael Temme
Institute of Flight Guidance, Department Controller Assistance, German Aerospace Center (DLR)
Lilienthalplatz 7, 38108 Braunschweig, Germany, Oliver.Ohneiser@DLR.de

***Abstract*—Adverse weather conditions can have major impacts on air traffic and its control (ATC) in terms of safety and capacity. However, today's ATC situation data displays hardly contain any weather information. Controllers only manually react on external forecasts or status reports to re-plan the aircraft flight paths. The project "Meteorology for Air Traffic Management" (MET4ATM) uses nowcast and forecast weather data for aircraft re-routing calculations and visualization of weather in the radar display. If our Arrival Manager (AMAN) detects that a planned four-dimensional aircraft trajectory is affected by severe weather, it will consider the respective weather polygon, severity, moving direction, and extension for an aircraft re-routing via a detour point. However, the weather situational awareness of the controller would still not be given without further information on the radar display. Therefore, we describe advanced visualization techniques, i.e. a morphing algorithm, to let nowcast polygons become forecast polygons over time until the next weather data update appears. Our implemented prototype highlights weather and re-route affected aircraft, presents smoothly moving weather polygons on the radar display, and gives concrete 4D-trajectory based advisories for re-routing taking the complete arrival stream into account. This continuous support will help controllers to optimize high dynamic air traffic flow even if aircraft do not completely follow the automatically generated plan. In this way the whole severe weather approach operations remain supported by the controller assistance systems and therefore within standard processes.**

Keywords—Air Traffic Controller, Severe Weather, Radar Display, Human Machine Interface, 4D-Trajectory, Advisories, MET4ATM

I. INTRODUCTION

Adverse weather conditions can have major impacts on air traffic control (ATC) in terms of safety and capacity [1]. Thus, it is of utmost importance to support both air traffic stakeholders – on board and on ground. Severe weather is often limited to regional areas, but this might have significant impact on air traffic though.

On more than 9% of all days in the first half of year 2018, Munich airport was affected by thunderstorms around or in the near Alps [2]. Weather was responsible for 8.3% of all flight delays in Europe at the same time [2]. Around 11% of all European departure flight delays result from adverse weather; even 45% of the delays were caused by weather at the airport Frankfurt/Main in 2010 [3].

The experiences of last decades show that the influence of weather on the air traffic system cannot be controlled completely. However, it is possible to proactively optimize the traffic control using appropriate support systems with correspondingly reliable and accurate weather forecasts.

However, today's controller assistance tools hardly integrate any weather information [4]. This paper introduces an arrival manager (AMAN) with enhancements to take severe weather areas in the vicinity of airports for the inbound air traffic organization into account. The AMAN 4-Dimensional Cooperative Arrival Manager (4D-CARMA) generates 4D-trajectories for all arrivals of an airport based on radar and flight plan data. It integrates all trajectories into the inbound traffic stream of one or more runways. Re-routed trajectories, derived guidance advisories, and some weather data itself can then be presented on the controller's human machine interface (HMI). Primarily, this paper handles different visualization techniques to support controllers with accurate current and forecast weather information as well as to give hints about possible implications.

Section II describes related work regarding weather awareness and re-routing in air traffic as well as existing weather displays. This section also outlines the MET4ATM project trying to overcome some drawbacks with respect to weather information for air traffic controllers and the used data structure. Section III explains different concepts and designs to visualize weather data in an air traffic controller (ATCO) display. Section IV demonstrates indicators for ATCOs to detect weather affected aircraft and the integration of calculated 4D-trajectories in this display to avoid severe weather. Section V shows the implemented ATCO radar display with the integrated weather visualization and severe weather avoidance indications. This is followed by a summary and an outlook in Section VI.

II. RELATED WORK AND MET4ATM PROJECT

Dealing with weather in ATC has multiple dimensions such as accurate weather data, on-ground- and on-board-systems, visualization of weather, determination of weather effects, re-routing of aircraft, and monitoring of weather affected air traffic. Aspects of this list have been considered and are outlined in the following.

978-1-7281-0649-6/19/$31.00 ©2019 IEEE

A. Weather Awareness

Weather awareness is not only an important topic for the ground side of air traffic. It is also necessary to understand factors for pilots' decisions and their situation awareness when avoiding severe weather conditions in general [5][6][7] or even with support functionalities to highlight important weather characteristics [8][9]. Ahlstrom points out that "research is lacking on the weather information needs" for Terminal Maneuvering Area (TMA) controllers [10].

Endsley performed a study on weather impact awareness with respect to the next five minutes including 20 ATCOs [11]. In almost 40% of the used scenarios, the ATCOs were not fully able to identify the weather effects [11].

B. Severe Weather Assumptions

Cumulus Nimbus cloud formations (CB) represent a challenge for air traffic in particular during the approach phase of flight. They are structures with a three-dimensional expansion, which should be represented mathematically by a three dimensional polyhedron. In dependence of the degree of latitude and the connected altitude of the troposphere, CBs start around some hundred meters above earth's surface and reach heights of eight to thirteen kilometers [12].

For this reason, during approach it is not possible for aircraft to fly under or over thunderstorms to avoid heavy wind shears, lightning, and hail. Instead, they have to fly around the CBs with a sufficient safety distance. For this reason, it is adequate to consider only the maximal two-dimensional expansion of the cloud formations and represent them as a mathematical polygon for trajectory calculation as lateral dodging is the only way to avoid the hazards of a thunderstorm. In doing so, memory requirements as well as computational calculation time and effort can be reduced.

The weather module of flexiGuide computes time-dependent volumes (4D-polyhedron) marking extreme weather in the vicinity of airports and in extended TMAs to avoid dangerous passings [13]. Their mathematical extension stretches from earth's surface to a height of twelve kilometers.

Unfortunately, with the today available measurement and tracking techniques, it is not possible, to supply thunderstorm areas – classified as severe – with a unique identifier. This is because the most active zones in a thunderstorm front may vary in periods of minutes. Detecting a particularly active area in the south of a 30 miles long front, this area may become less severe after ten minutes, when coincidently another area more in the north starts to develop strong sheer winds and lightning. Sometimes, two formally separated thunderstorms merge and form a front afterwards. Other long broad weather fronts lose their energy and degrade in some smaller extreme weather areas. With these drawbacks in the forecast, it is a real challenge for displaying weather movement on a controller's HMI.

C. Re-Routing Techniques

A simplified detour minimization along aircraft trajectories with on-board calculation has been analyzed in [14]. Furthermore, a two-dimensional rerouting of aircraft around severe weather was described without considering effects on trajectory re-routing conflicts [15].

Different algorithms have been compared for weather re-routing of arrival traffic against actual avoidance paths resulting in a potential for increased capacity [16][17]. Another route re-planning method to avoid convective en-route weather as well as prohibited, restricted, and danger area based on a cellular automat was investigated in [18]. Dynamic adverse weather re-routing in ATC approach can also be faced with a more strategic stochastic model in contrast to static or no re-routing [19].

D. Weather Displays

Many different weather displays exist for air traffic environments. The ROMATSA (Romanian air navigation service provider) ATC radar display is able to overlay the reflectivity of radar areas in many different colors. However, the view tends to over-cluttering.

The FAA (Federal Aviation Administration) reports about their Integrated Terminal Weather System presenting information about storm cells also with different colors on a map [20]. The FAA AIRWOLF system can present weather conflict alerts [21] and greyed weather hazard areas on the radar display [22]. The FAA storm motion tool also includes extrapolated storm cell positions via dotted lines in the radar display [23]. Storm cell now- and forecasts are categorized and visualized with colors on a radar image in [24].

Different shading options to display contours of storm cells are presented in [25]. A wind display integrated in the aircraft radar display and a three-dimensional layer-cake thunderstorm model deliver a huge amount of information that seems to not properly support active controlling of air traffic [26]. Time slots of weather effects on the air traffic flow can be shown on a TMA timeline [27].

In recent years, a cooperation of the German air traffic service provider Deutsche Flugsicherung GmbH (DFS) and the German Meteorological Service (DWD) has developed a new weather display system called "Graphical Meteorological Display System" (GMET) for air traffic controllers [36] and has taken it over for testing in Munich. It can display current data on lightning and precipitation in a separate controller display, allowing both ATCOs and supervisors to individually adjust the amount of weather information in the radar display. If desired, the weather displays also show the situation of the immediate past and are continuously updated.

E. The Project MET4ATM

The project "Meteorology for Air Traffic Management" (MET4ATM) aims to deliver air traffic control support systems for planning and guiding all traffic approaching an airport, taking dynamic storm cells into account [28].

Consistent meteorological data about current and forecasted weather – providing constant coverage with regard to both time and space – are a prerequisite for accurate decision support functionalities. The meteorological data is translated into 4D-polyhedrons respectively -polygons in order to calculate 4D-trajectories for safe arrival routes around severe weather.

The ATCO support consists of (1) visualization of current and forecasted categorized weather phenomena integrated into ATCO's situation data display, (2) calculation of four-dimensional trajectories to avoid severe weather such as thunderstorms and hail, (3) concrete adequately presented guidance advisories how and when to efficiently reroute the complete relevant air traffic, and (4) sequencing information for the complete arrival stream. This paper concentrates on the visualization aspects of (1) and (3) and will just give a short overview of calculation methods relevant for (2), (3) and (4).

F. Data Structure for Current and Forecast Weather

Actually, there exists no standardized data format for three dimensional weather nowcast and forecast information to use them in air traffic management environment.

In MET4ATM, the project-partner Leonardo defined an Extended Markup Language (XML) structure, which contains all relevant measurement and forecast data, independent of the extension of the study area, number of cloud formations, as well as count and time steps of the forecasts. The meteorological XML-tags contain data about timestamps for measurement and valid times, a severity classification, position, speed, and direction of draught of the geometric center of gravity, all with an additional tag for the employed units.

The CBs themselves are defined through a different number of layers, mathematical outlined with the three dimensional points in space, formatting a closed polyline of a polygon. All polygons of one CB together describe the three dimensional appearance of the severe weather area.

The given weather data is converted into the DLR AMAN database format. Two tables for the meteorological polygons and the related polygon points contain a polygon id, severity, speed, direction, and forecast minutes with validity of the polygon as well as the latitude and longitude of all polygon points and the centroid.

III. Weather Information Visualization Techniques

There are a number of different possibilities to display reasonable weather data on an air traffic controller display. Furthermore, this can be the main situation data display with all radar targets or an auxiliary screen.

First, ATCOs are more likely interested in severe weather areas including phenomena such as thunderstorm and hail as they have the greatest influence on aircraft. In this context moving direction, extension, and severity need to be considered.

Secondly, weather visualization is important with respect to current data and forecasted weather for a situation analysis. The forecasted weather again can be of interest with different time horizons, e.g. looking 5, 10, 15, 30 or more minutes into the future. Hence, also the moving behavior of weather cells over time given by a continuously forecast visualization (replay tool) can be useful.

Thirdly, visualization characteristics such as colors and shadings are important to not hide relevant traffic information. Fourthly, weather can be displayed in a two-, three- or four-dimensional view depending on the controller working position and ATCO needs.

Fifthly, ATCOs should have the ability to switch the weather visualization completely on and off respectively to adjust various settings such as reflectivity in dBZ, safety buffers or forecast times and overlay characteristics of weather data.

The following sections present visualization techniques for static, dynamic, and morphing two-dimensional weather cells integrated into a radar display. This is based on the availability of respective data, i.e. the current weather and forecasted weather in 5-minutes-steps up to one hour look-ahead time. This data – including current and forecast weather – is updated every five minutes.

A. Static Weather Cells

The easiest way to display current weather is to use the latitude and longitude of all nowcast polygon points and keep the shown polygon static until the next data update in five minutes. The next visualization update will be generated based on the weather data update.

This visualization is quite reliable as there is no dependency on forecast data that might have lower accuracy. However, the ATCO might see a big jump of weather polygon positions after updating and might rely on already outdated information. Furthermore, this visualization technique hardly gives an opportunity to estimate future weather behavior.

B. Weather Cells Movement with Current Vectors

Dynamic movement of weather polygons is an advanced option to visualize future weather behavior. The display appearance at the time of weather data update is the same as for static weather cells.

However, the polygon is moving with the constant speed and direction given in the data. The visualization update rate can e.g. be once per second or every five seconds adapted to the radar update rate. Hence, polygons move continuously without changing their shape.

At the time of the next weather data update there might as well be a small jump of polygons as the future position based on old data is different from the next nowcast. Hence, for minimizing the polygon jumps in case of data update, the forecast data should be used, too.

C. Time Based Weather Cells Transformation from Nowcast to Forecast Data

The forecast data polygon positions are seen as the final polygon appearance at the time of the next data update (in five minutes). The first polygon appearance is still the same as for the static weather cell. However, the polygon now modifies its shape and number of edges like a continuously interpolation between the nowcast and the forecast.

This helps ATCOs to even better anticipate future weather situations. The polygon jumps should be small if the forecast data was good. However, the needed morphing algorithm does not work very simple for all types of theoretical polygons. The

following sections III.D and III.E describe our developed morphing algorithms for realistic weather polygons.

D. Morphing Algorithm Considerations

Since thunderstorm cells should be visualized on a 2D display, the current and forecasted thunderstorm cells are represented by discrete sets of simple polygons. These polygons need identification numbers with a unique correlation between sets.

A (closed) simple polygon Pol is defined by the ordered set of its vertexes. It consists of all line segments consecutively connecting the vertexes, which do not intersect and bound the connected interior area. Polygons with self-intersections as well as polygons with holes are not considered here, since they are not suitable for the representation of thunderstorm cells.

Our goal was to perform a realistic interpolation of transformation of a set of polygons at the current moment in time t into a set of forecasted polygons at time $t+t_0$ over the time period t_0. The interpolation should have small time step (≤ 5 seconds). The solution should be as simple as possible in terms of computational effort.

Since thunderstorm cells can appear, disappear, merge together or split into small parts over time, it was necessary (and sufficient) to develop approaches to perform the following main interpolations of:

1. A transformation of one polygon into another polygon

2. A transformation of one polygon into several polygons and vice versa

3. A transformation of one polygon in simultaneous splitting and merging cases of this polygon.

The formulated problem belongs to the class of 2D polygon morphing problems. The problem is well studied, especially for complex shapes, because of their wide applicability in computer graphics and animation. As a consequence, these solutions imply significant computational effort and/or are developed for some initial objects that have properties which are irrelevant for visualization of thunderstorm cells. The detailed overview of the literature is out of scope of this paper. Here we refer works related to polygon morphing. An overview on the existing methods is given, for instance, in [29].

Approaches for polygon morphing consist of two main steps: *mapping of polygons by some characteristic/feature points* and *specification of interpolation ways/curves*. The last step is more complicated, because it is often desired to retain some characteristic features of the considered objects during interpolation. One of the main goals defining interpolation ways is to avoid local self-intersections of the polygon boundaries.

There are many heuristics presented in the literature. Guaranteed intersection-free polygon morphing described in [30] relies on analytical basis. However, the approach uses significant number of interior points and exterior Steiner vertices that raises its complexity. Morphing simple polygons with the same number of sites that are corresponding parallel are explored in [31]. Usually, morphing algorithms require user-assisted correspondence of the morphing objects.

[32] introduces an intuitive polygon morphing, however, the source and destination polygons must spatially overlap. [33] deals with the application of 2D polygonal morphing techniques to create spatiotemporal data representations of moving objects continuously over time. The movement of icebergs in the Antarctic seas is used as case study and the data sources are sequences of satellite images capturing the position and shape of the icebergs at different dates. The authors apply perceptually-based approach proposed in [34] that determines the so-called feature points of the morphing objects. The main challenges are determining the feature points and the correspondences between feature points. Since this work investigates visualization of moving gaseous objects, rotations and similarities have an insignificant role compare to solid objects [33].

Taking into account relative small time steps between consequent sets of thunderstorm cells compared to their movement and transformation, linear interpolation ways appear to be the best choice for our approximation requirements.

E. Morphing Algorithm Functionality

Let us consider the first transformation mentioned above of one polygon into another. This type of transformation provides the basis for two remaining mentioned types of transformations and, as follows, for the morphing of the set of polygons at the time t into the set of polygons at the time $t + t_0$ over the time period t_0.

a) Decomposition of Polygon Boundaries

To perform morphing of one polygon into another, polygon boundaries are decomposed into 4 corresponding pairs to get a more natural interpolation of the transformation that avoids significant turns and rotations of the boundary during interpolation. For this purpose 4 sets of vertexes with minimal and maximal abscissa and with minimal and maximal ordinate have to be found for the current and for the forecasted polygon. Each set can consist of more than one point and some sets can coincide. In the case, the considered set consists of more than one point, one can choose a point from this set, which provides balanced length of the boundary parts. The simplest solution is to take the first point from each set in the vertex index in increase direction. The obtained in such a way corresponding boundary parts marked by an identical color are illustrated in Fig. 1. Here Pol^t and Pol^{t+t_0} are polygons at the time moments t and $t + t_0$, E_i^t, $i = 1, \ldots, 8$ and $E_j^{t+t_0}$, $j = 1, \ldots, 9$ are the corresponding vertexes and E^{*Top}, E^{*Right}, $E^{*Bottom}$, E^{*Left} are subsets of vertexes with minimal or maximal first or second coordinate.

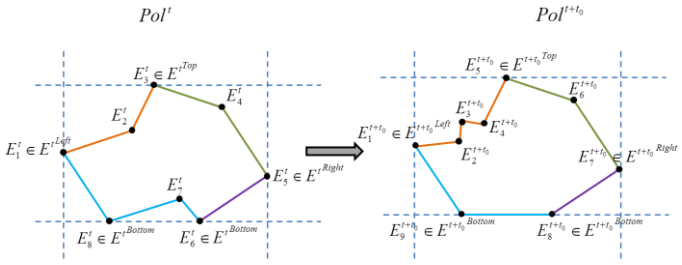


Fig. 1. Decomposition of polygon boundaries into 4 parts (marked by different colors).

After this decomposition step, morphing of the boundary parts is performed by the linear interpolation of the vertexes in s discrete steps, i.e. the time period t_0 is divided into s parts of the duration t_0/s. For this purpose, the boundary parts are balanced in order to establish one-to-one correspondence between them. This means that the boundary part with the smallest number of vertexes becomes additional points uniformly distributed along the edges so that on each edge the quotient number of additional points plus one additional point on the remainder number of edges in the vertex index increase direction are selected. The described approach is illustrated in Fig. 2. Here the piecewise linear curve A_1A_3 is transformed into the piecewise linear curve B_1B_{10}. The quotient of division 10 by 3 is equal to 3 and the remainder is 1.

Therefore, the edge A_1A_2 becomes 3+1 additional points and on the edge A_2A_3 3 additional equidistant points are selected. They are marked red in Fig. 2.

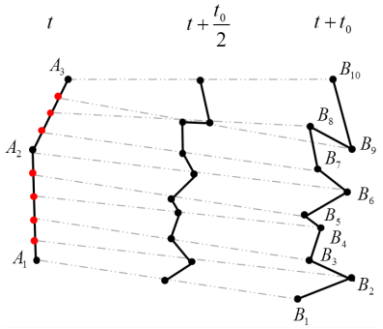


Fig. 2. One-to-one transformation with the help of additional points.

b) Morphing of a Single Polygon

The described approach provides an identical number of "vertexes" on both linear curves to perform one-to-one linear interpolation of the vertexes in s discrete steps. The interpolated curve at the time moment $t+\frac{t_0}{2}$ is illustrated in Fig. 2. As a result, morphing of Pol^t into Pol^{t+t_0} is realized by the interpolation of the transformation of four corresponding boundary parts. The interpolation at the time moment $t+\frac{t_0}{2}$ is illustrated in Fig. 3.

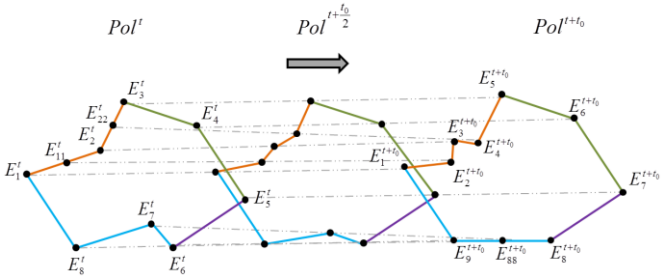


Fig. 3. Morphing of one polygon into another.

The input data, we have, represent current and forecasted locations of thunderstorm cells, i.e. the current situation and a forecasted result of transformation after the time period t_0 are available only. Although, here is a correlation of identification numbers between two consecutive (in time) sets of polygons, there is no information in the input data in a merging or splitting case, which part of a pre-image corresponds to a transformed polygon and vice versa. There is no way to exactly follow the decomposition process for available input data.

In the case when the forecasted decomposed polygons are mapped to some parts of the current polygon, the free corridors appearing in the visualization do not reflect the actual situation and can deviate from it significantly. Therefore, from the safety point of view aircraft should not be directed through these corridors. Additionally, between the forecasts are short up to 10-minutes time periods. Based on all said above, mapping of polygons without decomposition of the current polygon in the splitting case and without decomposition of forecasted polygon in the merging case can be taken as a reasonable solution.

c) Morphing of Polygons with Splitting and Merging

Fig. 4 illustrates the mapping of polygons in the splitting case to perform the interpolation of their transformation. The polygon $Pol^{t|1}$ at the current time t decomposes over the time period t_0 in 4 parts $Pol^{t+t_0|1^{(t|1)}}$,…, $Pol^{t+t_0|4^{(t|1)}}$.

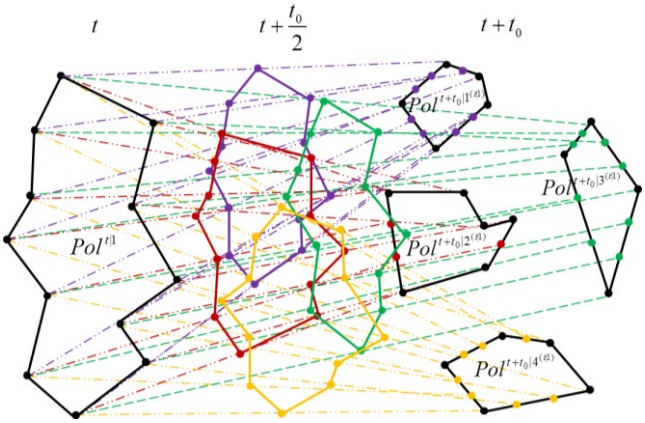


Fig. 4. Visualization of interpolation without decomposition of the current polygon.

The current polygon is mapped to all forecasted polygons and morphing of one polygon into another as described above is performed. Fig. 4 shows interpolation lines, additional points and the constellation of the polygons after the time period $\frac{t_0}{2}$ colored corresponding to the mapping of the polygons. Hereby an approach to transform one polygon into several polygons and vice versa is developed.

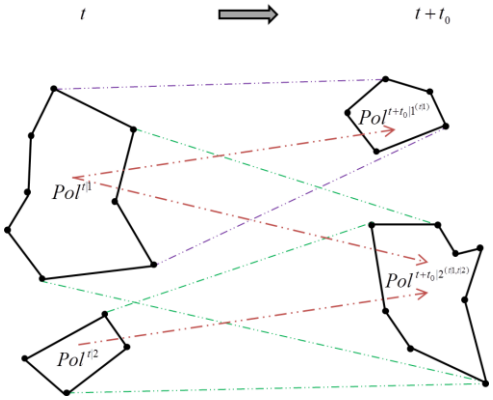


Fig. 5. Interpolation in simultaneous splitting and merging cases.

The mapping of the whole polygons can be also used to perform a transformation of one polygon in the case of simultaneous splitting and merging, i.e. for the case, when a part of one polygon splits from the polygon and another part of the polygon merges with some other polygon simultaneously. Mapping without decomposition is illustrated in Fig. 5.

Hence, the presented morphing approach is applicable for the general interpolation of transformation between two sets of closed polygons representing thunderstorm cells over some (short) period of time.

IV. Severe Weather Indication and Avoidance Trajectories

The displaying of extensive information as expanded bad weather areas on a flight radar HMI without cluttering and covering essential aircraft label are a challenge when trying to raise the controller's weather situation awareness. The 4D-trajectory approach calculation with its specific constraints due to speed and altitude reduction phases away from the standard arrival routes requests for new concepts as well.

A. Prototypic Situation Data Display RadarVision

The prototypic DLR air traffic situation data display RadarVision is used for visualization. RadarVision consist of a radar display (Fig. 6) and a timeline as well as menu buttons. The radar display presents all relevant waypoints, runways, aircraft symbols and labels, trajectories as well as some relevant weather data.

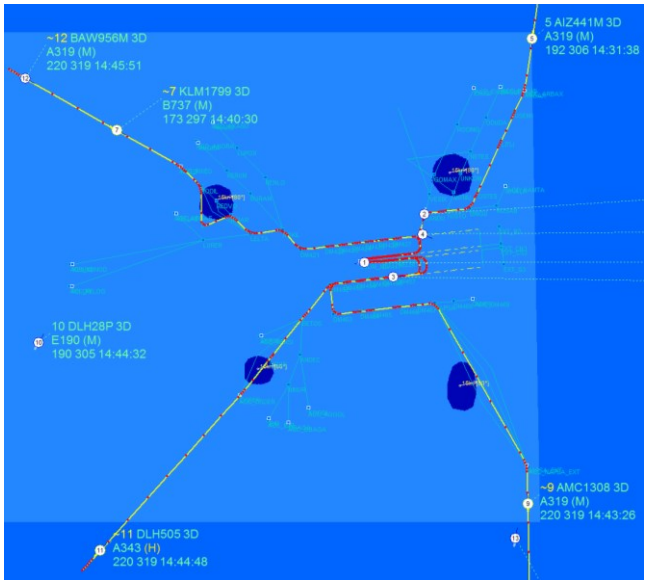


Fig. 6. Visualization of aircraft with their 4D-trajectories and radar label information avoiding thunderstorm cells when approaching Munich airport.

B. Indicators for Aircraft Affected by Severe Weather

A basic level of bad weather support is given to the ATCO by indicating aircraft whose trajectories are affected by severe weather. Therefore, RadarVision displays a yellow tilde symbol (respectively Greek omega symbol in a controller view not shown) at the beginning of the first label line.

Six aircraft of Fig. 6 do have this symbol to draw the ATCO's attention to those flights. The symbol appears as soon as the AMAN calculation determines the necessity for a re-routing. It will disappear as soon as the severe weather region has been passed.

C. 4D-Trajectory Calculation

The trajectory calculation starts with a determination of the shortest route an aircraft can take to the runway according to the standard terminal arrival routes (STAR) without taking severe weather conditions into account. Afterwards a time based trajectory is calculated, which gives information about where an aircraft will probably be at a certain point in time. With these information relevant weather polygons are selected and checked for possible conflicts with the aircraft trajectories. Relevant weather polygons can be current and forecasted weather cells.

To determine possible conflicts, a trajectory is split into smaller segments, which cover a short period of time. A conflict occurs when the time period and positions that are covered by a trajectory segment partly overlap with an area that is covered by a relevant weather polygon at a certain point in time. For any detected conflict, a detour has to be created, which is always based on three points:

- **Start and end of detour**: Two existing waypoints which are located outside of the conflicting weather polygon. Usually the start and end of the selected trajectory segment. If one of the points is inside the conflicting weather polygon, the closest waypoint before respectively after the weather polygon is selected.
- **Detour point**: A new calculated waypoint that connects the start and end waypoint of the detour and resolves the detected conflict with the weather polygon.

Calculation of the detour point is illustrated in Fig. 7. To find this point three subsidiary lines are introduced: the line through the start (P^1) and end point (P^2) of the detour and two orthogonal to this line at the points P^1 and P^2 (Fig. 7). These lines divide the plane into six parts shown in Fig. 7. The parts are labeled "left" and "right" with respect to the flight direction.

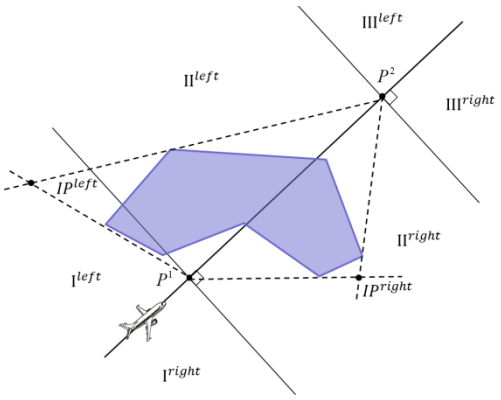


Fig. 7. Calculation of the detour points.

There is a maximum of two efficient possible detours: one left and one right from the polygon with respect to the flight direction. To find the detour points, an intersection point of the

tangent lines through the way points P^1 and P^2 to the left and right parts of the polygon has to be calculated (Fig. 7).

When the intersection point IP^{left} is located in the part II^{left} or, correspondingly, the intersection point IP^{right} is located in the part II^{right}, it is suitable as a detour point. In the case of two suitable detour points, the short distance rule is used to choose one of them. If there are no suitable detour points, a new start or end point for the detour calculation is selected until at least one suitable detour point is found.

D. 4D-Trajectory Visualization

The RadarVision display user can apply a mouse-over function on the aircraft radar symbol to show re-planned routes that comprise detour points as explained above (Fig. 8).

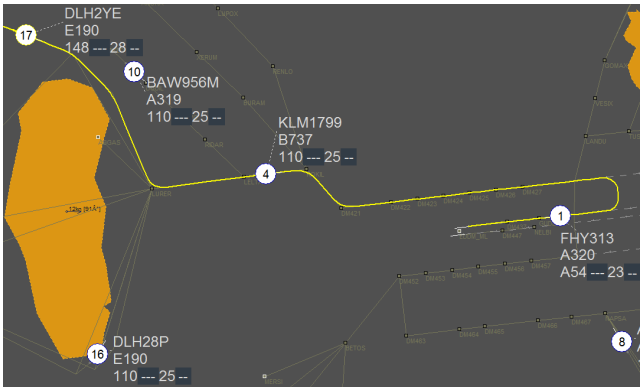


Fig. 8. Visualization of aircraft (circles) with arrival sequence numbers, orange weather polygons, and a yellow mouse-over displayed re-routing trajectory to one of the parallel runways.

It is also possible to switch the 4D-trajectory visualization completely off or on. A further software menu on the aircraft symbol offers the altitude and ground speed profile of the complete flight (Fig. 9).

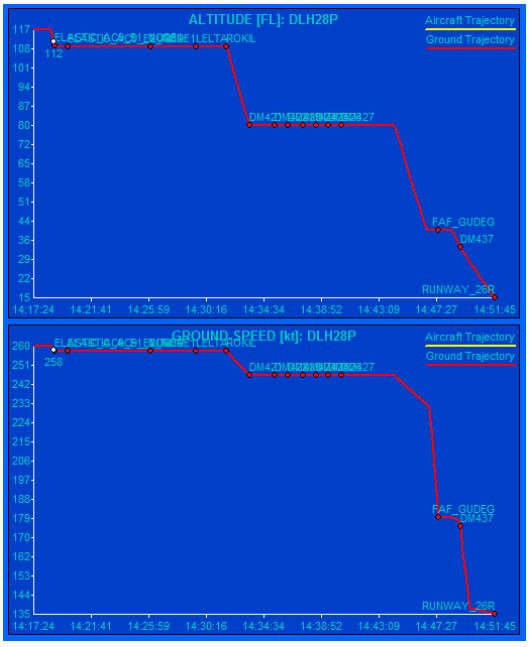


Fig. 9. Altitude and speed profile of a weather affected aircraft. The trajectroy was calculated by the DLR AMAN 4D-CARMA.

This shows that latitude, longitude, altitude, and time are still supported in the arrival stream planning for all aircraft independently of being weather and re-routing affected or not. Therefore, sequence numbers for the touchdown of arriving aircraft can be calculated and shown in the aircraft radar symbol (Fig. 8) or label (Fig. 6).

E. Re-Routing Advisory

The calculated detour point to avoid flying through a weather polygon is left or right of the originally planned trajectory as described above. Thus, the ATCO needs to give re-routing clearances to the respective aircraft pilot. The re-routing consists of a track towards the detour point and then back to the original route again. Our AMAN 4D-CARMA generates two controller command suggestions (advisories) to support the ATCO with re-routing (Fig. 10 and Fig. 11).

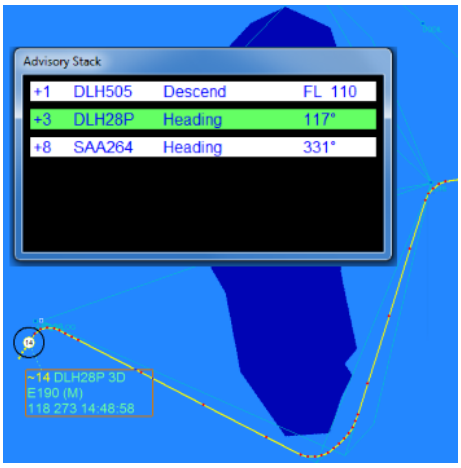


Fig. 10. Heading advisory for DLH28P to reach the detour point.

Normally, DLH28P would fly north-eastwards in Fig. 10. However, the dark blue weather polygon shall be avoided. Hence, there is a re-planning of the trajectory flying a southern detour.

The respective trajectory changes are derived along the calculated 4D-trajectory from the current aircraft position to touchdown. The AMAN's advisory generator module detects a curve by multiple heading changes in the planned trajectory. If this curve was integrated due to weather restrictions, a heading command with the target heading at the planned time of leaving the original route will be generated (e.g. "HEADING 117°" in Fig. 10).

Following the planning points on the trajectory, there will be two further new curves directly at the detour point (center area of bottom of Fig. 10) and the re-joining waypoint on the original route (center area of right side of Fig. 10). The algorithm determines the planned time of reaching the detour point curve and the name respectively position of the re-joining point curve to generate the second command advisory. This way, the ATCO only needs to issue two clearances for re-routing and re-joining the original route. This procedure is also beneficial for simulators in case of automatically following the advisories.

However, it is also technically feasible to replace the DIRECT_TO command by two further HEADING commands. For some trajectories that include long curves near the detour point, there might also be a further heading advisory.

The generated advisories follow the ontology format of ATC commands and are displayed in a similar way as shown in Fig. 10 and Fig. 11 [35]. The first command is a heading command, e.g. DLH28P HEADING 117 RIGHT. The second command indicates the path back to a waypoint on the original route, e.g. DLH123 DIRECT_TO LURER.

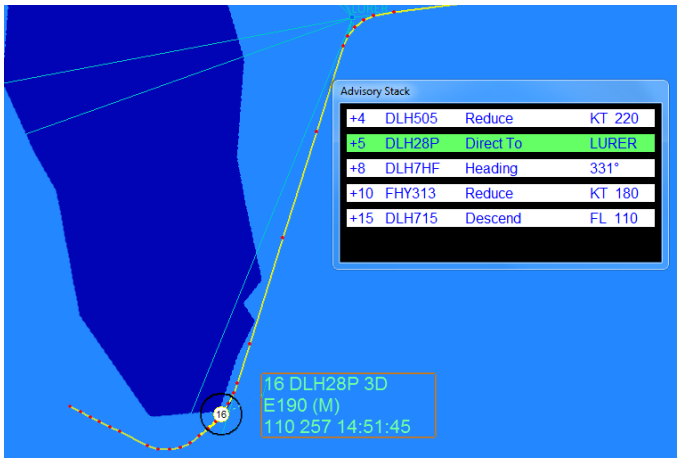


Fig. 11. Direct to waypoint advisory to re-join original route.

After the aircraft passed the detour point (compare Fig. 11), the weather affection symbol in the aircraft label disappears. Furthermore, the arrival sequence number of the aircraft at the beginning of the first label line might change due to different re-routings of aircraft in the complete approach area.

These advisories are shown in a so-called advisory stack that comprises command suggestions for the complete arrival stream. The stack also shows a countdown to issue the listed clearance suggestions with timely accuracy. Furthermore, the ATCO can use a mouse-over functionality. When hovering, the mouse over the advisory, the respective aircraft will be highlighted on the radar screen and vice versa.

V. Simulation Runs with Implemented ATCO Weather Display

A. *Use Case Munich Approach*

The Munich airport is an international airport in Bavaria, Germany. In terms of passenger numbers, it is the second busiest airport in Germany and the seventh busiest in Europe [37]. The airspace was modeled based on Aeronautical Information Publication (AIP). There are three main approach directions within the extended TMA. A parallel runway configuration with runways 26L and 26R was modeled. The scenario simulation runtime is one hour and includes 35 arrival aircraft (7 Heavy, 28 Medium) without departures or overflights for the weather re-routing scenario. The traffic is loosely based on real traffic mix. The convective cell data for the scenario is based on data provided by German Weather Service.

B. *Automatic Real Time Simulation for Severe Weather Avoidance*

The scenario described in section A of this chapter was used in automatic real time simulations to test the avoidance of severe weather conditions. This included the detour calculation as well as the visualization of the detour and severe weather conditions.

For simulation purposes, the radar simulator Arros, developed by DLR, was used. The simulator takes initial radar data for given aircraft into account and creates further radar data based on the trajectories that it is provided with. This meant for the simulations that all aircraft would follow the trajectories with the calculated detours. To visualize the radar data, trajectories, and severe weather conditions the prototypic radar display RadarVision was used. The outputs of the display have already been described in sections IV.A to IV.E.

The first tests with this setup by ATC experts showed that the chosen approach delivers good results for the detour calculation for all severe weather conditions used in the scenario. Also the presented visual output with complex moving polygon structures, information about weather influenced aircraft and 4D-trjaectories with possible detours for every aircraft showed the potential to provide a benefit to air traffic controllers.

VI. Summary and Outlook

In the project “Meteorology for Air Traffic Management” (MET4ATM), the DLR developed a controller support algorithm to generate 4D-trajectories around bad weather areas for individual aircraft in approach. After detecting severe weather on preplanned standard arrival routes, the arrival manager calculates deviation routes, integrates the aircraft into the current arrival stream to one or more finals of the airport, calculates target times for arbitrary waypoints and the runways, and supports the ATCO with advisories for pilots to follow the deviation trajectories around the bad weather areas. Therefore, an arrival manager’s trajectory generator takes current weather measurements and qualified forecasts into account. It also supports ATCOs to guide pilots where to fly around potential dangerous thunderstorms. In addition, the nowcast and forecast weather positions and moving behaviors are used to visualize the most dangerous areas on controller’s radar display.

The AMAN calculates the real time positions and shapes of the cells on the basis of the latest available measurements and forecasts. The HMI is updated with the depiction in five-seconds-steps. This is the same rate as aircraft positions are refreshed and therefore controllers are accustomed to get a new situation picture in this speed to assess the traffic and future weather situations. This gives the controller a feeling for the attitude of severe storm cells and bad weather fronts.

For trajectory calculation as well as for the weather area morphing on the display, the predicted speed, direction, and shape alteration of thunderstorms are taken into account. In this way, controllers are not only dependent on radio feedbacks of the flight crews to get a weather picture from the surrounding of an airport. The controllers can now move from a reactive to a much more active way of guidance and inbound stream organization.

The project MET4ATM focused on technology and software development. The use cases contained different traffic and weather scenarios, which verified the trajectory and target time calculations in medium and high density traffic situations for different flight distances to the test airport Munich in Germany.

First, it was proven that individual 4D-trajectories with typical constraints of approach procedures can be generated to avoid dynamic thunderstorm cells and fronts for the several aircraft. Secondly, aircraft trajectories were integrated into the arrival stream on finals. Thirdly, the validation of guidance suggestions and display aids have to be demonstrated in future human-in-the-loop validations.

For this, different traffic and weather scenarios will be arranged. To verify the benefit and effectiveness, ATCOs will have to guide aircraft around severe weather to an airport surrounding with and without the MET4ATM support tools. Several indicators have to be defined and subsequently achieved to prove the helpfulness of the new controller support system.

Currently, the AMAN algorithms generate a deviation trajectory around severe weather, most often with additional route length. In the next generation of trajectory deviation algorithms and controller support functions, we plan to integrate estimations for controllers and pilots, how long the detour around the weather will last. In a next step, the algorithm will decide, if a deviation especially around extended bad weather fronts makes sense, or whether holdings for a little waiting time are the better solution and later using gaps with reduced severity in the frontal system to continue the flight.

ACKNOWLEDGMENT

The work was conducted during the project MET4ATM funded by the Federal Ministry for Economic Affairs and Energy/LuFo. MET4ATM comprises five project partners: SELEX ES GmbH (belongs to Leonardo Germany GmbH; Coordinator), DLR Institute of Flight Guidance, DLR Air Transportation Systems, Harris Orthogon GmbH, and MeteoSolutions GmbH.

REFERENCES

[1] U. Ahlstrom, "Work domain analysis for air traffic controller weather displays," in: Journal of Safety Research, Volume 36, Issue 2, 2005, pp. 159-169.

[2] Gewerkschaft der Flugsicherung e.V., "Adverse Weather – New procedures help optimise air traffic management over the Alps in adverse weather conditions," der flugleiter, 6, 2018, pp. 49-50.

[3] DFS, "Luftverkehr in Deutschland - Mobilitätsbericht 2010," 60-2290-187/02.11, Langen, Germany, DFS Deutsche Flugsicherung GmbH, 2011.

[4] U. Ahlstrom and F. Friedman-Berg, "Controller Scan-Path Behavior During Severe Weather Avoidance," DOT/FAA/TC-06/07, U.S. Department of Transportation, Federal Aviation Administration (FAA), 2006.

[5] C.B. Tienes, "Important Factors for a Pilot's Decision when Avoiding Severe Weather Conditions," Bachelor Thesis, Rhein-Waal University of Applied Sciences, 2018.

[6] U. Ahlstrom, O. Ohneiser, and E. Caddigan, "Portable Weather Applications for General Aviation Pilots," Human Factors, 58(6), 2016, pp. 864-885.

[7] M.-M. Temme and C. Tienes, "Factors for Pilot's Decision Making Process to Avoid Severe Weather during Enroute and Approach," 2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), 2018.

[8] U. Ahlstrom, E. Caddigan, K. Schulz, O. Ohneiser, R. Bastholm, and M. Dworsky, "The effect of weather state-change notifications on general aviation pilots' behavior, cognitive engagement, and weather situation awareness," DOT/FAA/TC-15/64, FAA William Hughes Technical Center, 2015.

[9] U. Ahlstrom, "Weather display symbology affects pilot behavior and decision-making," in: International Journal of Industrial Ergonomics, 50, 2015, pp. 73-96.

[10] U. Ahlstrom and P. Della Rocco, "TRACON Controller Weather Information Needs: I. Literature Review," DOT/FAA/CT-TN03/18, U.S. Department of Transportation, Federal Aviation Administration (FAA), 2003.

[11] M. Endsley and M.D. Rodgers, "Attention Distribution and Situation Awareness in Air Traffic Control," in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 1996.

[12] N. Mölders and H. Kramm, "Lectures in Meteorology," Springer International Publishing, Switzerland, 2014.

[13] C. Forster and A. Tafferner, "Nowcasting and forecasting thunderstorms for air traffic with an integrated forecast system based on observations and model data," WMO Symposium on Nowcasting, Whistler, B.C., Canada, 2009.

[14] C. Schilke and P. Hecker, "Dynamic Route Optimization Based on Adverse Weather Data," Fourth SESAR Innovation Days, Madrid, 25.-27.11.2014.

[15] T. Hauf, L. Sakiew, and M. Sauer, "Adverse weather diversion model DIVMET," Institut für Meteorologie und Klimatologie, Leibniz Universität Hannover, Hannover, Germany, in: Journal of Aerospace Operations, 2, 2013, pp. 115-133.

[16] J. Krozel, S. Penny, and J. Prete, "Comparison of Algorithms for Synthesizing Weather Avoidance Routes in Transition Airspace," Collection of Technical Papers - AIAA Guidance, Navigation, and Control Conference, 1, 2004.

[17] J. Krozel, T. Weidner, and G. Hunter, "Terminal area guidance incorporating heavy weather," in: AIAA Guidance, Navigation, and Control Conference, 1997, pp. 411-421.

[18] W. Lim and Z. Zhong, "Re-Planning of Flight Routes Avoiding Convective Weather and the 'Three Areas'," in: IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 3, March 2018, pp. 868-877.

[19] A. Mukherjee and M. Hansen, "A dynamic rerouting model for air traffic flow management," in: Transportation Research Part B: Methodological, 43, 1, 2009, pp. 159-171.

[20] J.E. Evans and E.R. Ducot, "The Integrated Terminal Weather System (ITWS)," Lincoln Lab. J. 7, 2, September 1995, pp. 449-474.

[21] U. Ahlstrom, J.T. DiRico, and K. Stephenson, "The AIRWOLF Tool to Support En Route Controller Weather Advisories," DOT/FAA/TC-10/09, U.S. Department of Transportation, Federal Aviation Administration (FAA), 2010.

[22] U. Ahlstrom and E. Jaggard, "Automatic Identification of Risky Weather Objects in Line of Flight (AIRWOLF)," in: Transportation Research Part C: Emerging Technologies, 18(2), 2010, pp. 187-192.

[23] U. Ahlstrom and F. Friedman-Berg, "Using eye movement activity as a correlate of cognitive workload," in: International Journal of Industrial Ergonomics, 36(7), 2006, pp. 623-636.

[24] C. Forster, A. Tafferner, and H.-D. Saffran, "Gewittervorhersage," promet, 39 (39), DWD, 2014, pp. 45-54.

[25] M.-M. Temme, O. Ohneiser, O. Gluchshenko, and A. Lau, "Konzept zur Umgehung von Konvektionszellen im En-Route Segment und im Anflug mithilfe von Lotsenunterstützungssystemen," Internal Report, DLR-IB-FL-BS-2016-335, 2016.

[26] J. Stobie and R. Gillen, "Integrating Convective Weather Forecasts With the Traffic Management Advisor (TMA)," Embry Riddle Aeronautical University, Daytona Beach, FL, Aviation, Range and Aerospace Meteorology Special Symposium on Weather-Air Traffic Management Integration, 2009.

[27] M. Robinson, H.J. Davison Reynolds, and J.E. Evans, “Traffic Management Advisor (TMA) Weather Integration,” Project Report ATC-364 Lincoln Laboratory MIT, prepared for FAA, 2010.

[28] Homepage of MET4ATM project on DLR Braunschweig website: https://www.dlr.de/fl/en/desktopdefault.aspx/tabid-1149/1737_read-47512/, 2018.

[29] M Malkova, “Morphing of geometrical objects in boundary representation: The State of the Art and the Concept of Ph.D. Thesis,” Technical Report No. DCSE/TR-2010-02, April 2010.

[30] C. Gotsman and V. Surazhsky, “Guaranteed intersection-free polygon morphing,” Computers & Graphics 25, 2001, pp. 67-75.

[31] L. Guibas, J. Hershberger, and S. Suri, “Morphing Simple Polygons,” in: Discrete Comput Geom, 24:1, 2000.

[32] M. Malkova, J. Parus, I. Kolingerova, and B. Benes, “An intuitive polygon morphing,” in: The Visual Computer, 26(3), 2009, pp. 205-215.

[33] J. Moreira, P. Dias, and P. Mesquita, “Morphing techniques for creating and representing spatiotemporal data in GIS,” International Environmental Modelling and Software Society (iEMSs), 7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA, D.P. Ames, N.W.T. Quinn, and A.E. Rizzoli (Eds.), 2014.

[34] L. Liu, G. Wang, B. Zhang, B. Guo, and H.-Y. Shum, “Perceptually Based Approach for Planar Shape Morphing,” in: Proc. 12th Conf. on Computer Graphics and Applications, IEEE, 2004, pp. 111-120.

[35] H. Helmke, M. Slotty, M. Poiger, D. Ferrer Herrer, O. Ohneiser, N. Vink, A. Cerna, P. Hartikainen, B. Josefsson, D. Langr, R. García Lasheras, G. Marin, O. Georg Mevatne, S. Moos, M.N. Nilsson, and M. Boyero Pérez, “Ontology for Transcription of ATC Speech Commands of SESAR 2020 Solution PJ.16-04,” 37th AIAA/IEEE Digital Avionics Systems Conference (DASC), London, UK, 23.-27.09.2018.

[36] Geelvink, N. Grafische Wetterdatendarstellung: Für mehr Planungssicherheit und Kapazität in der Luft. Langen, Deutsche Flugsicherung GmbH, 2009.

[37] Maurus, K. and Belz, C, “Luftverkehr in Deutschland: Mobilitätsbericht 2017”, DFS Deutsche Flugsicherung, Langen, Germany, 2018.

Please have a Look here: Successful Guidance of Air Traffic Controller's Attention

Oliver Ohneiser; Hejar Gürlük; Malte-Levin Jauer
German Aerospace Center (DLR),
Institute of Flight Guidance,
Lilienthalplatz 7, 38108 Braunschweig, Germany
Oliver.Ohneiser@DLR.de

Ádám Szöllősi; Dóra Balló
HungaroControl Zrt.,
Igló u. 33-35,
1185 Budapest, Hungary

***Abstract*—Keeping the operator's attention on the right spot of the situation data display is one of the key factors to successfully guide air traffic. However, this becomes particularly difficult with complex and dense traffic situations displayed on larger screens. This paper describes our developed prototypic attention guidance (AG) system for air traffic controllers (ATCO). This system uses eye-tracking as an input for the ATCO's current attention. Different attention guidance was implemented for specific air traffic control (ATC) events such as handover and conflict alerts. For those events, different visual cues are presented step-wise within various levels of escalation in case the ATCO did not pay attention to ATC events. The AG system was tested in human-in-the-loop validation trials with five ATCOs. The simulated Hungarian Flight-Centric airspace was chosen as a test-case. The validation trials revealed promising results for the Solution controller working position (CWP), which was equipped with AG functionality. ATCOs reported less workload and improved situation awareness with the Solution CWP than without AG support. Increased acceptance and confidence with the Solution system were also reported. ATCOs felt strongly supported by our robust and smoothly interacting attention guidance system encouraging further development of our prototype towards operational use.**

Keywords—Air Traffic Controller; Attention Guidance; Controller Working Position; Human Machine Interface; Eye-Tracking; Visual Cues; Flight-Centric Air Traffic Control

I. INTRODUCTION

Air traffic controllers are scanning their situation data display continuously in order to keep up situation awareness and handle all relevant events safely and timely with respect to urgency and importance. Thus, ATCOs have to determine and prioritize ATC events and plan their controller tasks accordingly. As ATC events are mostly connected to aircraft on the radar screen, there are relevant screen sections that need to be looked at and paid attention to.

This task becomes more difficult with more complex and dense traffic, particularly when displayed on large format monitors. Current human machine interfaces (HMI) often provide single types of alerts to warn the ATCO. These mostly visual alerts can be over-looked or may be too intrusive with regards to the aforementioned challenges.

The feedback cycle providing the ATCO's reaction to the HMI only starts with the conflict resolving. In conclusion, there is a demand for smooth and non-intrusive guiding of ATCO's attention to the relevant HMI spots.

In our prototypic AG system, this is done via an eye-tracking system determining which area the ATCO currently looks at. The corresponding data feeds a trigger algorithm for relevant ATC events. It is then used to potentially escalate visual cues on the HMI in different steps if the ATCO does not pay attention to the relevant spots.

This paper reviews related work on the topic of attention and its guidance in chapter II. Chapter III outlines the concept for an Attention Guidance (AG) prototype at a Controller Working Position (CWP) and its implementation within the given Flight-Centric Air Traffic Control (FC-ATC) use case. The study setup to reveal benefits and drawbacks of the AG prototype is outlined in Chapter IV. Chapter V presents the results of the AG Human-in-the-Loop study. Those results are discussed in Chapter VI. Chapter VII summarizes, concludes, and gives an outlook on future work.

II. RELATED WORK

A. Operator Attention

According to Broadbent's filter metaphor [1], the attention of a human operator represents a filter to the environment. This filter reduces irrelevant input to focus on relevant – potentially multiple – input streams [2]. This selective filter avoids overload of the human's brain [1]. Many aspects such as task demands, operator's situation understanding, different channels and senses as well as the related perceptual limits influence the effectivity of the filter [2]. Operator's attention is connected to distinguishing, remembering, reacting in a certain amount of time, perceiving, and conceiving [3].

The operator's attention and thus filtering is a prerequisite for proper situation awareness seen as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" following Endsley [4]. As the situation awareness itself is limited, meaning that better understanding of some elements decreases understanding of others, this factor is especially important in complex environments [5].

Air traffic control can be seen as such a complex environment where visual attention plays a key role for proper situation awareness. There are six tasks that ATCOs need to fulfill [2] related to visual attention. (1) Scanning and orienting in a goal-directed and undirected way. (2) Supervisory control to assure that aircraft characteristics stay within a required range.

founding members
9th SESAR Innovation Days
2nd – 5th December 2019
ISSN 0770-1268

(3) ATCOs need to notice unexpected events like emergencies when monitoring the radar screen. (4) ATCOs are searching for specific aircraft to issue clearances. (5) ATCOs read all relevant information such as radar labels shown on the situation data display. (6) After a clearance has been given to an aircraft there must be a follow-up check if the current flight behavior is conform to the desired behavior. The above knowledge about visual attention and situation awareness needs to be taken into account when attempting to guide operator's attention.

B. *Guiding Attention*

In the SESAR2020 exploratory research project Mitigating Negative Impacts of Monitoring high levels of Automation (MINIMA), some eye-tracking based guidance of attention has already been done [6]. A semi-transparent circle was used to highlight aircraft on the radar screen that have not been looked at for a certain amount of time. However, this highlighting of inattention did not take the relevance and necessity to pay attention to this specific aircraft into account. A study on attention in a supervisory task with large displays revealed that movements were better for notifications than (animated) color [7]. Faster perception of notified aircraft was achieved by background opacity, concentric circles, and pulsating boxes. However, the intrusiveness differed on the use of color [7].

Peripheral cues should be presented next to the target stimulus if ATCO's view is anywhere else in order to enable proper cue recognition. Those cues need to be salient enough to be recognized [2] especially if they are far away from the current gaze focus. However, cues should not mask other information. They should also not be displayed too often or too long as this might distract ATCOs. Reliable cues can be achieved via integration of multiple sensor data and a model-based AG system. If cues are reliable, exogenous cues are preferred over endogenous cues as they are processed faster by humans [8] [9]. These aspects were considered for the AG concept and their implementation is described in the following section.

III. Attention Guidance Concept and Implementation

A. *Basic Assumptions for the AG Concept*

The display spot, where (a) the ATCO's gaze targets at respectively (b) the mouse-cursor on the radar display (when moved), are assumed as the ATCO's current area of visual attention. Thus, this area is determined (a) via an infrared contact-free eye-tracker mounted at the bottom of a 40 inch screen respectively (b) via mouse data positions. An assistance system of the FC-ATC environment calculates the next relevant ATC events – so called triggers. If there are multiple unsolved ATC events, the AG system prioritizes them with respect to urgency and importance. In most cases, the ATC event with highest priority requires the ATCO's attention. If the current and preceding area of attention of the ATCO does not match the expected area of attention, the AG system checks if the minimum look-ahead time of some seconds or minutes depending on the ATC event type has passed. If so, the AG system raises the escalation level [10] and raises it even further if the ATCO still does not pay attention.

B. *Escalation Levels and De-Escalation for ATC Events*

The AG system comprises four different escalation levels (0 to 3) if there is a relevant ATC event. The current AG implementation includes short-term conflict alerts (STCA), medium-term conflict alerts (MTCA; with and without right of way) as well as handover events. Basic level 0 identifies the state without additional visual cues. As short-term conflicts are very important, they do not have an escalation level 0, but are directly escalated to level 1. In general, there is a rectangle frame around the ATC event affected aircraft radar label in escalation level 1. The frame is accompanied by a round semi-transparent flashlight effect in level 2. This flash-light effect obtains a colored "glowing" circular frame in escalation level 3. The higher the level, the more salient the visual cue (compare figure 1).

Figure 1. Screenshots of the three escalation levels' visual cues for a handover event.

If the controller notices all visual cues related to an ATC event, the escalation level will immediately be set to zero and the cue disappears. In case of a handover event it is just this aircraft that needs to be noticed. In case of an aircraft conflict all involved aircraft need to be noticed. There are two options for an ATCO to notice a visual cue. The primary option is to look at the aircraft radar label or head symbol. The eye-tracker will then detect the ATCO's gaze and initiate the de-escalation. The eye-tracker has a certain detection range of a few centimeters, so that multiple aircraft can be noticed at a time. This bigger detection range allows for more body movement of the controller that otherwise would result in less robust recognition. The secondary option is to move the mouse cursor over the radar label to indicate the noticing of the event. For more details on the AG concept, please refer to [11].

C. *ATC Event Resolution*

Even if the ATC event was noticed, the AG system will remain active in the background until the event was actually resolved. The resolution is carried out by the ATCO usually by issuing clearances into the aircraft label or by coordinating with other ATCOs. If the event resolution takes longer than a certain ATC event type dependent threshold time and the controller again did not notice the involved aircraft for a certain amount of time, another escalation through the levels will follow. The non-resolution can have two reasons. Either, the ATCO forgot or did not actively notice the event at all. However, from a safety perspective, this does not matter a lot due to the re-escalation.

D. *Use Case: Flight-Centric Air Traffic Control*

SESAR2020 (Single European Sky Air Traffic Management Research Programme) foresaw the integration and validation of an Attention Guidance prototype as part of PJ.16-04-03 within the Flight-Centric ATC environment of PJ.10-01b.

founding members
9th SESAR Innovation Days
2nd – 5th December 2019
ISSN 0770-1268

The Flight-Centric ATC concept focuses on ATCO's responsibility for a number of aircraft instead of geographic airspace sectors. This concept has been researched at DLR in cooperation with the German ANSP DFS Deutsche Flugsicherung GmbH since 2008 [12]. The general feasibility has been proven for the upper airspace area [13]. Furthermore, assignment strategies have been analyzed [14] so that the incoming traffic is balanced between CWPs in current FC-ATC software. ATCO support tools for conflict detection and planning are another essential factor to enable the FC-ATC concept [15].

It has to be noted that the Flight-Centric ATC part just served as a use case to integrate and present the Attention Guidance functionality. Thus, results reported in this paper focus on the comparison of a CWP equipped with an AG system compared to a CWP without an AG system and do not consider benefits or drawbacks from the FC-ATC environment. All ATC events, related times, and the radar appearance of basic escalation level 0 are part of FC-ATC.

IV. Human-in-the-Loop Study with Attention Guidance Prototype

A. Validation Setup with Software, Hardware, and Simulation Exercise Staff

The Human-in-the-Loop Study to evaluate the DLR Attention Guidance prototype took place on January 17th, 2019 at HungaroControl premises in Budapest, Hungary.

The final software setup consisted of the FC-ATC software (provided by the DLR pilot assistance department) on the one hand and the AG software (provided by the DLR controller assistance department) on the other hand. The FC-ATC part included the traffic simulation itself, the aircraft assignment, automatic conflict solving options, communication infrastructure, and finally the situation data display.

The FC-ATC software was connected to the AG system, which is the logical core part of the presented concept. From a software engineering perspective, the FC-ATC software can be seen as the "model" and the "view", whereas the AG system is the "controller" governing the current appearance of the HMI.

The visual cues of the escalation levels are shown on the ATCO display as the handover example in figure 1 points out. Other ATC events are displayed in a very similar but differently colored way. Details can be found in [11].

Five Flight-Centric ATC CWPs with height movable chairs for the eye-tracking calibration were available as demonstrated in figure 2. The ATCOs communicated with five simulation pilots. The traffic – respectively its coordination – was automated. The simulation crew was available in the background in case of upcoming questions.

The average age of the five participating HungaroControl ATCOs was 33.2 years (standard deviation, SD: 10.1 years), with an average job experience as a controller of 7.4 years (SD: 10 years). All of them were en-route area control center (ACC) controllers for Hungary (LHCC flight information region, FIR, Budapest).

Figure 2. SESAR2020 PJ.16-04-03 Attention Guidance Human-in-the-Loop validation exercise EXE-16.04-TRL4-TVALP-310 with DLR AG prototype and five ATCOs at HungaroControl Simulation Hub in Budapest, Hungary.

B. Simulation Run Conditions

For the simulation runs two different conditions were designed in order to compare effects on human performance. One condition depicted a Baseline Flight Centric CWP without AG functionality, whereas the Solution CWP represented the other experimental condition encompassing the Flight Centric CWP with AG functionality. Both conditions incorporated a high density Flight-Centric ATC traffic scenario in order to raise the probability of provoking higher numbers of aircraft conflicts or missed aircraft handovers. For a traffic setup time and familiarization, both simulation runs started with a 7.5 minutes initialization phase running in double simulation speed. Afterwards, the runs were directly continued in real time (announced by the simulation staff) until one hour simulation run time was completed.

As all five ATCOs changed their CWP for the second run, they had to handle different assigned aircraft of the complete Hungarian airspace and thus different air traffic situations in both runs to avoid a scenario learning effect. A restart of the display was necessary for ATCOs 3 and 4 in the Solution condition. This restart lasted 23 respectively 27 seconds, but should not have significantly affected any of the results compared to the simulation duration of 3,000 seconds.

C. Organizational Preparation of Simulation Runs

The ATCOs already received a pre-briefing document some weeks in advance comprising AG prototype functionalities and the schedule. The trials started with a briefing explaining the purpose of the study and an overview of the AG system functionalities including escalation levels and visual cues. Furthermore, participant agreement sheets needed to be signed and a demographics questionnaire to be filled. Afterwards, a short training run familiarized the ATCOs with the eye-tracking calibration process and the further visual cues in case of ATC event escalation. As the ATCOs were already trained with the Flight-Centric ATC CWP the days before, the training concentrated on the appearance of visual cues and how to let them disappear.

D. Eye-Tracking Calibration Process

The eye-tracking calibration was done for each controller before participating in the Solution simulation run. First, a comfortable and technically appropriate seating position regarding chair height and distance to the screen needed to be found.

founding members
9th SESAR Innovation Days
2nd – 5th December 2019
ISSN 0770-1268

After that, the ATCO's gaze had to be fixed on four specific spots on the screen for some seconds in order to let the eye-tracker's infrared sensors learn the pupils' positions.

Finally, aircraft (radar labels) that are recognized as being noticed by the ATCO on the radar display were highlighted in yellow. This feature for eye-tracking recognition demonstration was switched off for the validation runs and just served for calibration transparency reasons.

E. AG Data Acquisition Activities: Simulation Runs, Questionnaires, and Debriefing

During the following first simulation runs two ATCOs worked with the Solution CWP and three ATCOs with the Baseline CWP. ATCOs filled out the common PJ.16-04 human performance questionnaire after the run and before the subsequent break. Then, eye-tracking calibration was executed for the other ATCOs; this time, three of them worked with the Solution and two ATCOs with the Baseline CWP. This order was chosen to balance the run sequence. The second round of questionnaires again included the human performance questionnaire. In addition, the tailor-made Attention Guidance parts were filled out by the ATCOs.

The questionnaire items, statements, scales, and other details are explained in results Section V.A. The final group debriefing followed a semi-structured interview method. Log files of eye-tracking and mouse data of ATCOs as well as ATC event data were recorded during the Solution run. The complete validation exercise lasted slightly more than 3.5 hours in the afternoon.

V. Results of Attention Guidance Validation Exercise

This section presents the results with respect to human performance questionnaires, AG log files, and more general statements. Values for questionnaire ratings and times are reported as arithmetic averages with standard deviations SD:

$$SD = \sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$$

where n is the number of values, $\bar{x}$ is the arithmetic average, and x is the rating value. The above equation is used, as values are a random sample of the population.

A. Subjective Controller Ratings on Human Performance and the AG concept

Five questionnaire parts mainly with Likert scales [16] have been answered by the ATCOs two times (after each simulation run). The questionnaires were extracted from the EUROCONTROL Human Performance repository [17]. In addition, a more general AG questionnaire was answered just once after finishing both simulation runs to compare the Solution system with the Baseline system.

1) Situation Awareness

First, ATCOs filled in the China Lakes Scale [18] and rated their situation awareness on a decision tree from low (1) to high (10). Average situation awareness score was 8.25 (SD: 0.96) for the Baseline system, but 9.0 (SD: 0.82) for the Solution system.

Secondly, ATCOs rated five items about - traffic understanding, aircraft messages, coordination, and identification during the run from bad (1) to good (7). The awareness score was 5.84 (SD: 1.06) for the Baseline system and 5.72 (SD: 1.19) for the Solution system.

2) Workload

First, ATCOs filled in the Bedford Workload Scale [19] from easy (1) to hard (10) for peak and average workload experienced during the run. The peak workload was 5.4 (SD: 1.34) for the Baseline system and 4.8 (SD: 1.34) for the Solution system, respectively. The average workload was 3.4 (SD: 1.14) for the Baseline system and 3.4 (SD: 0.55) for the Solution system, respectively, as shown in figure 3.

Secondly, ATCOs rated an ATC Workload Scale with nine items from easy (1) until hard (10) for multitasking, planning, decision making, team awareness, information processing, attention direction, problem solving, memory management, and maintaining awareness as experienced during the run. The average workload score was 5.16 (SD: 0.71) for the Baseline system and 3.96 (SD: 0.5) for the Solution system, respectively (see figure 3).

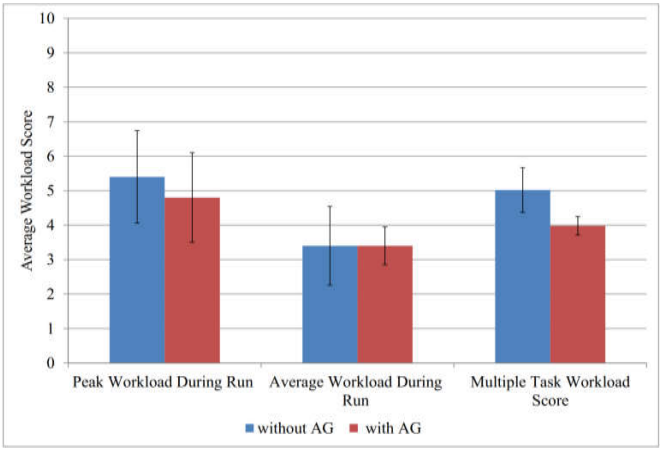


Figure 3. ATCO workload ratings for Baseline without AG (blue) and Solution with AG (red) with standard deviations as black vertical lines.

3) Usability and Controlling Tasks

ATCOs had to rate 19 different statements regarding usability and controlling tasks about air traffic control functionalities on a scale from "strongly disagree" (1) to "strongly agree" (5) or even "not applicable". The average for the Baseline system was 3.86 (SD: 0.69), for the Solution system 4.0 (SD: 0.53).

The five statements about separation determination, coordination of in- and outbound traffic, regular scanning cycle, conflict detection, and separation assurance had a 0.4 points better value for Solution than Baseline in average. The statement that ATCOs were able to rapidly prioritize alerts was even rated 1.25 points better in Solution than Baseline.

4) User Acceptance

The adapted Controller Acceptance Rating Scale (CARS [20]) delivered an average value between bad (1) and good (10).

ISSN 0770-1268

For the five ATCOs it was at 2.8 (SD: 1.64) for the Baseline system, but 7.4 (SD: 3.24) for the Solution system (figure 4).

5) User Confidence

Four statements on confidence with a scale from "completely disagree" (1) to "completely agree" (10) had to be rated. Statements asked if the tool supports work, if ATCOs feel adequately trained, if information is suitable for their tasks, and if an overall confidence is given.

User confidence was at an average of 4.5 (SD: 3.3) for the Baseline system and 5.95 (SD: 2.95) for the Solution system, respectively, as presented in figure 4. One ATCO noted that his score for the Solution system was low as he could not split the FC-ATC and AG part for his rating.

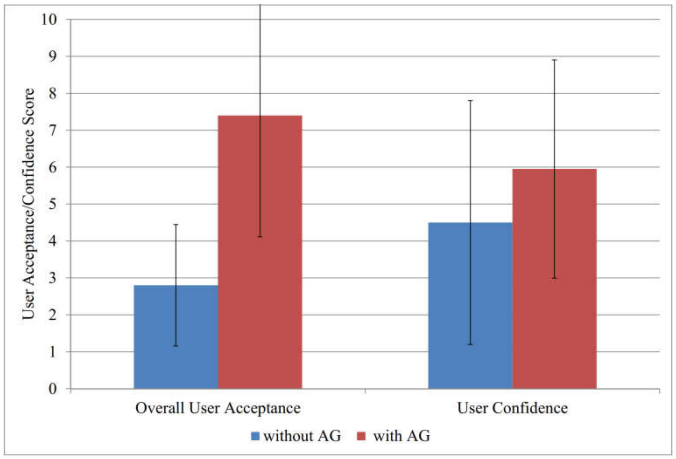


Figure 4. ATCO acceptance and confidence for Baseline without AG (blue) and Solution with AG (red) with standard deviations as black vertical lines.

6) Tailor-made Attention Guidance Rating

The ATCOs rated six statements about the AG concept and the hardware setup on a scale from "I do not agree at all" (1) to "I totally agree" (5). The statements were about if the display of AG escalation levels for STCA/MTCA/handover are understood easily, if the AG logic is transparent, if the eye-tracking works reliably, and if the radar screen is sufficiently large. ATCOs gave 30 single rating scores with an average of 4.73 (SD: 0.25).

7) Summarized and Normed AG Rating

Parts 1 to 6 comprise nine different questionnaire sections with 47 rating items. For better readability and comparability, all scales have been normed (scale from 1 to 10) and some scales have been inverted so that a higher score is always better than a lower score in figure 5.

Hence, the Usability and Controlling Tasks part, as well as the tailor-made AG part have been normalized by multiplying with 2 to enlarge the rating scale from 1 to 5 up to 1 to 10. The second situation awareness part with a scale from 1 to 7 has been multiplied with a factor of 10/7. The workload ratings have been inverted (10 to 1; 9 to 2; …; 1 to 10) as lower workload scores indicate better results.

This inverted rating was combined with the other workload parts into a single so called "Relax Score". Furthermore, both situation awareness parts as well as the usability and controlling part were combined to a single score as well as another single score for the user acceptance and confidence parts as figure 5 shows.

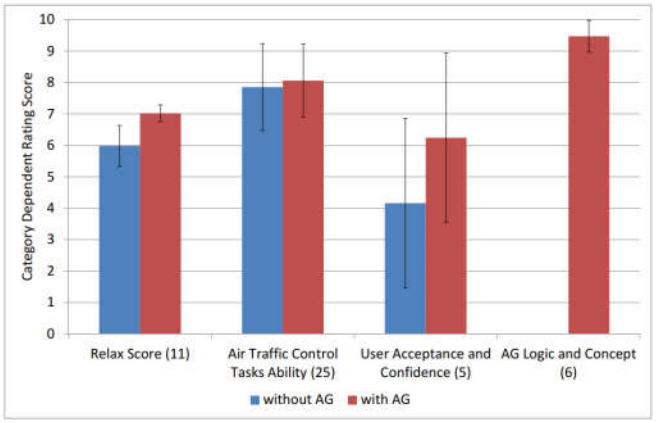


Figure 5. Combined questionnaire parts with number of single rating items in brackets for both conditions: Baseline without AG (blue) and Solution with AG (red) with standard deviations as black vertical lines.

B. Log File Results on ATC Events and Aircraft Noticing

1) Number of Escalated ATC Events

114 ATC events in total have been escalated in the five simulation runs of 50 analyzed net minutes duration per each of the five ATCOs. This means 22.8 escalated ATC events per ATCO (SD: 2.9). 108 of the escalated ATC events were noticed by the ATCOs detected via eye-tracker, just three ATCOs noticed two escalated handovers, each detected via mouse-over functionality.

95 of those ATC events were handovers (83.3%), 15 medium-term conflict alerts (13.2%; thereof 3 with right of way and 12 without right of way), and four short-term conflict alerts (3.5%). Just one ATCO had no STCA during the high-density Flight-Centric ATC scenario; however this ATCO had the most MTCAs. 75 ATC events were escalated until level 1 (65.8%), 30 events until level 2 (26.3%), and only 9 events until the highest escalation level 3 (7.9%) as shown in figure 6.

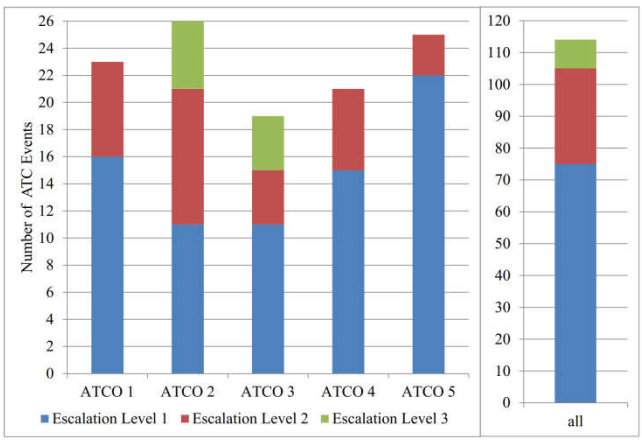


Figure 6. Number of ATC events escalated to level 1/2/3 (blue/red/green) until detection per ATCO and over all ATCOs.

founding members
9th SESAR Innovation Days
2nd – 5th December 2019
ISSN 0770-1268

Regarding just handover escalations, 62.1% (59) were noticed in level 1, 28.4% (27) in level 2, and 9.5% (9) only in level 3. 16 conflict alerts (both medium- and short-term) have been noticed in level 1, the other three in level 2. Just two ATCOs experienced ATC events that were escalated up to the highest level 3 (see ATCO 2 and 3 in figure 6).

All escalated ATC events have been resolved during the simulation time except of 10 handovers and 6 medium-term conflict alerts that only appeared during the last minutes. Hence, there would have been time to solve these conflicts if the simulation would have continued.

2) Visual Escalation Cue Noticing Times

In average it took an ATCO 8.3 seconds to notice an escalated ATC event (SD: 10.2s). For all types of conflicts – so without handovers – the average time was 3.1 seconds (SD: 3.0s). From the highest escalation level, a general ATC event was escalated to (1, 2 or 3), it took the ATCO only 3.6 seconds (SD: 5.6s) to notice the visual cue. The noticing time for conflict alerts from the highest escalation level for all ATCOs was only 2.3 seconds (SD: 1.8s). As one ATCO needed more than twice as much time to notice escalated ATC events compared to the average of the others, this heavily influences the above reported average times (see figure 7).

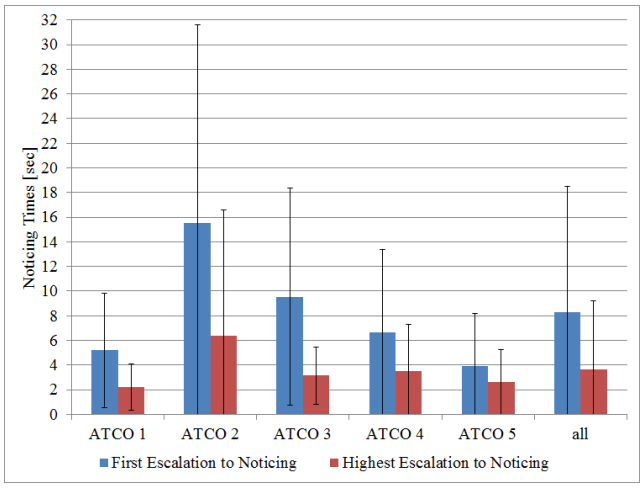


Figure 7. Average time needed to notice an escalated ATC event per ATCO and over all ATCOs (blue left columns). Average time needed to notice an ATC event after its highest escalation level per ATCO and over all ATCOs (red right columns); with positive/negative standard deviation (black lines).

C. AG Concept Results and General Comments

1) Tailor-made Attention Guidance Questionnaire

ATCOs were also asked to complete eleven open questions comparing the Baseline and the Solution once after finishing both simulation runs. Four ATCOs preferred the Solution system over the Baseline system (one ATCO gave no answer on this).

This paper does not analyze the FC-ATC part that served as a Baseline system. Hence, only the following summarized view on the ATCO's questionnaire statements 1 to 5 is given for completeness reasons: Information filtering and situation awareness were poor for the time with all potential subsequent effects on safety.

Some meaningful and mostly positive answers about the Solution system (questionnaire statements 6 to 10) are paraphrased in the following. When using the Solution system, ATCOs liked best that incoming and outgoing traffic as well as conflicts were highlighted if not scanned or being forgotten. In addition, the general AG idea was liked. The system was rated as simple to use and really helped to find blind spots.

The majority of ATCOs did not mention any disadvantage for the Solution system. Moreover, they only saw advantages as the AG system complements existing features. However, one ATCO reported that the system sometimes requires taking the focus away from an area or problem which needs to be focused on. Thus, it might be better suited only for training new ATCOs. Others especially wanted to work with the visual guidance tool (Solution system) in dense traffic situations or when they are more tired in order to draw the attention back. The only issue with the AG system might exist if somebody does not want to get eye-tracked. Some ATCOs wished that the AG system would already have been implemented in their CWP as it really provided them assistance.

2) Debriefing Comments

In the common verbal debriefing session all five ATCOs preferred the Solution to the Baseline. Some feedback sentences are paraphrased in the following. ATCOs found the visual cues to be non-intrusive. They also thought that eye-tracking worked really robust after calibration and thus the interaction was fine. The AG functionality really supported ATCOs to have a look at HMI spots that they would not have looked at this timely. The AG system was assessed as already ready to be used in operational life. Furthermore, most ATCOs would like to have AG in their CWP, in the near future, regardless of Flight Centric ATC environment.

ATCOs wanted to have individual settings for aircraft highlighting in case of handovers, i.e. a possibility to switch on and off respectively a different visual cue style. Some preferred to only highlight aircraft when they left their sector without being handed over to the next position or entered their sector without being handed over to them respectively aircraft being assumed or not assumed. Others also liked the highlighting before aircraft enter or exit their sector. The possibility to switch off the complete AG functionality existed, but was not explained to the ATCOs as exactly this functionality needed to be evaluated.

The mouse-over functionality was only used in seldom cases as the eye-tracking was very accurate. ATCOs also found that the debugging functionalities of the eye-tracking could be very useful for training purposes. The instructor could improve scan patterns of trainees better when visualizing the gazes. This could lead to a whole new methodology in ATCO training.

ATCOs felt well supported by the AG system and reported unanimously that AG provided additional benefit to their HMI working routines compared to working at the FC-ATC CWP without AG assistance. ATCOs reported they were fine with respect to the timing behavior of cues except of MTCAs that should only be escalated with eight minutes look-ahead time to avoid false alerts. ATCOs rarely observed intrusiveness if they intensively "worked" at a different spot.

Then it might be slightly disturbing if their peripheral attention was tracked to look somewhere else. Many ATCOs said there was nothing annoying at all with the AG functionality.

VI. Discussion of Attention Guidance Study Results

A. Human Performance Results Discussion

The human performance results showed ATCOs' preference for the Solution system with attention guidance functionality over the Baseline system. However, no statistical significance was tested due to the limited number of test subjects (n=5). Therefore, results showed tendencies in ATCO ratings, but have differences in mean values bigger than the corresponding standard deviations in some of the categories indicating a strong tendency.

The Solution system had a better score for overall situation awareness, but no great difference between compared systems in the second situation awareness questionnaire. The Solution system showed slightly better usability and support for controlling tasks than the Baseline system in average. This small difference in the rating might also be affected by the FC-ATC part which in general was not perceived to provide good situational awareness.

Also ATCOs did not perceive significant differences in the average workload during the simulation runs. This seems to be logic as the scenarios were the same and ATCOs just had to handle different aircraft but roughly the same number of aircraft in their runs. However, the peak workload was much lower for the Solution system. Furthermore, the workload scale with nine questionnaire items showed a much lower score for the Solution system. Hence, workload seems to be lower with AG especially in dense and complex traffic operations. ATCOs may handle ATC events earlier without getting into stress situations, indicating a confirmation of the envisioned benefits of the AG system.

User acceptance and user confidence were rated much better for the Solution system, too. This can be a prerequisite when introducing this new system later on.

B. Log File Results Discussion

Almost two thirds of escalated ATC events have already been noticed in escalation level 1. Roughly another quarter was noticed in level 2. This shows the effectivity of visual cues to guide ATCO's attention, while still remaining non-intrusive as the controller statements show.

The evaluation of log files additionally confirms the suitability of the developed AG prototype for the improvement of situation awareness and timely handling of ATC events: All escalation levels did occur during the simulations and therefore helped the operator noticing ATC events he/she was previously unaware of.

Furthermore, the decreasing number of occurrences of higher escalation levels (see figure 6, right) and the reduced noticing time of events with higher salience (e.g. conflicts) undermines the suitability of the HMI design in terms of showing the most important information first.

As in general, the reaction times are within a one digit number of seconds and ATCOs reported being pointed to spots that they would not have realized so fast otherwise, demonstrates the efficiency potential of the AG system. However, some reaction times were also slower than ten seconds. Probably, ATCOs saw something in their peripheral view, however, finished their task at their current area of interest, but were aware of that they need to shift their attention afterwards. This fact can especially be assumed for less time-critical handover events. ATCOs also relied on the eye-tracking system for aircraft noticing rather than using the mouse-over functionality.

To sum up the time aspect, we took effective measures for the implementation of our AG prototype to ensure that ATC events with high priority are noticed by the controller in a timely manner to support safety.

C. Tailor-made Attention Guidance Questionnaire and Debriefing Comments Discussion

The results of the tailor-made AG ratings showed that the ATCOs had a clear understanding of the AG logic and that the system is viewed as being robust in this early stage of development. ATCOs' feedback on the AG functionality was almost completely positive. The only negative statements comprised the guidance of attention in situations where the ATCO likes to keep attention anywhere else and theoretical concerns on data privacy due to eye-tracking.

ATCOs did not feel patronized, but really felt supported individually by AG. The Solution system helped to put attention on important display areas that otherwise would have been looked at only later. ATCOs also experienced a well-working assistance as it was robust and non-intrusive. They even wanted to have it for their conventional CWP and formulated ideas for enhancements and further ATC events that could be included with respect to attention guidance. The ATCO statements stand for themselves and support the core AG validation result even more than just the subjective questionnaire ratings.

VII. Summary and Outlook

A. Summary of Attention Guidance Validation Trials

Our attention guidance prototype – based on prioritized ATC events and eye-tracking data of ATCOs – was successfully implemented and tested in the FC-ATC environment. The validation exercise of the AG prototype revealed very motivating results. ATCOs felt supported by the visual cues of escalated ATC events for handovers, medium- and short-term conflicts. As they were reminded of conflicts in case of non-resolution, AG may also serve as an additional safety net.

The event noticing times also depending on the escalation level were in the range of a radar update (few seconds). Even if not significant in all categories, relax score, ATC tasks ability (also comprising situation awareness), as well as user acceptance and confidence were higher using the Solution CWP with AG functionalities. ATCOs also rated the AG logic and concept very high.

founding members
9th SESAR Innovation Days
2nd – 5th December 2019
ISSN 0770-1268

The debriefing feedback was really encouraging. It hardly happens that ATCOs wish to have a new functionality – in their daily life CWP – to be noted that these were just first trials of a prototype. Furthermore, the used low budget eye-tracker and the few adaptations that would be necessary to integrate an Attention Guidance system into a CWP promise to deliver reasonable support for air traffic controllers.

B. Outlook on Future Work

Further ideas of which visual cues to escalate additionally or which aspects to be customizable have been developed. ATCOs uttered the idea to adjust some of the visual cues by their own. The personalization of AG settings as outlined in section V.C.2 is easily doable from a technical point of view, but needs further analysis with respect to a common CWP functionality basis.

Some ATCOs wanted to reduce the escalation time in advance of a medium-term conflict as some of them could be false alerts. ATCOs even wished to use the visual cues also for other ATC events. Escalation should be done in case of wrong Mode-S settings in the cockpit, for route adherence monitoring (RAM), approaching restricted areas, cleared flight level conformance alarm (CLAM) events, and if the current flight level is different to the exit flight level with the aircraft being close to the exit point.

The trigger algorithm could be adapted so that whenever the operator notices an important event, a "working time" is defined. Only after this working time has passed, the trigger logic will continue generating visual cues for the high-priority events. This could avoid unintended guiding of ATCO's attention. One ATCO zoomed far out of his airspace looking on the whole European map to check whether all aircraft in the very center are detected as being noticed by him at the same time. ATCOs did not retry this during the simulation runs. However, it is of course one aspect to adapt the eye-tracking noticing area depending on the radar display zoom step.

The majority of ATCOs also wanted to have a visual cue if an aircraft was not looked at for a longer period of time. This should also be valid for all radar targets on the display that are not correlated with a flight plan. However, airspace regions should rather not be highlighted in general if there was a lack of attention. Although, this indicator could be tested in a further study after implementing the respective functionality.

The AG concept will be adapted to other laboratories and training CWP environments like (multiple remote) tower in the future. It will also be enhanced with additional auditory cues. Furthermore, a coupling of the ATCO authentication with the pre-defined user profile settings could ease the use. Despite all those ideas for further refinement (respectively enhancement) of the AG concept and its implementation, ATCOs found the AG system already ready for the next step towards operationalization.

ACKNOWLEDGMENT

The PJ.16-04 CWP HMI project also comprising the Attention Guidance activity (PJ.16-04-03) has received funding from the SESAR Joint Undertaking under the European Union's grant agreement No. 734141.

REFERENCES

[1] D. Broadbent, "Perception and Communication," London: Pergamon Press, 1958.

[2] C.D. Wickens, "Engineering Psychology and Human Performance," 4th Edition, Psychology Press, 2015.

[3] W. James, "The principles of psychology," 2016, website cited May 14, 2019, URL: http://ebooks.adelaide.edu.au/j/james/william/principles/chapter11.html.

[4] M.R. Endsley, "Situation Awareness Global Assessment Technique (SAGAT)," in "Proceedings of the IEEE 1988 National Aerospace and Electronics Conference: NAECON," 1988, p. 792.

[5] M.R. Endsley, "Designing for Situation Awareness in Complex Systems," in "Proceedings of the 2nd International Workshop on Symbiosis of Humans, Artifacts and Environment," Kyoto, Japan, 2001.

[6] O. Ohneiser, F. De Crescenzio, G. Di Flumeri, J. Kraemer, B. Berberian, S. Bagassi, N. Sciaraffa, P. Aricò, G. Borghini, and F. Babiloni, "Experimental Simulation Set-Up for Validating Out-Of-The-Loop Mitigation when Monitoring High Levels of Automation in Air Traffic Control," in "International Journal of Aerospace and Mechanical Engineering," 12 (4), 2018, pp. 307-318.

[7] J.-P. Imbert, H.M. Hodgetts, R. Parise, F. Vachon, F. Dehais, and S. Tremblay, "Attentional costs and failures in air traffic control notifications," in "Ergonomics," Taylor & Francis, 57 (12), 2014, pp. 1817-1832.

[8] J. Theeuwes and R. Godjin, "Parallel Allocation of Attention Prior to the Execution of Saccade Sequences," in "Journal of Experimental Psychology, Human Perception and Performance," 29, 2003, pp. 882-896.

[9] J. Jonides, "Voluntary versus automatic control over the mind's eye's movement," in "Attention and Performance," 9, 1987, pp. 187-203.

[10] H. Springborn, "Design and Assessment of Methods of Attention Guidance for the Sector-Less Air Traffic Management Controller Working Position," Master's Thesis, FH Joanneum, Graz, Austria, 2017.

[11] O. Ohneiser, M.-L. Jauer, H. Gürlük, and H. Springborn, "Attention Guidance Prototype for a Sectorless Air Traffic Management Controller Working Position," German Aerospace Congress (DLRK), Deutsche Gesellschaft für Luft- und Raumfahrt - Lilienthal-Oberth e.V., 4.-6. Sep 2018, Friedrichshafen, Germany.

[12] B. Korn, C. Edinger, S. Tittel, D. Kügler, T. Pütz, O. Hassa, and B. Mohrhard, "Sectorless ATM — A Concept to Increase En-Route Efficiency," in "Proceedings of the 28th Digital Avionics Systems Conference (DASC) 2009", Orlando, FL, USA, 2009.

[13] M. Biella, B. Birkmeier, B. Korn, C. Edinger, S. Tittel, and D. Kügler, "Operational Feasibility of Sectorless ATM," in "Proceedings of the International Conference of the European Aerospace Societies (CEAS) 2011", Venice, Italy, 2011.

[14] A.R. Schmitt, C. Edinger, and B. Korn "Balancing Controller Workload Within a Sectorless ATM Concept," CEAS Aeronautical Journal, 2, 2011, pp. 35-41.

[15] B. Birkmeier, "Feasibility Analysis of Sectorless and Partially Automated Air Traffic Management," PhD dissertation as DLR Forschungsbericht 2015-12, ISSN 1434-8454, 2015.

[16] R. Likert, "A Technique for the Measurement of Attitudes," in "Archives of Psychology 22," No. 140, 1932, pp. 5-55.

[17] EUROCONTROL, SESAR JU electronic "Human Performance repository," 2013, website cited May 14, 2019, URL: https://ext.eurocontrol.int/ehp/?q=Home.

[18] M.R. Endsley, "Measurement of Situation Awareness in Dynamic Systems," Human Factors, 37, 1995, pp 65-84.

[19] A.H. Roscoe and G.A. Ellis, "A subjective rating scale for assessing pilot workload in flight: A decade of practical use," RAE-TR-90019, Royal Aerospace Establishment Farnborough, United Kingdom, 1990.

[20] K.K. Lee, K. Kerns, R. Bone, and M. Nickelson, "The Development and Validation of the Controller Acceptance Rating Scale (CARS): Results of Empirical Research," Proceedings of the 4th USA/Europe Air Traffic Management R&D Seminar, Santa Fe, NM, USA, 2001.

aerospace

MDPI

Article

Integrating Eye- and Mouse-Tracking with Assistant Based Speech Recognition for Interaction at Controller Working Positions

Oliver Ohneiser [1,2,*], Jyothsna Adamala [3] and Ioan-Teodor Salomea [4]

[1] German Aerospace Center (DLR), Institute of Flight Guidance, Lilienthalplatz 7, 38108 Braunschweig, Germany
[2] Institute for Informatics, Clausthal University of Technology, Albrecht-von-Groddeck-Straße 7, 38678 Clausthal-Zellerfeld, Germany
[3] Faculty of Informatics, Automotive Software Engineering, Technische Universität Chemnitz, Straße der Nationen 62, 09111 Chemnitz, Germany; Jyothsna.Adamala@s2017.tu-chemnitz.de
[4] Faculty of Aerospace Engineering,"Politehnica" University of Bucharest, Str. Gh. Polizu No. 1, 1st District, 010737 Bucharest, Romania; Teodor.Salomea@euroavia-bucuresti.ro
* Correspondence: Oliver.Ohneiser@DLR.de; Tel.: +49-531-295-2566

Abstract: Assistant based speech recognition (ABSR) prototypes for air traffic controllers have demonstrated to reduce controller workload and aircraft flight times as a result. However, two aspects of ABSR could enhance benefits, i.e., (1) the predicted controller commands that speech recognition engines use can be more accurate, and (2) the confirmation process of ABSR recognition output, such as callsigns, command types, and values by the controller, can be less intrusive. Both tasks can be supported by unobtrusive eye- and mouse-tracking when using operators' gaze and interaction data. First, probabilities for predicted commands should consider controllers' visual focus on the situation data display. Controllers will more likely give commands to aircraft that they focus on or where there was a mouse interaction on the display. Furthermore, they will more likely give certain command types depending on the characteristics of multiple aircraft being scanned. Second, it can be determined via eye-tracking instead of additional mouse clicks if the displayed ABSR output has been checked by the controller and remains uncorrected for a certain amount of time. Then, the output is assumed to be correct and is usable by other air traffic control systems, e.g., short-term conflict alert. If the ABSR output remains unchecked, an attention guidance functionality triggers different escalation levels to display visual cues. In a one-shot experimental case study with two controllers for the two implemented techniques, (1) command prediction probabilities improved by a factor of four, (2) prediction error rates based on an accuracy metric for three most-probable aircraft decreased by a factor of 25 when combining eye- and mouse-tracking data, and (3) visual confirmation of ABSR output promises to be an alternative for manual confirmation.

Keywords: air traffic controller; human machine interaction; multimodality; eye-tracking; mouse-tracking; automatic speech recognition; controller command prediction; attention guidance

Citation: Ohneiser, O.; Adamala, J.; Salomea, I.-T. Integrating Eye- and Mouse-Tracking with Assistant Based Speech Recognition for Interaction at Controller Working Positions. *Aerospace* **2021**, *8*, 245. https://doi.org/10.3390/aerospace8090245

Academic Editor: Alexei Sharpanskykh

Received: 19 July 2021
Accepted: 1 September 2021
Published: 3 September 2021

1. Introduction

One central task of air traffic controllers (ATCos) is to issue verbal commands to aircraft pilots via radiotelephony in order to enable a safe, orderly, and expeditious flow of air traffic [1,2]. Usually, ATCos also need to enter this recently instructed command information into an electronic air traffic control (ATC) system such as aircraft radar labels or flight strips. This documentation supports ATCo hearbacks, i.e., to compare pilot's readbacks with ATCo instructions [3] and helps to monitor the aircraft status regarding the issued command characteristics.

If ATCo commands are issued via controller pilot data link communications (CPDLC)—being more common for non-time-critical commands in en-route phase—the content au-

tomatically feeds the ATC system and is uplinked to the aircraft pilot in order to be acknowledged. However, the traditional verbal way of ATCo-pilot communication that is assumed to remain in the medium-term future especially in highly dynamic and time-critical approach domain induces additional workload for the ATCo. This is because the ATCo needs to express the same information content twice—verbally for pilots via radiotelephony using standard phraseology according to ICAO (International Civil Aviation Organization) specifications [4] and manually for the ATC system.

Thus, automatically extracting the relevant command parts of verbal clearances to feed the electronic ATC systems without intense ATCo effort became a highly relevant technological topic in ATC. As a first step, automatic speech recognition (ASR) helps to provide the uttered words of ATC communication in written form. In addition, automatic command extraction from ATC utterances is also needed to understand the meaning of written word sequences. This language understanding task [5] can be heavily supported by using context knowledge about airspace situation, aircraft information, weather, etc. as provided through command predictions by an assistant system and used by an ASR engine.

Such assistant based speech recognition (ABSR) systems have proven to be a lightweight and easy-to-use technology to fulfill the task of ATC command recognition [6]. ABSR systems have also shown to improve air traffic management (ATM) efficiency and save aircraft fuel as ATCos can better guide air traffic with reduced workload [6]. However, ABSR command predictions have varying levels of accuracy, e.g., depending on individual ATCo habits and situations. Thus, it would be beneficial to know what part of the overall situation the ATCo currently processes—cognitively or manually.

Current prototypic ABSR implementations for ATC approach require a manual confirmation of ABSR output or a correction of recognized values, respectively [6]. Confirmation clicks via mouse are even needed if the ABSR system has low error rates [6]. Therefore, ATCos in ABSR studies are open to automatically accept ABSR output after a threshold time. However, this would also mean that sometimes unchecked and potentially erroneous ABSR output would also get automatically accepted.

Benefits of multimodal and more natural interaction at a controller working position (CWP) have already been investigated, i.e., to combine interaction technologies such as speech recognition and eye-tracking with each other to support ATCo tasks [7]. Hence, integrating further unobtrusive sensor data from eye- and mouse-tracking with ABSR and reasonably using these modalities' benefits promises to further improve efficiency of ATCos' CWP interaction.

The four derived research objectives are to (1) collect eye and mouse movement data of ATCos while monitoring radar traffic and prepare raw data for further applications, (2) extract relevant information from aforementioned interaction modalities and develop a framework to integrate the interaction data into an existing ABSR system to improve the overall performance, (3) develop and implement a method to calculate probabilities for predicted ATCo commands based on aircraft level and evaluate their quality, and (4) develop a CWP system to enable unobtrusive (visual) ABSR output confirmation and evaluate its usefulness.

Operator interaction data from eye- and mouse-tracking can support two important steps of ABSR applications as will be shown in this paper: (1) predict more accurate ATCo commands in order to reduce command recognition error rates, (2) check implicit ATCo confirmation of presented ABSR output or escalate attention guidance mechanisms to enforce ABSR output check. These two conceptual enhancements have been implemented, tested, and evaluated. The one-shot experimental case study with two controllers in a human-in-the-loop simulation of an ATC approach scenario at DLR Braunschweig in May 2021 revealed promising results—even if not significant due to the limited number of study subjects—to further refine the integrated use of interaction data: (1) command predictions on aircraft callsign level got more accurate by a factor of four, (2) combination of eye- and mouse-tracking metrics was superior over single modality metrics with an

improvement factor of 25 for prediction error rates, and (3) ABSR output confirmation by ATCos worked feasibly just by using gaze information.

Section 2 outlines related work on eye- and mouse-tracking as well as speech recognition and combinations of modalities relevant for ATC systems. Both, the baseline CWP and our CWP prototype with integrated eye- and mouse-tracking for ABSR output confirmation are described in Section 3. Section 4 explains the concept of assigning individual probabilities to command predictions based on ATCo interaction data. The study setup, methods, and subject data are explained in Section 5. The results of the study as sketched above are presented and discussed per conceptual enhancement in Section 6. Section 7 concludes and discusses the results more generally. Finally, Section 8 outlines future work.

2. Related Work on Speech Recognition, Eye-Tracking, and Mouse-Tracking

The following subsections give evidence to the use and benefits of speech recognition, eye-tracking, and mouse-tracking prototypes and applications as well as analyzes how the modalities can be used together and benefit from each other, respectively.

2.1. Related Work on Automatic Speech Recognition (ASR)

ASR means to convert speech, i.e., audio signals, into a sequence of words, commonly referred to as transcription. This transcription contains all uttered words and has special transcription rules for spelled letters, truncated and non-understandable words, human noise, and different versions of English or even non-English words [8]. The next important step is the language understanding, i.e., to transform the sequence of words into machine-readable semantic meaning, commonly referred to as annotation.

Speech recognition found its way into daily life as Amazon Alexa, Apple's Siri®, Google Assistant, or Microsoft's Cortana show. ASR activities in ATC [9] and using contextual knowledge to improve ASR began decades ago [10]. The mandatory use of ICAO standard phraseology, which limits the number of words and structures, helps to analyze verbal ATC communication [4]. However, transcription and especially annotation is more complex, because ATC radiotelephony users often deviate from the phraseology. Many European air navigation service providers and air traffic management system providers agreed on an ontology for annotating ATC utterances in a consortium led by DLR to enable better interoperability [11]. This ontology dramatically eases semantic interpretation especially when ATCos or pilots deviate from standard phraseology.

Assistant based speech recognition (ABSR) has proven to be a good approach [12] to achieve low ATC command recognition error rates [6]. In ABSR systems, ASR engines are supported by hypotheses about the next ATC commands, so called ATCo command prediction, that reduce the ASR engine's search space [13]. With this technology, command recognition error rates of below 2% are possible [14]. The command annotations can be used for further applications such as radar label maintenance to reduce ATCo workload [13], workload assessment [15], safety nets [16,17], arrival management planning input [18,19], or ATC simulation and training support [20,21]. The most advanced command prediction techniques base on machine learning and cover all relevant flight phases in the approach, en-route, and tower environment [22–24]. The command prediction error rate of an early implementation for multiple remote tower simulation command predictions was below 10% [25]. An ATC command prediction error rate of even 0.3% has been achieved for simulated Prague approach environment [26].

Another relevant metric is the portion of predicted commands, i.e., the number of predicted commands divided by the total number of commands per aircraft callsign, that an ATCo could theoretically issue. The lower the portion of predicted commands, the less alternatives that an ASR engine needs to choose from. For example, 144 heading commands are modeled as being usually possible with the qualifiers *RIGHT* and *LEFT* for the value range from 005, 010 to 355, 360. For the multiple remote tower environment, a context portion predicted of below 10% was achieved [25].

Currently, besides some statistical approaches, actually issued ATC commands were either predicted or were not predicted at all by an ABSR system, i.e., for comparison reasons we assume that predictions have a probability of one divided by the number of all predicted commands (uniform probability) or of zero. However, information about the certainty of different words and commands can support the ASR engine to choose the correct words [27,28].

2.2. Related Work on Eye-Tracking

Eye-tracking is a technology based on sensors to determine a human's gaze point and gaze movements as well as pupil size [29,30]. Most modern eye-trackers emit near-infrared light that is reflected by the eye's pupil and cornea [31]. These reflections can be measured with an infrared camera to derive the human's gaze points and further eye-tracking metrics [32]. Such eye tracking techniques do not distract the people involved because infrared light is invisible to the human eye.

Eye-tracking devices can be mounted on the head or can be worn as glasses with the advantage of free movement for the human user, but with the disadvantage of being more intrusive on the human's body [33]. Other eye-trackers can solely be mounted on a monitor. However, this leads to a restricted range of gaze detection. In a calibration process, the pupils' and corneas' reflection are matched with the screen coordinates that the human would be focusing on.

A number of metrics regarding eye-tracking have been established for further interaction analysis. A gaze point is a single point of gaze measurement that is often recorded with 50–60 Hz. A fixation is a cluster of subsequent gaze points defined through spatial thresholds and timely dwell times, such as 200–300 ms. There are many different algorithms for eye-tracking fixation identification based on spatial and temporal information [34]. Fixations indicate well the human's visual attention [31]. Given the fixation, the dwell time—hereinafter referred to as fixation duration—can also be measured [35]. The rapid eye movement segments between fixations are called saccades. The sequence of fixations and saccades is called scan path and is important to estimate user behavior in analyzing screen content [36]. Analyzing such scan pattern can help to train highly specialist screen users such as ATCos [37,38].

For the purpose of gaze analysis, certain spots of a screen are defined as areas of interest (AoI). An AoI is defined as "physical location, where specific task-related information can be found" [39]. The time spent on an AoI as a sum of fixations can be used to derive the human's attention or situational awareness in a broader view. This data is often presented as colored heat maps of human's gaze points on screen [40].

Eye-tracking is already widely used to analyze human's behavior on websites, e.g., using fixation count and fixation duration to predict customer interest and choices [41,42]. The time-to-first-fixation of an AoI was found to not support customer intention prediction [41].

In another study about eye-tracking based intent prediction with a support vector machine, a customer request prediction accuracy above 75% was achieved almost 2 s before the customer request towards a worker for an ingredient was uttered verbally [43]. Again, the fixation count and fixation duration (initial and in total) were considered. Furthermore, the fixation time was analyzed, i.e., how recent did the fixation happen on an AoI. Support vector machines using visual attention data have also been used successfully to predict human behavior in problem-solving tasks [44]. Hence, eye-tracking data can enable benefits in online applications, but also with offline analysis after recording [45].

Different research prototypes incorporating eye-tracking have already been developed for ATC [46–49]. Eye-tracking data assist to guide human ATC operators' attention via visual cues based on the desired and actual area of attention [50–52]. A combination of eye-tracking and electroencephalography was even used to control vigilance and attention of ATCos [53]. One important advantage of eye-tracking methods for ATCos is the potential to relieve them from tasks that would otherwise have to be done by hand [54].

2.3. Related Work on Mouse-Tracking

Mouse-tracking is a cheap and simple hardware-based method to acquire information that can be translated into visual attention later on. Human computer users can move a mouse to position a cursor on screen, can perform clicks with left and right mouse button, and scroll with a mouse wheel if applicable. The main mouse functions are metaphors of humans pointing to things (cursor) or touching things (selection of screen items with clicks) with their fingers or hands. Hence, mouse usage generates a variety of input data for the computer when users select text, hover over icons, or click to start events. Furthermore, this kind of tracking is unintrusive [55].

Mouse-tracking data for user intent prediction can be captured with a relatively low rate of 10 Hz [56]. Mouse cursor trajectories support understanding human decision processes [57,58]. Mouse movement paths seem to be more important than speed and acceleration of mouse movements in order to anticipate user decisions similar to the scan path in eye-tracking [59]. The cognitive processes related to eye- and mouse-tracking are similar as it is assumed in both cases to indicate visual attention [60]. Humans tend to use the mouse cursor for examining screen content, e.g., text reading and highlighting as well as interaction with screen content, but they may also ignore the mouse if it does not seem to be useful [61,62]. When clicking with the mouse, humans follow the mouse cursor even more visually compared to just move the mouse [56]. In more than two-thirds of the cases, the human watches the mouse cursor region on screen after a mouse saccade [63]. In more than 80% of the cases, if screen areas are examined visually, they are also examined with the mouse. Similarly, if they are not examined visually, they are also ignored with the mouse [63].

2.4. Multimodal Integration of Different Modalities Related to Human-Machine Interaction

Different approaches combine multiple interaction modalities to be used either independently of each other or to combine the advantages of them.

Eye-tracking can be used to re-assign probabilities of speech recognition hypotheses or to adapt the language model, respectively, by considering human's visual attention leading to significant decrease in word error rate [64]. However, achieved better recognition accuracy with such technique was connected more to the visual field than to the visual focus [65]. Eye-tracking and other non-verbal modalities have been combined to make speech recognition more robust against noise [66]. Eye-tracking was also found to be complementary to speech recognition for affect recognition in a gaming environment's multimodal interface [67] and for tracking reading progress [68].

The multimodal CWP prototype "TriControl" combines speech recognition, eye-tracking, and multi-touch sensing to issue ATCo commands [69]. The three main parts of an ATC command—callsign, command type, and command value—are entered into the ATC system via three different modalities, i.e., by looking at an aircraft radar label for the callsign, performing defined multi-touch gestures for the command type, and by uttering only the command value [70]. These three command parts are put together, confirmed, and sent to the aircraft via data link or electronically read, e.g., by looking at aircraft callsign "SAS818", swiping down for command type "DESCEND", and uttering "four thousand" for a command value of 4000 ft [71]. The possibility to work with different modalities in parallel enables faster and more intuitive interaction especially for approach ATCos [7].

Human-machine interfaces (HMI) that offer multiple modalities are called multimodal HMIs [72–75]. Multimodal HMIs can have several advantages such as robustness [76,77], quick, safe, and reliable use [78,79], individualized use [80], natural and intuitive interaction [81,82], workload reduction [83], and adaptation for certain human needs in environments like system control [84]. Human HMI users often change between multimodal and unimodal use [85,86]. Some tend to prefer multimodal interaction if well-designed [76], others prefer unimodal interaction especially in phases of low cognitive workload [87]. An example HMI for cars also offers speech, gaze, and gestures for system input [88].

Examples of multimodal research prototypes in ATC, e.g., combine gestures with speech recognition [89] or eye-tracking [90]. Additionally, in SESAR (Single European Sky ATM Research Programme) speech recognition and eye-tracking for attention guidance have been investigated and were found to be important future CWP technologies [91,92].

3. Description of Controller Working Position Prototype with Integrated Eye- and Mouse-Tracking for ABSR Output Confirmation

3.1. Description of the Baseline Controller Working Position (Mouse-Click Trigger)

ATCos will be using the same basic CWP setup to evaluate the baseline and our solution system. The baseline includes the common interaction method with using symbols to be clicked in the aircraft radar label. The newly implemented solution system works by just looking or mouse-hovering at the aircraft radar label to start the ABSR output confirmation process. Hence, the majority of ATCos' tasks are the same in baseline and solution run as detailed in Section 5.2. ATCos have to monitor air traffic in approach phase with the given situation data display (see Figure 1).

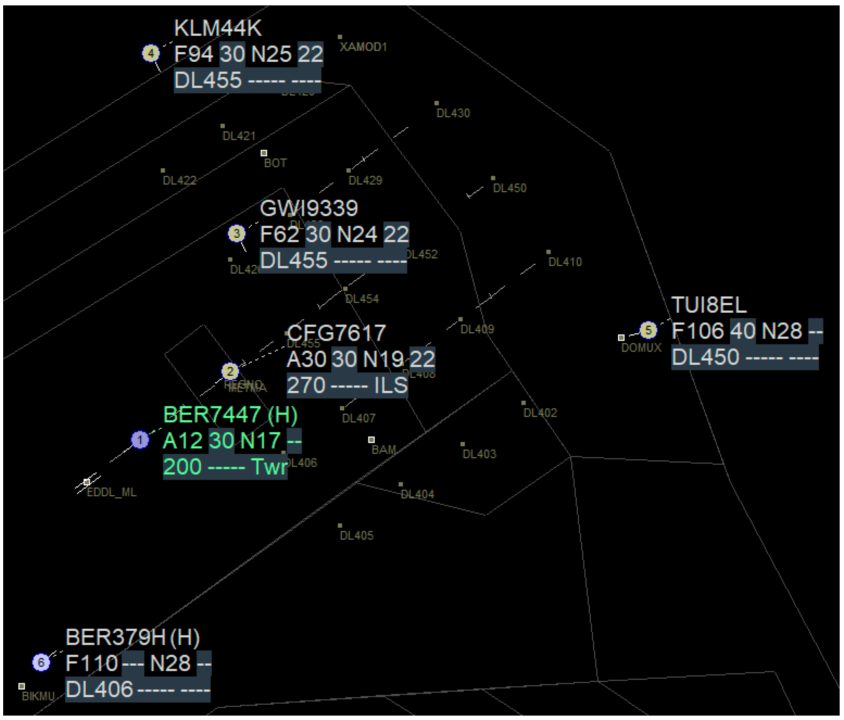


Figure 1. Aircraft radar labels next to aircraft circle icons (containing sequence numbers) flying within Düsseldorf approach airspace shown on DLR's radar display RadarVision [93]. The five shaded label cells in the second and third label lines may depict the last ATCo command value for a certain command type (altitude, speed, direction, rate of altitude change, miscellaneous).

The first label line in any of the labels in Figure 1 indicates the callsign and the weight category in brackets. "medium" is the default weight class category. The second line shows (1) flight level (first letter is "F") or altitude in hundreds of feet (first letter "A"), (2) the last given or recognized altitude command, (3) the speed in tens of knots ("N"), and (4) the last given or recognized speed command. The third line displays last issued heading/waypoint ("270"/"DL455") clearances, rate of climb/descent with an arrow if applicable, and any other miscellaneous recently given command content such as an ILS-clearance ("ILS") or handover to tower ("Twr"). The label example in Figure 2 also shows an optional fourth label line activated by mouse-over function with current heading ("053") and aircraft type ("A319").

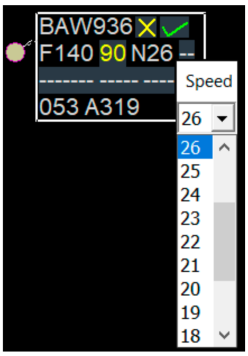


Figure 2. Baseline aircraft radar label with white frame and yellow ABSR output value expecting manual ATCo confirmation through mouse click on green check mark (or rejection on yellow cross) and drop-down menu to change misrecognized or not recognized speed value.

Based on the air traffic situation and the ATCos' situational awareness, ATCos issue commands to aircraft pilots. The primary way to issue commands shall be the acoustic modality, i.e., to press a foot switch (push-to-talk), utter commands/clearances, and release the foot switch again. The recorded verbal utterance is analyzed in the speech recognition process by the ABSR system. The ABSR output is presented as yellow value in one of the five shaded aircraft radar label cells (see yellow flight level "90" in Figure 2). Clicking on one of the five shaded cells will open a drop-down menu to enable manual correction of the ABSR output. The first line of the aircraft radar label also shows a green check mark and a yellow cross to completely accept or reject all shown ABSR output for this aircraft, respectively. The former should ultimately be clicked if all ABSR output shown in the label is correct. All label values will then turn into white. Hence, the ABSR output confirmation by ATCos is triggered by mouse-clicks. In earlier trials with the same configuration, ATCos complained about the need to always click on the check mark given the high command recognition rate of the ABSR system. Furthermore, they need to move the mouse cursor—and thus also their gaze—to a less important area in the corner of the aircraft radar label. This causes additional manual and cognitive workloads. ATCos would rather just see the highlighted ABSR output that enters the ATC system directly if there is no ATCo intervention in a certain amount of time.

3.2. Description of the Solution Controller Working Position (Attention Trigger)

Based on the aforementioned ATCo recommendation, we modified the concept of ABSR output confirmation [94]. However, as a safety net, we still want to check if the ATCo at least noticed the ABSR output and did not intervene in a certain amount of time.

Thus, to avoid manual workload for ABSR output confirmation, the visual attention shall be used as a trigger in the confirmation process without the need for mouse clicks. One pre-assumption is that the ATCo has his/her visual attention at the spot he/she is looking at. This might not always be true, e.g., in case of staring at a certain position without presuming anything. However, this is a valid approximation to support ATCos in a visual task [50]. An infrared eye-tracker mounted on the bottom of the situation data display continuously records the ATCos' gaze points. The software module *ModEyeGaze* tries to match these gaze points with relevant objects displayed on the screen. These objects can be aircraft icons, aircraft labels, and airspace points.

The accuracy of eye-tracking is not of utmost importance, i.e., an accuracy of pixels is not required as it is not important to determine if the ATCo is looking at the speed or the altitude field in a label. An accuracy of roughly less than 1 cm is feasible to match the gaze points with displayed objects such as aircraft radar labels given a further visual threshold. Furthermore, a dwell time is defined in order to calculate a fixation on a displayed object. This avoids too many fixations in case the ATCo is just quickly shifting his/her view to the other side of the display. Like in the baseline system, yellow ABSR output values will

appear in the aircraft radar label immediately after the speech recognition process ends (see yellow values in Figure 3).

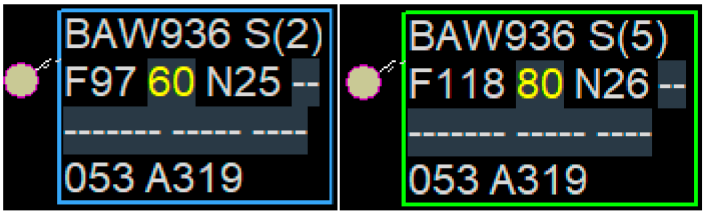


Figure 3. Solution aircraft radar labels with yellow ABSR output expecting attention-based ATCo confirmation and colored label frames in different states; left: light blue frame in saliency level "2" as visual check gaze for ABSR output is pending, right: green frame in saliency level "5" as visual check gaze has confirmed and time for potential manual ASBR output correction is running.

Peripheral cues are used to guide the operator's attention [95]. More precise, different saliency levels of labels are applied depending on the visual check status by the ATCo to smoothly guide the ATCos' attention to the relevant spots. All aircraft labels are in the default saliency level *transparent* ("−1") initially. As soon as yellow ABSR output appears in a label, eye-tracking data analysis will be activated. The layout is as shown in Figure 2 of baseline system, but without the cross and check mark. The saliency level of the label will be escalated further every 5 s if *ModEyeGaze* does not detect an ATCo fixation on a highlighted aircraft radar label.

The label status is switched to saliency level *white* ("0"), i.e., a white label frame will be drawn. Saliency level *yellow* ("1") with a yellow label frame is activated 5 s after the start of saliency level *white* to get the ATCo's attention. Accordingly, saliency levels *light blue* ("2") (see left label of Figure 3) followed by *dark blue* ("3") are activated later after a gap of 5 s each. Thus, if there was no visual scan of the ABSR output (aircraft radar label) for 25 s after the appearance of the ABSR output value in yellow, the ABSR output will be rejected (*saliency level 4*) and does not enter the ATC system. The label's saliency level will revert to *transparent* ("−1") afterwards.

If *ModEyeGaze* detects an ATCo fixation on an aircraft radar label that has at least one unchecked yellow ABSR output value independent of the current saliency level, saliency level *green* ("5") will be activated, i.e., a green label frame (see right label of Figure 3) will remain until the end of the maximum time for optional correction (10 s). If the correction time has passed, all visible yellow values in the aircraft radar label will enter the ATC system and the label will revert to saliency level *transparent* ("−1") with all label values displayed in white color.

Eye-tracking as a technology might be more error-prone than manual system operator input especially if ATCos heavily move around with body and head compared to the calibration seating position. Therefore, mouse interaction data with the situation data display is used as a backup. The frequency of mouse usage by the ATCos depends on the CWP interaction design. However, as this data is just used as a backup data input, it is of less importance if the mouse is really used. Accordingly, if the mouse cursor is moved on an aircraft radar label that currently displays yellow ABSR output values and the mouse-over time exceeds a certain threshold time, this is determined as a match as if the ATCo would have looked at the label. Hence, the label frame turns green and counts down the remaining time for optional ABSR output value correction.

As system operators often carry their gaze, i.e., their visual attention, along with the mouse cursor, the gaze- or mouse-over initiated check of the solution system is called "attention triggered".

4. Description of Command Prediction Rescoring with Integrated Eye- and Mouse-Tracking

The second use case for operator gaze and interaction data is the enhancement of ATCo command prediction quality [96]. The implemented algorithm will be tested on the baseline run (Section 5.1), but also works if the ABSR output confirmation is used as in the solution system explained in Section 5.2. DLR's command hypotheses generator predicts ATCo commands for the speech recognition engine for given timeticks as shown below in Table 1.

Table 1. Examples for controller command predictions in ontology format with higher probability for aircraft that recently received ATCo attention.

Aircraft Callsign	Command Type	Second Type	Command Value	Unit	Qualifier	Uniform Probability	Re-Assigned Probability
AFR641P	HEADING		260		RIGHT	0.1	0.02
AFR641P	CLEARED	ILS	RW23R			0.1	0.02
AFR641P	DESCEND		4000	ft		0.1	0.02
BAW936	TRANSITION		DOMUX 23			0.1	0.06
DLH5MA	DESCEND		80	FL		0.1	0.23
DLH5MA	REDUCE		200	kt		0.1	0.23
DLH5MA	INFORMATION	QNH	1013			0.1	0.23
KLM1853	CONTACT		TOWER			0.1	0.03
KLM1853	CONTACT_ FREQUENCY		118.300			0.1	0.03
UAE57	DIRECT_TO		DL455		none	0.1	0.13

In Table 1's example, five different aircraft callsigns are predicted to possibly receive an ATCo command in the near future. For those callsigns different command types and values are reasonable due to their current airspace position and current motion characteristics. Hence, the number of predicted commands per aircraft can vary. In the basic ABSR implementation, no probability values are used, i.e., all predicted commands (here: 10 different ones) are assumed to have the same probability $P(cmd)_u$ (here 0.1). The basic advantage of this command prediction for the speech recognition engine is to know beforehand about commands that may be uttered (e.g., "AFR641P DESCEND 4000 ft") and to know, which will probably not be uttered (e.g., "KLM1853 DESCEND 4000 ft"). However, there might exist further data that even state which of the predicted commands are more likely to be uttered than others, i.e., to re-assign probabilities for command predictions with higher weightings for some aircraft commands (exemplarily underlined in column "Re-assigned Probability" with $P(cmd)_{ra}$ of Table 1). From an implementation point of view, the term *assignment* is more correct than *re-assignment*. However, the latter term better emphasizes to compare individualized probabilities against uniform probabilities for command predictions as outlined above.

It is important to note that the re-assignment does not intend to further predict yet unpredicted commands or to delete some predicted commands. Hence, as in the basic implementation, it can still happen that the ATCo issued a command to aircraft callsign "DAL27V", which is not a predicted aircraft callsign in the example of Table 1.

The basic pre-assumption is again: "the visual attention is where the ATCo looks at". However, some derived assumptions need to be made for this concept, i.e., display spots—including aircraft—that get more attention from the ATCo than others will more likely be involved in very near-term future ATC commands that the ATCo will issue. We assume that an ATCo will more likely give a command to an aircraft that he/she currently looks at or recently looked at—maybe even a multiple of times—as compared to an aircraft that was never looked at in the recent past by the ATCo, as determined by eye-tracking and

ModEyeGaze. In Table 1's example, we assume that DLH5MA and UAE57 have recently been looked at. Thus, predicted commands that include these aircraft callsigns receive probabilities above the "uniform" probability average for all commands. This implies that the probabilities for all the other aircraft needs to be reduced and re-assigned (AFR641P, BAW936, KLM1853).

Mouse interaction is again used as backup sensor data, i.e., if the ATCo moved the mouse and rested over an aircraft radar label recently or clicked very close by, this is considered to be similar to the visual attention via eye-tracking. For all interaction data stored in a data base, i.e., the combination of eye-tracking recorded with 60 Hz and mouse-interaction data recorded with 10 Hz (except the mouse clicks), different ratios will be tested. The most recent data from the last five to ten seconds for eye-tracking and the most recent data from the last three seconds for mouse-tracking is used in our concept due to expert feedback and initial feasibility testing. Three parameters of the recent past seconds will be considered for re-calculating probabilities: gaze duration on aircraft, gaze counts on aircraft, and mouse movements related to aircraft shown on a radar display.

4.1. Command Probability Calculation Based on ATCo Interaction Data (Aircraft Level)

The calculation of probabilities for command predictions with respect to different aircraft based on ATCo interaction data will be explained in the following. The total command probability $P(cmd)$ for a single command can be calculated with individual weightages W for each of the three interaction data metrics that sum up to one:

$$P(cmd) = W_{ETfix_{dur}} \cdot P(cmd)_{ETfix_{dur}} + W_{ETfix_{cnt}} \cdot P(cmd)_{ETfix_{cnt}} + W_{MTint} \cdot P(cmd)_{MTint.} \quad (1)$$

These metrics are called eye-tracking gaze fixation duration ($ETfix_{dur}$), eye-tracking gaze fixation count ($ETfix_{cnt}$), as well as mouse interaction data ($MTint$) and will be explained in Sections 4.2 and 4.3.

4.2. Command Probability Calculation Based on Eye-Tracker Data (Aircraft Level)

The total probability of an aircraft receiving an ATC command in the near future should be extremely high in case the ATCo looked at this aircraft for a long amount of time in the recent past. This mathematical weightage can be best expressed with an exponential function instead of a linear function. Thus, the re-calculation of probability P per command (cmd) for a concrete aircraft (A/C_k) based on eye-tracking gaze fixation duration ($ETfix_{dur}$) is given by:

$$P\left(cmd_{A/C_k}\right)_{ETfix_{dur}} = \frac{e^{dur}{}_{A/C_k}}{\sum_{i=1}^{\#A/C}\left(\#cmd_{A/C_i}\, e^{dur}{}_{A/C_i}\right)}. \quad (2)$$

The parameter dur is the time spent on an aircraft during the last five seconds, $\#cmd_{A/Ci}$ represents the number of predicted commands per aircraft with all aircraft from iterator start i = 1 to the number of considered aircraft ($\#A/C$) being summed up.

The eye-tracking gaze fixation count ($ETfix_{cnt}$) in Equation (3) is considered in a linear way as the number of fixations on an aircraft is not assumed to be as an extreme indicator as the duration for an aircraft to receive the next ATC command. It is calculated with the following equation where cnt is the number of fixations for the specific aircraft in the last ten seconds:

$$P\left(cmd_{A/C_k}\right)_{ETfix_{cnt}} = \frac{cnt}{\sum_{i=1}^{\#A/C}\left(\#cmd_{A/C_i}\, cnt\right)}. \quad (3)$$

Both eye-tracking probabilities (ET) can be combined to a single probability with an appropriate weight.

4.3. Command Probability Calculation Based on Mouse-Tracker Data and Combination of Interaction Data (Aircraft Level)

Mouse-tracking (MT) data are considered by Euclidian distance between the position of closest aircraft radar icon and position of mouse cursor/click. This closest aircraft

influences the mouse interaction weighting score *miw* to be (a) 5 if the aircraft has been visited with the mouse cursor for at least 300 ms or (b) 10 if the ATCo left/right clicked close to this aircraft as a sign of more active interaction with the aircraft's characteristics. The command probability based on mouse interaction data (*MTint*) in Equation (4) is only considered for an aircraft (*A/C*) if *miw* is greater than zero, i.e., if any mouse interaction close to the analyzed aircraft has taken place:

$$P\left(cmd_{A/C_k}\right)_{MTint} = \frac{e^{miw}{}_{A/C_k}}{\sum_{i=1}^{\#A/C}\left(\#cmd_{A/C_i}\, e^{miw}{}_{A/C_i}\right)}. \quad (4)$$

Inactive mouse interaction can result from the CWP design or from individual preferences of the ATCo. Unlike ET, positions of aircraft radar labels are not considered for MT as labels may overlap and may be moved away just for readability even if the labels are far away from aircraft icons and contain relevant information why the ATCo looks there.

4.4. Air Traffic Situation Dependent Command Probability Combined with Interaction Data (Command Type Level)

We further assume that scanning different aircraft in the recent past leads to dedicated command types if some of the scanned aircraft have certain characteristics. For example, if the ATCo scans an aircraft close to the runway, the likelihood of a *CONTACT* command to the tower increases. If the ATCo fixes the gaze on a certain waypoint and on an aircraft for which this waypoint has been predicted as a command value, the likelihood for a *DIRECT_TO* command to this waypoint increases. Furthermore, if an approach ATCo scans two or more aircraft at similar altitudes, the likelihood of commands from the categories of altitude change commands, direction change commands, or speed change commands can be adjusted as shown in Figure 4 based on ATCo feedback. For example, if scanned aircraft in similar altitudes have converging headings and are in close proximity, altitude change commands would be re-assigned with higher probabilities than heading change commands and especially than speed change commands. If these aircraft are not in close proximity, the speed difference might decide about prioritizing heading or speed change commands. Individual air traffic situations require individual decisions about ATC commands as well as individual conflict detection and resolution strategies [97], but slightly different probabilities on command type level can help to predict commands better on average.

If in Table 1's example DLH5MA was recently scanned, having the same altitude and intersecting path with another aircraft, the *DESCEND* command might be re-assigned with higher probability, e.g., 0.39 as compared to 0.15 for each of the *REDUCE* and *INFORMATION QNH* commands.

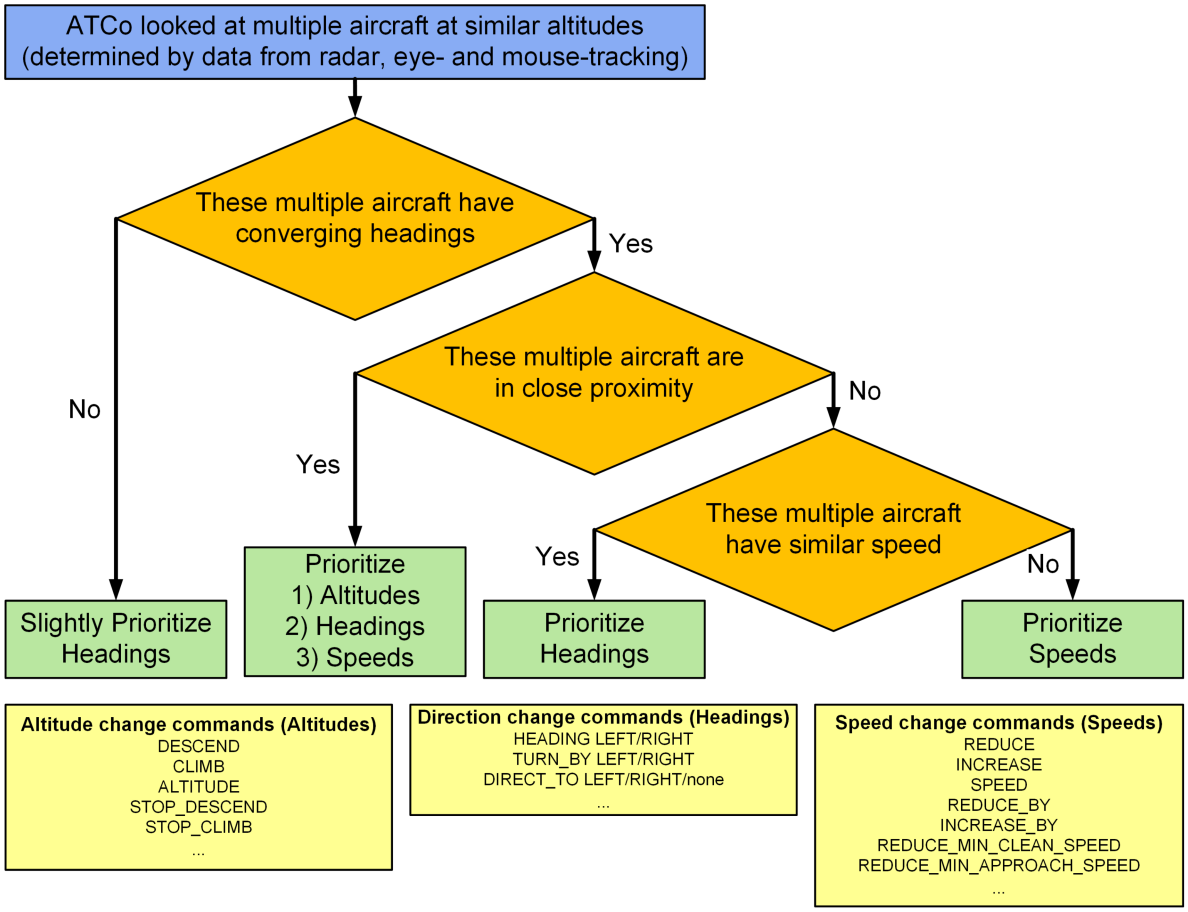


Figure 4. Flow chart to determine priorities for ATC command types based on aircraft scanned by ATCos.

5. One-Shot Experimental Case Study with Controllers in Simulation Environment

For a quantitative and qualitative evaluation on how DLR's ABSR application benefits from the use of eye- and mouse-tracking interaction data, relevant data from the simulation trials of a one-shot experimental case study was recorded in log files and data bases. This data comprises of:

- Positions of aircraft icons and aircraft radar labels with their states as shown on the situation data display
- Verbal utterances with automatic transcriptions, annotations, and instruction methods
- Eye gaze data with timeticks and fixation positions/durations
- Mouse interaction data with timeticks, click positions, and movements
- Answers of online questionnaires

5.1. Study Setup and Schedule for Evaluation of Eye- and Mouse-Tracking Support for Speech Recognition

In May 2021 we conducted an early interaction study at DLR Braunschweig with two controllers living close by—as COVID-19 restrictions prohibited trials with international ATCos. Hence, there was no scientific sampling and recruitment process. The study subjects were both male, roughly at the same age, wore a face mask (due to Covid-19 hygienic protocol), and spoke English with a German accent being relevant for speech recognition. Furthermore, both subjects wore glasses which is relevant for eye-tracking. One of the participants was an active licensed ATCo for tower and approach and the other participant was a former ATCo trainee for Düsseldorf approach area. Both subjects were not involved in the research activities and received the main part of the study information

only in the briefing session. The complete hardware setup of the prototypic CWP can be seen in Figure 5.

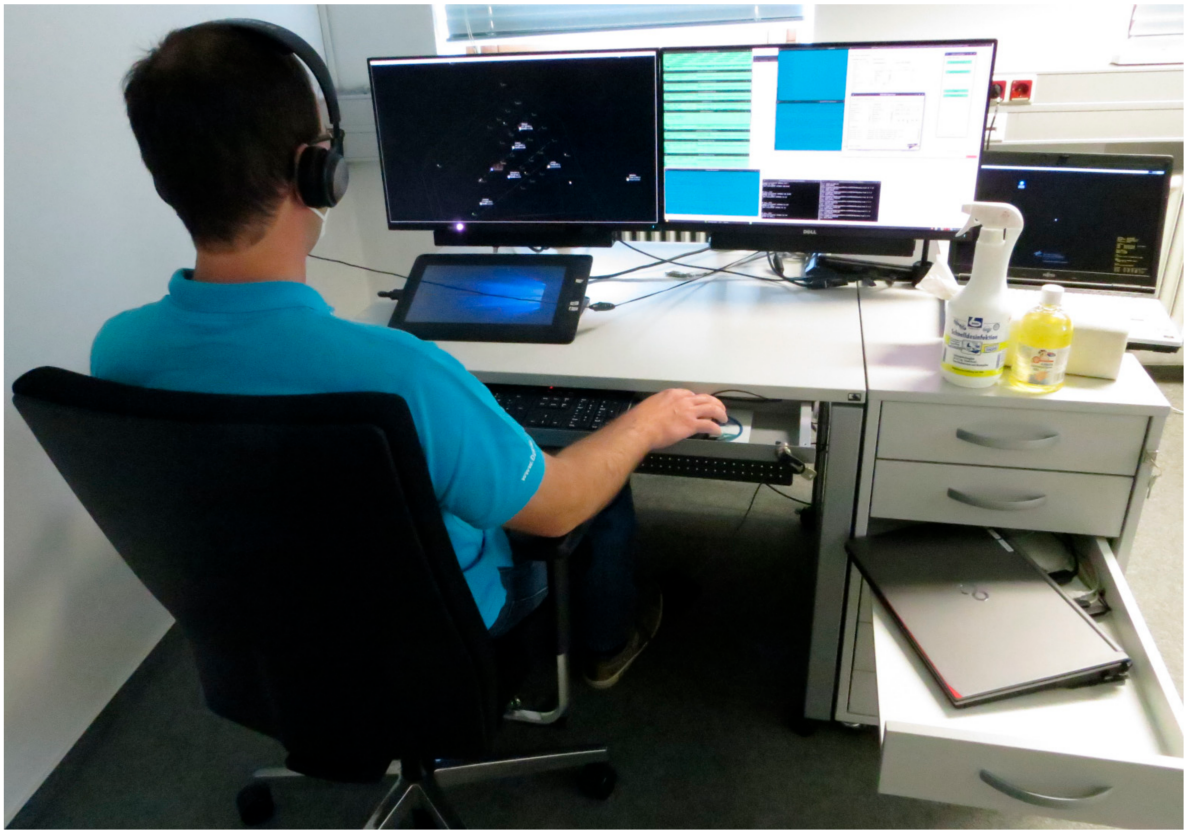


Figure 5. Study participant during simulation trials using an eye-tracking supported attention guidance system for assistant based speech recognition.

The subject used a foot switch to enable and disable voice recording (push-to-talk). The voice itself was recorded via the headset. The mouse placed to the right of the keyboard could be used to manually correct ABSR output or give commands via mouse. The leftmost monitor shows the situation data display with aircraft radar data in Düsseldorf approach airspace. The eye-tracker is mounted onto the bottom of this monitor. All other devices were not relevant for the subject's work during the scenario, but to run the simulation. The right monitor presents software module output of the arrival manager, the speech recognition engine, and the air traffic simulator running on the two Linux laptops on the right side of the photograph. The situation data display and the eye-tracking system runs on a Windows laptop (hardly visible below the right monitor). The disinfection material placed on the desk was used before a new operator started working on the CWP prototype to fulfill the hygienic protocol.

The software setup of the human-in-the-loop simulation comprised of an air traffic scenario for Düsseldorf approach (ICAO airport code EDDL). The only active runway was 23R. The duration of the scenario was one hour and included 38 approaching aircraft without considering departures. Seven aircraft were of weight category "heavy", all others were "medium" class aircraft. The participants had to handle the traffic being a "Complete Approach" controller, i.e., combined pickup/feeder ATCo in Europe or combined feeder/final ATCo in the US, respectively. This setup was similar to the earlier AcListant® [14,18], AcListant®-Strips [13], and TriControl [7,71] trials.

The four-hour-schedule of the study started with a 30-min briefing about the tasks to perform and included an eye-tracking calibration exercise. Two training runs for baseline and solution condition with roughly 20 min each and individual short breaks between

simulation runs followed. The baseline and solution runs themselves lasted up to one hour each—conducted in alternate order for the different participants to avoid bias. During the final half an hour, participants had to fill a questionnaire as well as needed to answer open questions and give comments during a debriefing.

5.2. Subjects Tasks and Execution of Simulation Study

The ATCos' task was to issue ATC commands primarily via voice by using the push-to-talk functionality. An example would be the following transcription of words: "lufthansa five mike alfa descend flight level seven zero turn right heading three six zero". If relevant parts of this utterance are correctly recognized by the speech recognition engine, the semantic representation of the utterance as per the agreed ontology, also known as the annotations would be displayed as follows: "DLH5MA DESCEND 70 FL, DLH5MA HEADING 360 RIGHT". These commands are converted to the necessary format for the air traffic simulator which itself changes the motion of the relevant aircraft. Hence, there are no active simulation pilots during the runs (amongst other reasons due to COVID-19 restrictions). All commands recognized by ABSR will be executed by the simulator. In almost all cases, misrecognized commands have not been shown as ABSR output, because they have been invalidated beforehand as not being plausible, due to reasons such as missing a correct callsign or a command value being out of a reasonable range.

Some technical problems of the CWP system that occurred during baseline and solution runs need to be mentioned that probably also affected the rating of the tested features. There was an operating system latency of roughly one second due to a laptop docking station issue that was only found after the trials. With this, there was a slight lag for the output display to appear, i.e., the confirmation saliency level, the ABSR output or the zoomed situation data display region appeared later than expected/theoretically possible. Furthermore, some commands have not been properly forwarded to the traffic simulator, i.e., altitude commands between 4000 and 6000 feet, *DIRECT_TO*-commands, and some *ILS* clearances were affected. Nevertheless, all traffic could be handled and could be guided to land on the runway. As the flown trajectory did not matter for data analysis, but only the relevant eye- and mouse-tracking data, as well as the given ATC commands, the technical problems mentioned above should not heavily influence the basic conclusions of the simulation runs.

6. Results Regarding Effectivity of Eye- and Mouse-Tracking to Support Speech Recognition Applications

Data of two baseline and two solution runs has been recorded. Only the middle 45 min of the runs were analyzed to avoid data of a "slow start" and "scenario fading out". As Table 2 shows, ATCos issued 180 ATC commands per run on an average considering both modalities. Roughly 125 of these 180 ATC commands were recognized from slightly more than 100 speech utterances on an average, i.e., 1.3 ATC commands per speech utterance. The remaining 55 ATC commands were instructed via mouse in roughly 49 mouse issuing occasions, i.e., 1.1 ATC commands per mouse issuing occasion.

Table 2. Number (#) of actually issued ATC commands per run and command modality.

Run	# Actually Issued ATC Commands via Mouse	# Actually Issued ATC Commands via Speech	# Actually Issued ATC Commands per Run	# Speech Utterances/Mouse Issuing Occasions per Run
Baseline	88	105	193	154
Solution	22	146	168	144
All	55	125	180	149

In baseline runs, roughly 105 and 88 commands were issued via voice and mouse, respectively. The different types of issued ATC commands—by using both modalities with some misrecognitions—were *ALTITUDE* (36.4%, mainly *DESCEND*), *HEADING* (34%),

CLEARED ILS (13.6%), *SPEED* (6.6%, mainly *REDUCE*), *CONTACT* (6.5%), and others including *DIRECT_TO* (3%).

Multiple thousand gaze fixations have been determined by the eye-tracking algorithm per run. A total of 42% of those fixations were on aircraft radar labels, 23% on aircraft radar icons, and 35% on airspace waypoints. In the baseline scenario, on an average more than 6000 mouse movements, around 250 left clicks, and less than ten right clicks on the situation data display have been captured per run.

6.1. Enhancement of Probabilities for Speech Recognition Hypotheses by Eye- and Mouse-Tracking Data

This section compares the re-assigned ATCo command prediction probabilities with the uniform probabilities of the basic ABSR system implementation. The first part of the analysis concentrates on the benefits of re-assigned probabilities for different aircraft callsigns of command predictions while the second part also investigates re-assigned probabilities for different command types of single aircraft command prediction sets.

There are two basic result areas for the analysis. First, a factor showing the improvement in prediction accuracy as compared to the basic ABSR implementation, i.e., if the factor is greater than 1, the enhanced implementation outperforms the basic. Second, a four-field confusion matrix that helps to classify predicted and actually issued commands, i.e., the percentage of correct command predictions can be derived.

6.1.1. Conditions and Metrics for Evaluating Prediction Probabilities on Aircraft Callsign Level

The recorded data is analyzed (1) for three conditions of eye- and mouse-tracking metrics as well as for two combinations of them, (2) for input modalities speech, mouse, and both combined, and (3) for the four simulation runs.

As explained above, the terms baseline and solution are right for the task of non-manual ABSR output check, but may be misleading for the task of analyzing the re-assignment of command prediction probabilities. However, the display appearance was slightly different in the two runs—cross and check mark in the first aircraft radar label line were not shown for solution runs unlike in baseline runs as explained in Section 3.2. Nevertheless, data from baseline and solution runs can loosely be compared with each other for a few special analyses. Therefore, the simulation runs are abbreviated as B ("baseline") and S ("solution"). Mouse-tracker data only exists for the B runs as mouse-tracking has only been implemented for S runs' setup; eye-tracker data exists for all runs.

The average improvement factor is calculated as shown in Equation (5) to sketch the enhancement of the probability (P) re-assignment (ra) concept compared to uniform (u) probabilities per command (cmd):

$$Improvement\ Factor = \frac{P(cmd)_{ra}}{P(cmd)_{u}}. \quad (5)$$

Five conditions or condition combinations, respectively, for the re-assignment of prediction probabilities based on aircraft level were analyzed with their influence on the prediction accuracy:

1. Only eye-tracking fixation duration of last 5 s to be considered ($ETfix_{dur}$)
2. Only eye-tracking fixation counts of last 10 s to be considered ($ETfix_{cnt}$)
3. Only mouse-tracking interaction data of last 3 s to be considered (MTint)
4. Combining (1) and (2) with 50% weightage each (ET)
5. Combining (4) with 70% weightage and (3) with 30% weightage (ET+MT).

From Equation (6) and using the definition in Table 3, *Accuracy* is defined as the percentage of correctly predicted ATCo commands. In other words, it is the number of commands predicted with above-average probabilities (compared to uniform average probabilities) which were actually issued plus the number of commands predicted with

average or below-average probabilities which were not issued divided by the number of all predicted commands:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}. \quad (6)$$

Table 3. Confusion matrix of ATC commands predicted vs. actually issued commands.

	Command Issued YES	**Command Issued NO**
Command Predicted YES	True Positive (TP)	False Positive (FP)
Command Predicted NO	False Negative (FN)	True Negative (TN)

More precisely, the following *Accuracy* values always consider Top *N* aircraft, e.g., for Top 2 A/C, the two aircraft callsigns that have the highest re-assigned probability compared to the other aircraft. Hence, if the ATCo actually issues a command to one of the two highest-ranked aircraft in terms of prediction probability, it is a TP. If the ATCo issues a command to the third ranked aircraft, it would be a FN. An aircraft is a FP if its callsign was predicted with above-average probability, but is not affected by the ATC command at the timetick it was issued. Finally, a callsign is said to be a TN if the used callsign was predicted with average or below-average probability and was not issued a command by the ATCo. As noted above, gazes on aircraft only influence the command prediction probability of callsigns if commands with the aircraft callsigns have been predicted in the basic implementation, i.e., in 3.2% of the cases aircraft callsigns receive a command that was not predicted. As it was neither predicted in the basic implementation, nor in the enhanced implementation, this has no negative influence on the defined *Accuracy*. Hence, if *N* is set to the maximum number of aircraft, *Accuracy* for *Top N* will be 100%.

Usually, there is a high one-digit number of aircraft to be considered at the same time as these are the aircraft under ATCo's responsibility. However, commands are only predicted for some of those aircraft as prediction for other aircraft might temporarily not be reasonable due to their motion characteristics. So, for each point in time when the ATCo issues one or multiple commands, there are usually multiple aircraft to be considered. For the four conducted simulation runs, commands have been predicted for 7.8 aircraft on an average at a time. Hence, for 149 prediction timeticks (100 speech utterances plus 49 mouse issuing occasions) almost 1200 aircraft callsigns have been predicted in total per run. Based on experiments, it is thus most reasonable to consider the Top 3 A/C only. Top 3 A/C are selected as shown in Table 4.

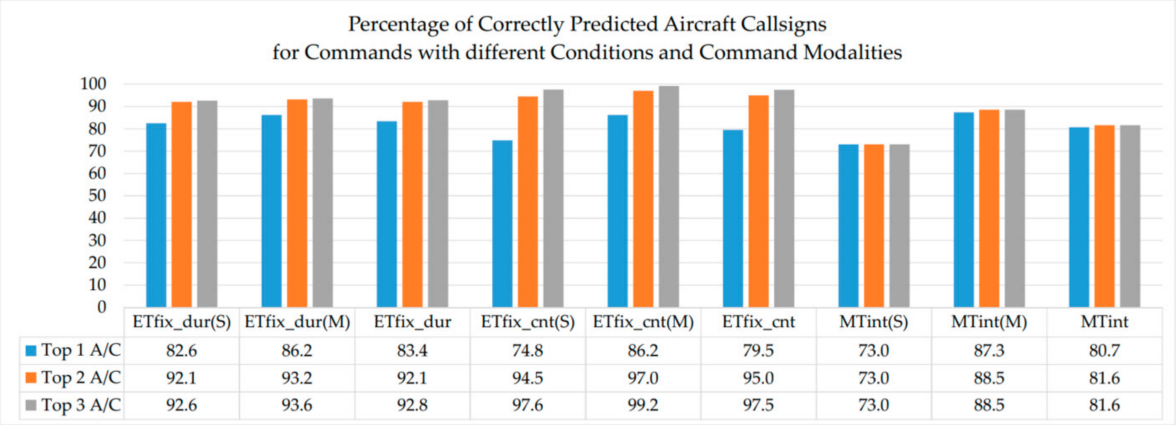
Table 4. Example of prediction sets for Top *N* A/C based on Table 1.

	***N* Highest Prediction Probability for Aircraft Callsign**	**Probability Sum**
Top 1 A/C	{DLH5MA}	0.69 [1]
Top 2 A/C	{DLH5MA; UAE57}	0.82 [2]
Top 3 A/C	{DLH5MA; UAE57} [3]	0.82

[1] 3 × 0.23 (for the three commands of DLH5MA); [2] 3 × 0.23 + 0.13 (for the three commands of DLH5MA and the one command of UAE57); [3] neither of the three further aircraft {AFR641P; BAW936; KLM1853} is considered for Top 3 as they all have the same overall probability sum of 0.06 in Table 1 and there would be no single choice aircraft.

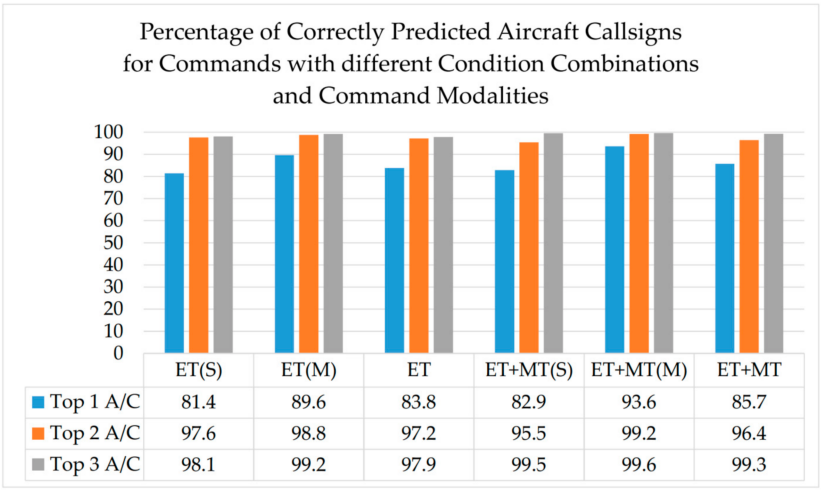
6.1.2. Accuracy of Aircraft Callsign Prediction for ATC Commands Based on Interaction Data

The percentages of correctly predicted aircraft callsigns for ATC commands based on Top 1/2/3 A/C for the input modalities speech (S), mouse (M), and both combined, considering the five interaction conditions are shown in Figures 6 and 7 for both B runs in average.



	ETfix_dur(S)	ETfix_dur(M)	ETfix_dur	ETfix_cnt(S)	ETfix_cnt(M)	ETfix_cnt	MTint(S)	MTint(M)	MTint
Top 1 A/C	82.6	86.2	83.4	74.8	86.2	79.5	73.0	87.3	80.7
Top 2 A/C	92.1	93.2	92.1	94.5	97.0	95.0	73.0	88.5	81.6
Top 3 A/C	92.6	93.6	92.8	97.6	99.2	97.5	73.0	88.5	81.6

Figure 6. Correctly predicted aircraft callsigns for ATC commands when considering Top 1/2/3 aircraft for single interaction data conditions per command modality (speech: S; mouse: M; combined).



	ET(S)	ET(M)	ET	ET+MT(S)	ET+MT(M)	ET+MT
Top 1 A/C	81.4	89.6	83.8	82.9	93.6	85.7
Top 2 A/C	97.6	98.8	97.2	95.5	99.2	96.4
Top 3 A/C	98.1	99.2	97.9	99.5	99.6	99.3

Figure 7. Correctly predicted aircraft callsigns for ATC commands when considering Top 1/2/3 aircraft for combined interaction data conditions per command modality (speech: S; mouse: M; combined).

The number of correctly predicted aircraft callsigns increases for the analyzed stand-alone conditions from Top 1 A/C to Top 3 A/C (see Figure 6). The gaze fixation duration metric alone achieves accuracy results above 80% for Top 1 A/C which further increases to around 93% for both Top 2 and Top 3 A/C. The gaze count metric is slightly less accurate in predicting Top 1 A/C as compared to gaze fixation duration metric, but significantly improves the accuracy to around 95% for Top 2 A/C and 98% for Top 3 A/C (see Figure 6). The mouse interaction metric behaves almost in the same for all the three Top A/C categories with accuracies between 73% and 89% (see Figure 6), i.e., the ATCo either has just moved the mouse to the aircraft, which gets the next command or the mouse is not moved at all to that aircraft during the last ten seconds. For all three metrics, aircraft callsigns are predicted more accurately if ATC commands are given via mouse (M) rather than speech (S).

When combining the two eye-tracking metrics or even combining all three interaction metrics, the accuracy of probabilities for aircraft callsign prediction improves significantly (see Figure 7). Independent of the command modality used, from the average values we see that an accuracy rate of 84% and 86% for ET and ET+MT for Top 1 A/C, 97% and 96% for ET and ET+MT for Top 2 A/C, and 98% and 99% for ET and ET+MT for Top 3 A/C was achieved. This implies that the prediction error rates decrease significantly from 16% to 2% (factor of 8 improvement) when Top 3 A/C is predicted as compared to Top 1 A/C for the case when just ET was used. Similarly, when both ET and MT was

used, the prediction error rates decrease from 14% to 1% (factor of 14 improvement) when Top 3 A/C is predicted as compared to Top 1 A/C. Another impressive result is to compare the prediction error rates for speech modality of the three single modalities for Top 3 A/C of 7.4% ($ETfix_{dur}$), 2.4% ($ETfix_{cnt}$), and 27% (MTint) with the prediction error rate of the combined condition ET+MT(S) of 0.5%—up to a factor of 54 improvement. Overall, it is a factor of 25 improvement when comparing the average prediction error rate of the three single modalities (12.3%) to the combined condition for Top 3 A/C *Accuracy*.

6.1.3. Improvement Factor for Predicted ATC Commands Based on Interaction Data

The improvement factor for all five conditions and command modalities vary between 3.4 and 6.4 as shown in Figures 8 and 9 for both B runs on average. Again, as for the Top A/C analysis, the factor is higher with mouse as command modality. The metrics gaze fixation duration and fixation count achieve improvement factors above 5 and around 4, respectively. The metric mouse interaction is more dependent on the command modality with a factor of 4.9 over all commands. Yet, all the factors illustrated in Figure 8 indicates that the re-assigned probabilities are much better on average as compared to the basic uniform probabilities.

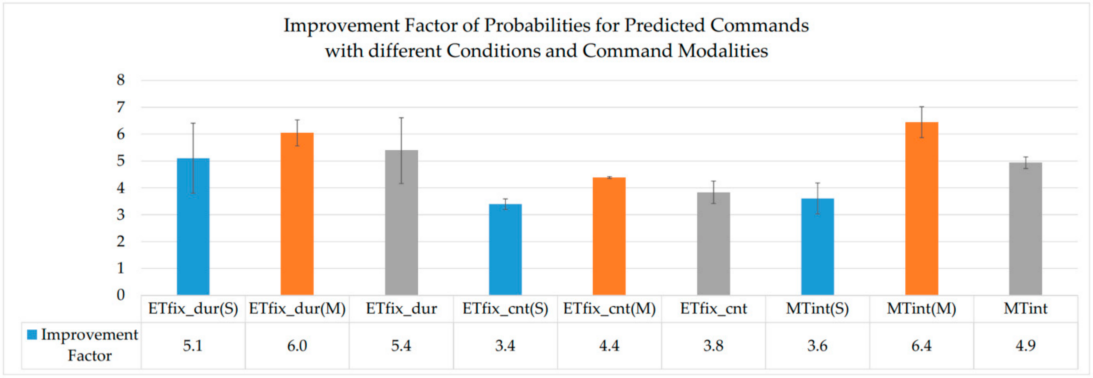


Figure 8. Improvement factors for command prediction probabilities for single interaction data conditions per command modality (speech: S; mouse: M; combined) with positive and negative standard deviation of the two average values per run (black lines).

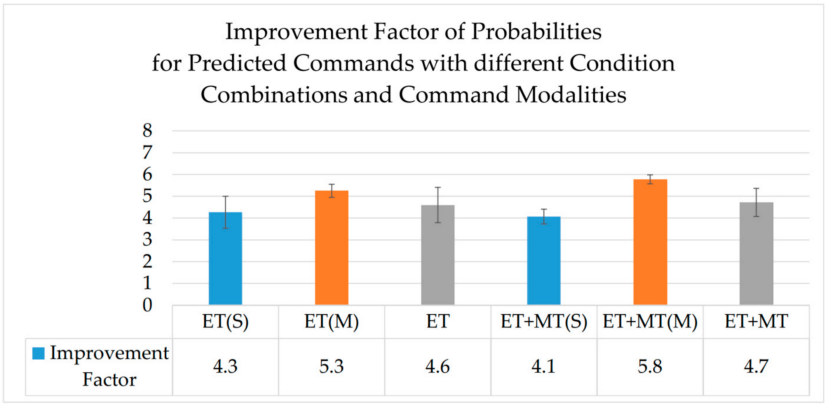


Figure 9. Improvement factors for command prediction probabilities for combined interaction data conditions per command modality (speech: S; mouse: M; combined) with positive and negative standard deviation of the two average values per run (black lines).

When combining the eye-tracking metrics and also further integrating the mouse-tracking metric, the average factors for ET and ET+MT are 4.6 and 4.7, respectively.

6.1.4. Detailed Analysis of Specific Results and Discussion on Probability Re-Assignment Quality

Given the above numbers, it is of interest which of the results per condition and per command modality should be interpreted as the core result. As ATCos usually issue commands via speech and the combination of using all three interaction metrics from eye- and mouse-tracking demonstrated to be the most feasible option under the given circumstances, the values for ET+MT(S) should be selected as core results. Thus, an improvement factor of 4.1 (3.7 and 4.4 for the two controllers each per run) is achieved. Furthermore, above 99.5% of aircraft callsigns for ATC commands have been correctly predicted for Top 3 A/C (95.5% for Top 2 A/C and 82.9% for Top 1 A/C). For one ATCo, prediction of Top 3 A/C even reached an accuracy of 100%. For the condition ET+MT(S) with speech command modality, 92% of improvement factors per speech utterance are greater than 1 showing a positive effect of the investigated re-assignment probability implementation.

When correlating Top 1 A/C data from mouse-tracking and eye-tracking, around 66% (two thirds) of predicted aircraft callsigns for ATC commands match, with similar numbers for correct and wrong predictions. When correlating Top 1 A/C data from mouse-tracking and Top 2 A/C data from eye-tracking, 79% of all predicted aircraft callsigns for ATC commands match—83% for correct predictions and 69% for wrong predictions. Hence, there is a slight potential to further filter out wrong predictions by analyzing and comparing single conditions.

The improvement factors for all B-runs analyzed independent of controller, condition, and command modality are always greater than 3 showing a good robustness of the enhanced command prediction probabilities when using ATCo interaction data. The greatest improvement factor for a single run was 7 for one controller in condition with mouse-tracking data only and commands issued via mouse (MTint(M)). If ATCos issue commands via speech, they could basically be looking anywhere. If ATCos issue commands via mouse, they are more or less forced to look at the aircraft radar label and they are definitely forced to move the mouse onto the label to open the intended drop-down menus and select the right values. So, a factor of around 7 seems to indicate the greatest possible factor when considering interaction data. However, the use of mouse-tracking data depends on the CWP and command modality design.

Probabilities of ATC commands derived from interaction data when issuing commands via mouse (ET+MT(M)) and data link can still be used for plausibility checking of command contents. When analyzing all four runs together (2xB, 2xS) for all command modalities and the condition ET, we still achieve 75.1% for Top1 A/C, 90.6% for Top 2 A/C, 93.9% for Top 3 A/C and an improvement factor of 4.1 even if the concept was not intended to be applied on the S-runs.

Some further results for other conditions and modalities are also noteworthy. When considering Top 1 A/C for $\text{ETfix}_{\text{dur}}$ in S-runs with commands issued via mouse, there exists no correctly predicted aircraft callsigns. This is a conceptional issue as the commands are only issued after the time for optional manual correction has passed—quite a long time after visually checking the aircraft radar label values inserted via mouse before. The improvement factor and the accuracy increase when the analysis duration is extended, i.e., by looking more into the past to gather interaction data. However, this fact together with the high percentages of B-runs prove the pre-assumptions very well that upcoming ATCo actions are connected to gazes and even non-visual checking is related to hardly any ATCo action concerning a displayed aircraft.

6.1.5. Re-Assigned Prediction Probability Evaluation on Command Type Level

As described in Section 4, the concept of re-assigning prediction probabilities encompasses aircraft callsign level and command type level. However, only aircraft callsign level has been implemented so far. To estimate the further benefits of the command type level, we applied a generalized post-analysis on the command prediction results with re-assigned probabilities. More precisely, we increase the probabilities of command types

that were issued more often and decrease probabilities of command types that were seldom issued. According to the analysis at the beginning of Section 6, we again re-assign the probabilities of the three most often used command types. Thus, for analysis, *DESCEND*, *HEADING*, and *CLEARED ILS* commands have twice as high probability as all other command types for the same aircraft callsign. This reveals an assumed benefit of having different probabilities even for command types.

With this analysis, the improvement factor will further increase by 0.4 when considering different command types for each aircraft callsign. However, it must be mentioned that the analysis approach is just based on statistical incidence, while the concept approach bases on concrete air traffic situations that can be determined via surveillance data. Hence, it is unclear if the improvement factor will in reality be higher or lower than 0.4. Furthermore, it is unclear what the effect on ABSR output will be for command types that occur less frequently, e.g., only less than every tenth command. Though, some of these less frequently occurring command types such as *CONTACT* can be predicted quite reliably in space and time. Hence, it is assumed that a positive influence and an improvement factor increase of more than 0.4 is achievable when implementing the re-assigned probability on command type level.

6.2. Using Gazes for Confirmation with Potential Visual Attention Guidance for Speech Recognition Output

In the solution runs 146 ATC commands have been extracted on average from speech utterances. The number of relevant speech utterances is only 123 as often multiple ATC commands were given to aircraft in single utterances. All 123 speech recognition outputs for verbal utterances have been acknowledged via gaze on an aircraft radar label, i.e., the ATCo visually checked one or more at the same time yellow highlighted ABSR output values in a single aircraft radar label. Also, the escalation of saliency levels to enforce the ABSR output check technically worked without any problems. Roughly 120,000 peripheral views on elements at the situation data display have been calculated.

6.2.1. Quantitative Questionnaire Results and Discussion

The two subjects rated higher workload for the solution run than for the baseline run, i.e., average Bedford scale workload [98] was 4 for baseline and 7 for solution as well as Raw NASA-TLX scale [99,100] without weighted ratings was 35 for baseline and 51 for solution. The overall score of the system usability scale (SUS) was 77 (range "good") [101,102]. The ratings for robustness and reliability of the tested system were around the scale mean value. These numbers and the following qualitative feedback should not be generalized given only two study subjects, but can indicate a tendency.

6.2.2. Qualitative Questionnaire Results and Discussion

The different frame colors around aircraft radar labels of higher saliency levels seldom appeared for the two subjects as the solution system almost always detected the subjects' gaze at the colored frame in the first saliency level. So, the colors, numbers, and durations of the additional saliency levels could hardly be correctly judged with regards to usefulness. Nevertheless, the eye-tracking based attention guidance for ABSR output was judged to give a medium added value on a scale from very low to very high. Moving and freezing of gazes at a certain aircraft radar label was perceived as physically demanding to some extent. However, the responsiveness of the system given the hardware latency strongly impacted the controlling task in baseline and solution run.

Subjects felt that they had sufficient amount of time to correct the presented ABSR output after the aircraft radar label frame turned green for the confirmation saliency level. The duration for escalating to a higher saliency level should not be changed due to the subjects' ratings. However, the duration of displaying the green aircraft radar label frame in the confirmation saliency level could be reduced. Both subjects voted to decrease the number of different saliency levels. Three different levels are sufficient due to the subjects' opinion. The aircraft radar label frames were found to be unobtrusive, but sometimes

there were too many green frames at the same time, because the ATCo issued many ATC commands in a short amount of time. The maximum number of visible green frames could be reduced to three. The green frames indicate the time to correct the ABSR output after looking at the label. However, the expectation related to a highlighting frame would be that visual attention is required which is not the case. So, it could be a good idea to completely eliminate the green frame when looking away to only let the yellow highlighted ABSR output value remain for a few seconds without an aircraft radar label frame.

After manually clicking check mark and cross in the baseline run subjects felt to have cognitively finished their checking task. This feeling was different for the visual check as the response state, i.e., yellow ABSR output turning white still takes some time as there is still some time remaining for possible correction.

Also, the threshold times for saliency levels could be dependent on the number of highlighted aircraft radar label frames. One subject wished to have check mark and cross even next to the visual ABSR confirmation to be able to return to the default saliency level earlier. Furthermore, parallelly checking ABSR output and pilot readback might be difficult as one or both of them could contain errors and "appear" at the same time. In case of multiple commands in the same transmission or multiple transmissions shortly after each other for the same aircraft it was not clear which elements were already accepted and which were not.

This feedback shows basic feasibility of the visual confirmation concept and implementation without general showstoppers and encourages further advances based on reasonable suggestions.

7. Conclusions and Overall Discussion

The four general research objectives have been fulfilled, i.e., (1) eye and mouse movements of ATCos can be recorded and post-processed, (2) relevant information is extracted from such data and integrated into an ABSR system, (3) probabilities for predicted ATCo commands are calculated with good accuracy, and (4) ABSR output can be visually confirmed by ATCos in a CWP system prototype.

Eye- and mouse-tracking were rated to be unobtrusive and important features to easily support ABSR applications with more accurate data and interaction options. Visual confirmation of ABSR output technically worked and confirms that state-of-the-art eye-tracking accuracy is sufficient for applications in various domains and even in the safety-critical ATC domain.

Command prediction probabilities improved by a factor of four on average compared to an existing state-of-research prototype (basic implementation) and included more than 95% of correct aircraft callsigns for Top 2 A/C and even more than 99.5% of correct aircraft callsigns for Top 3 A/C analysis. Thus, Top 2 A/C seems to be sufficient to consider for probability re-assignment even if Top 3 A/C is slightly better. The combination of using all eye-tracking and mouse-tracking metrics together was superior over using some of these metrics alone with an improvement factor for the prediction error rate of 25. This confirms state-of-the-art knowledge that using multiple sensor data is superior to just using single sensor data. To the best of our knowledge no eye- and mouse-tracking based ATC command prediction system or prototype, as well as no visual ASR output confirmation exists in the academic world that could be compared with the results in this paper.

The command predictions support the ABSR engine to reduce command recognition error rates if timely considerable in the search space of the engine. Reduced error rates further enable benefits for speech recognition applications that may lead to reduced workload or increased accuracy of safety net functions. Hence, the concept of visual (and mouse-hover) confirmation should be refined and implementation should be advanced, the concept of re-assigned probabilities based on eye- and mouse-tracking data should be further implemented.

It has to be clearly stated that our one-shot experimental case study without any control group and many possible confounding variables has very low internal validity and

cannot reveal any cause-and-effect relationships. The reported results base on a sample size of just two study subjects and can therefore not be generalized. The reported results might be interpreted as a vague tendency on usefulness of implemented prototypes and indicate that it is worth to move forward with our research from pre-experimental design. Nevertheless, the results presented in this paper tremendously help to design a future broader true experimental design study with randomized groups and clearly defined independent and dependent variables after fixing the reported minor technical issues of the prototypic CWP.

For example, the study design should consider to let all saliency levels appear a number of times to be better judgeable. In addition, the duration of training runs should be extended to reduce the effect of subjects on results with being new and unfamiliar with the elements of the prototypic CWP.

The two controllers had a different professional background, i.e., different number of years of experience as ATCo in approach or tower domain and different experience levels in ATC research. This background and the knowledge about actively participating in a study might have influenced their performance and their reported judgements in a positive or negative way. However, this influencing effect might be bigger for the conceptual element with visual ABSR output confirmation than for the visually nontransparent ATC command prediction rescoring.

It has also to be noted that the explained pre-assumptions about the connection between visual attention and spot of ATCos' gaze have limitation implications, i.e., the effects of implications are different for different CWPs, ATCos, and other aspects of the working environment. The reported qualitative and quantitative results enable to assess the two implemented techniques in a human-in-the-loop simulation trial with more ATCos in the near future. Then, it can also be determined in detail how much the improved command prediction probabilities help in terms of ASR engine's word error rate, ABSR system's command recognition rate, and further following measures such as ATCo workload when using the system.

All in all, this paper has given first evidence that using further interaction data of a controller working position such as eye-tracking and mouse-tracking can easily enhance existing ATC system prototypes or be integrated in advanced CWP prototypes as demonstrated with functionalities around an Assistant Based Speech Recognition system.

8. Outlook on Future Work

The following subsections sketch some future work per each of the two conceptual elements and in general related to CWP interaction.

8.1. Outlook on Command Prediction Probability Re-Assignment

Given an improved eye-tracker accuracy, e.g., with advanced devices, it could be checked whether the ATCo looked at, e.g., the label value for current speed of an aircraft. This would lead to an increased likelihood of speed commands for this aircraft or other aircraft being looked at in close timely proximity. The improvement factor for re-assigned ATCo command predictions might be further enhanced if the weighting, e.g., 35% $ETfix_{dur}$, 35% $ETfix_{cnt}$, and 30% MTint would be changed dynamically during a simulation run. If it is detected by the mouse-tracker, that the mouse is inactive or the human operator has many eye gaze saccades, the weighting could be adapted.

Legally collecting large amounts of relevant eye- and mouse-tracking data from CWPs—in laboratories or real-life—might be slightly easier than recording radiotelephony utterances due to privacy issues of personal data existing in some countries even if all interaction data could be used in anonymized form to derive patterns and human erroneous behavior. Machine learning on a huge amount of ATC interaction data from eye-tracking, mouse-tracking, and speech recordings could even more automatically individualize re-assigned probabilities for command predictions.

8.2. Outlook on ABSR Output Confirmation Mode

Saliency levels should be reduced in their number and re-designed in order to be less intrusive. Taking the existing attention guidance implementation as role model [50], the levels may escalate as follows: The default *transparent* saliency level remains unchanged as well as the first saliency level *white* directly appears with yellow ABSR output values. However, after a few seconds without attention-based trigger, a semi-transparent circle around the aircraft icon should appear. If this visual cue and the white label frame remain undetected, the semi-transparent circle could also receive a flashlight effect for some additional seconds as the highest saliency level. In case the ATCo's attention has been determined to have rested on a highlighted aircraft label, there should be no label frame of any color. The ABSR output value might stay yellow or become another color as visual feedback for checking status for the remaining optional correction time. If the correction time has passed or the highest saliency level duration has passed, all accepted label values turn to white. Furthermore, the optional time for correcting ABSR output should be dependent on the number of aircraft currently under responsibility, i.e., to give the ATCo more time if there are more aircraft to monitor and potential tasks to perform before correcting aircraft radar label input. Also, the time for escalation of saliency levels and the time for optional correction could be made command type specific. In situations of dense air traffic, it might be more important to confirm altitude and heading commands than to confirm *CONTACT* commands.

The feature of visual checking and confirmation via eye gaze could also be applied to other parts of CWPs. One example would be highlighted warnings, e.g., on automatically detected readback errors or medium-term conflict alerts with following escalation and de-escalation via attention guidance mechanisms. Another example is the acknowledgement of the final command in the TriControl prototype via gaze instead of a touch gesture.

8.3. Outlook on General Improvements for CWP Interaction

In general, the approximated ATCos' visual attention will be used to assist ATCos in a more convenient way, i.e., giving information at the time and spot that is deemed most reasonable given the current situation. Besides, even further sensors can be included to analyze the ATCos' CWP interaction, e.g., integrate an audio-visual speech recognition system into ABSR.

As a next concrete step, both conceptual techniques will be applied for upcoming ABSR studies in the approach, en-route, and even tower domain.

Author Contributions: O.O. was responsible for basic concept, supervision of J.A.'s master's thesis and I.-T.S.'s bachelor's thesis (for both theses including support for concept refinement, literature research, programming, testing, result analysis, etc.), conduction of the study, and concisely writing this article. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data are not publicly available due to included personal data of controllers.

Acknowledgments: We like to thank Hartmut Helmke, Shruthi Shetty, Robert Hunger, Michael Finke (all DLR, Germany) for their paper reviews as well as Irina Stefanescu (Technical University Bucharest, Romania) and Norbert Englisch (Technische Universität Chemnitz, Germany) for co-supervising the university theses.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. ICAO. *Air Traffic Services–Air Traffic Control Service, Flight Information Service, Alerting Service, International Civil Aviation Organization (ICAO), Annex 11*; ICAO: Montréal, QC, Canada, 2001.
2. Cardosi, K.M.; Brett, B.; Han, S. *An Analysis of TRACON (Terminal Radar Approach Control) Controller-Pilot Voice Communications, (DOT/FAA/AR-96/66)*; DOT FAA: Washington, DC, USA, 1996.
3. Skaltsas, G.; Rakas, J.; Karlaftis, M.G. An analysis of air traffic controller-pilot miscommunication in the NextGen environment. *J. Air Transp. Manag.* **2013**, *27*, 46–51. [CrossRef]
4. ICAO; ATM (Air Traffic Management). *Procedures for Air Navigation Services*; International Civil Aviation Organization (ICAO), DOC 4444 ATM/501; ICAO: Montréal, QC, Canada, 2007.
5. Lin, Y. Spoken Instruction Understanding in Air Traffic Control: Challenge, Technique, and Application. *Aerospace* **2021**, *8*, 65. [CrossRef]
6. Helmke, H.; Ohneiser, O.; Buxbaum, J.; Kern, C. Increasing ATM Efficiency with Assistant Based Speech Recognition. In Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 26–30 June 2017.
7. Ohneiser, O.; Jauer, M.-L.; Rein, J.R.; Wallace, M. Faster Command Input Using the Multimodal Controller Working Position "TriControl". *Aerospace* **2018**, *5*, 54. [CrossRef]
8. Ohneiser, O.; Sarfjoo, S.; Helmke, H.; Shetty, S.; Motlicek, P.; Kleinert, M.; Ehr, H.; Murauskas, Š. Robust Command Recognition for Lithuanian Air Traffic Control Tower Utterances. In Proceedings of the InterSpeech 2021, Brno, Czech Republic, 30 August–3 September 2021.
9. Connolly, D.W. Voice Data Entry in Air Traffic Control. In Proceedings of the Voice Technology for Interactive Real-Time Command/Control Systems Application, N93-72621, Moffett Field, CA, USA, 6–8 December 1977; pp. 171–196.
10. Young, S.R.; Ward, W.H.; Hauptmann, A.G. Layering predictions: Flexible use of dialog expectation in speech recognition. In Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI89), Morgan Kaufmann, Detroit, MI, USA, 20–25 August 1989; pp. 1543–1549.
11. Helmke, H.; Slotty, M.; Poiger, M.; Herrer, D.F.; Ohneiser, O.; Vink, N.; Cerna, A.; Hartikainen, P.; Josefsson, B.; Langr, D.; et al. Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ.16-04. In Proceedings of the IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), London, UK, 23–27 September 2018.
12. Rataj, J.; Helmke, H.; Ohneiser, O. AcListant with Continuous Learning: Speech Recognition in Air Traffic Control. In *Air Traffic Management and Systems IV, Selected Papers of the 6th ENRI International Workshop on ATM/CNS (EIWAC2019)*; Springer: Singapore, 2021; pp. 93–109.
13. Helmke, H.; Ohneiser, O.; Mühlhausen, T.; Wies, M. Reducing Controller Workload with Automatic Speech Recognition. In Proceedings of the 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 25–29 September 2016.
14. Helmke, H.; Rataj, J.; Mühlhausen, T.; Ohneiser, O.; Ehr, H.; Kleinert, M.; Oualil, Y.; Schulder, M. Assistant-Based Speech Recognition for ATM Applications. In Proceedings of the 11th USA/Europe Air Traffic Management Research and Development Seminar (ATM2015), Lisbon, Portugal, 23–26 June 2015.
15. Cordero, J.M.; Dorado, M.; de Pablo, J.M. Automated speech recognition in ATC environment. In Proceedings of the 2nd International Conference on Application and Theory of Automation in Command and Control Systems, London, UK, 29–31 May 2012; pp. 46–53.
16. Chen, S.; Kopald, H.D.; Elessawy, A.; Levonian, Z.; Tarakan, R.M. Speech inputs to surface safety logic systems. In Proceedings of the IEEE/AIAA 34th Digital Avionics Systems Conference (DASC), Prague, Czech Republic, 13–17 September 2015.
17. Chen, S.; Kopald, H.D.; Chong, R.; Wei, Y.; Levonian, Z. Read back error detection using automatic speech recognition. In Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 26–30 June 2017.
18. Gürlük, H.; Helmke, H.; Wies, M.; Ehr, H.; Kleinert, M.; Mühlhausen, T.; Muth, K.; Ohneiser, O. Assistant Based Speech Recognition—Another Pair of Eyes for the Arrival Manager. In Proceedings of the 34th Digital Avionics Systems Conference (DASC), Prague, Czech Republic, 13–17 September 2015.
19. Ohneiser, O.; Helmke, H.; Ehr, H.; Gürlük, H.; Hössl, M.; Mühlhausen, T.; Oualil, Y.; Schulder, M.; Schmidt, A.; Khan, A.; et al. Air Traffic Controller Support by Speech Recognition. In *Advances in Human Aspects of Transportation: Part II, Proceedings of the International Conference on Applied Human Factors and Ergonomics (AHFE), Krakow, Poland, 19–23 July 2014*; Stanton, N., Landry, S., Di Bucchianico, G., Vallicelli, A., Eds.; CRC Press: Boca Raton, FL, USA; pp. 492–503.
20. Updegrove, J.A.; Jafer, S. Optimization of Air Traffic Control Training at the Federal Aviation Administration Academy. *Aerospace* **2017**, *4*, 50. [CrossRef]
21. Schäfer, D. Context-Sensitive Speech Recognition in the Air Traffic Control Simulation. Ph.D. Thesis, University of Armed Forces, Munich, Germany, 2001.
22. Kleinert, M.; Helmke, H.; Siol, G.; Ehr, H.; Finke, M.; Srinivasamurthy, A.; Oualil, Y. Machine Learning of Controller Command Prediction Models from Recorded Radar Data and Controller Speech Utterances. In Proceedings of the 7th SESAR Innovation Days, Belgrade, Serbia, 28–30 November 2017.

23. Helmke, H.; Kleinert, M.; Ohneiser, O.; Ehr, H.; Shetty, S. Machine Learning of Air Traffic Controller Command Extraction Models for Speech Recognition Applications. In Proceedings of the IEEE/AIAA 39th Digital Avionics Systems Conference (DASC), Virtual, 11–16 October 2020.
24. SESAR2020-Exploratory Research Project HAAWAII (Highly Automated Air Traffic Controller Workstations with Artificial Intelligence Integration). Available online: https://www.haawaii.de (accessed on 19 August 2021).
25. Ohneiser, O.; Helmke, H.; Shetty, S.; Kleinert, M.; Ehr, H.; Murauskas, Š.; Pagirys, T. Prediction and extraction of tower controller commands for speech recognition applications. *J. Air Transp. Manag.* **2021**, *95*, 102089. [CrossRef]
26. Kleinert, M.; Helmke, H.; Moos, S.; Hlousek, P.; Windisch, C.; Ohneiser, O.; Ehr, H.; Labreuil, A. Reducing Controller Workload by Automatic Speech Recognition Assisted Radar Label Maintenance. In Proceedings of the 9th SESAR Innovation Days, Athens, Greece, 2–5 December 2019.
27. Nguyen, V.N.; Holone, H. N-best list re-ranking using syntactic score: A solution for improving speech recognition accuracy in air traffic control. In Proceedings of the 16th International Conference on Control, Automation and Systems (ICCAS), Gyeongju, Korea, 16–19 October 2016; pp. 1309–1314.
28. Shore, T.; Faubel, F.; Helmke, H.; Klakow, D. Knowledge-Based Word Lattice Rescoring in a Dynamic Context. In Proceedings of the Inter Speech 2012, Portland, OR, USA, 9–13 September 2012.
29. Punde, P.A.; Jadhav, M.E.; Manza, R.R. A study of Eye Tracking Technology and its applications. In Proceedings of the 1st International Conference on Intelligent Systems and Information Management (ICISIM), Maharashtra, India, 5–6 October 2017; pp. 86–90.
30. Farnsworth, B. What Is Eye Tracking and How Does It Work? Available online: https://imotions.com/blog/eye-tracking-work/ (accessed on 19 August 2021).
31. Farnsworth, B. 10 Most Used Eye Tracking Metrics and Terms. Available online: https://imotions.com/blog/10-terms-metrics-eye-tracking/ (accessed on 19 August 2021).
32. Bhattarai, R.; Phothisonothai, M. Eye-Tracking Based Visualizations and Metrics Analysis for Individual Eye Movement Patterns. In Proceedings of the 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), Chonburi, Thailand, 10–12 July 2019; pp. 381–384.
33. Poole, A.; Ball, L.J. Eye tracking in human-computer interaction and usability research: Current status and future prospects. In *Encyclopedia of Human Computer Interaction*; Idea Group Reference: Hershey, PA, USA, 2006; pp. 211–219.
34. Salvucci, D.; Goldberg, J.H. Identifying fixations and saccades in eye-tracking protocols. In Proceedings of the Eye Tracking Research & Application Symposium, ETRA 2000, Palm Beach Gardens, FL, USA, 6–8 November 2000.
35. Scholz, A. Eye Movements, Memory, and Thinking–Tracking Eye Movements to Reveal Memory Processes during Reasoning and Decision-Making. Ph.D. Thesis, Technische Universität Chemnitz, Chemnitz, Germany, 2015.
36. Lorigo, L.; Haridasan, M.; Brynjarsdóttir, H.; Xia, L.; Joachims, T.; Gay, G.; Granka, L.; Pellacini, F.; Pan, B. Eye Tracking and Online Search: Lessons Learned and Challenges Ahead. *J. Am. Soc. Inf. Sci. Technol.* **2008**, *59*, 1041–1052. [CrossRef]
37. Fraga, R.P.; Kang, Z.; Crutchfield, J.M.; Mandal, S. Visual Search and Conflict Mitigation Strategies Used by Expert en Route Air Traffic Controllers. *Aerospace* **2021**, *8*, 170. [CrossRef]
38. Kang, Z.; Mandal, S.; Dyer, J. Data Visualization Approaches in Eye Tracking to Support the Learning of Air Traffic Control Operations. In Proceedings of the National Training Aircraft Symposium, Daytona Beach, FL, USA, 14–16 August 2017.
39. Wickens, C.; Hollands, J.; Banbury, S.; Parasuraman, R. *Engineering Psychology and Human Performance*, 4th ed.; Pearson Education: Boston, MA, USA, 2013.
40. Zamani, H.; Abas, A.; Amin, M.K.M. Eye Tracking Application on Emotion Analysis for Marketing Strategy. *J. Telecommun. Electron. Comput. Eng.* **2016**, *8*, 87–91.
41. Goyal, S.; Miyapuram, K.P.; Lahiri, U. Predicting Consumer's Behavior Using Eye Tracking Data. In Proceedings of the 2nd International Conference on Soft Computing and Machine Intelligence (ISCMI), Hong Kong, China, 23–24 November 2015; pp. 126–129.
42. Sari, J.N.; Nugroho, L.; Santosa, P.; Ferdiana, R. The Measurement of Consumer Interest and Prediction of Product Selection in E-commerce Using Eye Tracking Method. *Int. J. Intell. Eng. Syst.* **2018**, *11*, 30–40. [CrossRef]
43. Huang, C.-M.; Andrist, S.; Sauppé, A.; Mutlu, B. Using gaze patterns to predict task intent in collaboration. *Front. Psychol.* **2015**, *6*, 1049. [CrossRef]
44. Eivazi, S.; Bednarik, R. Predicting Problem-Solving Behavior and Performance Levels from Visual Attention Data. In Proceedings of the 2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction, Palo Alto, CA, USA, 13 February 2011.
45. Duchowski, A.T. A breadth-first survey of eye-tracking applications. *Behav. Res. Methods Instrum. Comput.* **2002**, *34*, 455–470. [CrossRef]
46. Traoré, M.; Hurter, C. Exploratory study with eye tracking devices to build interactive systems for air traffic controllers. In Proceedings of the International Conference on Human-Computer Interaction in Aerospace (HCI-Aero'16), Paris, France, 14–16 September 2016; ACM: New York, NY, USA, 2016.
47. Merchant, S.; Schnell, T. Applying Eye Tracking as an Alternative Approach for Activation of Controls and Functions in Aircraft. In Proceedings of the 19th Digital Avionics Systems Conference (DASC), Philadelphia, PA, USA, 7–13 October 2000.
48. Alonso, R.; Causse, M.; Vachon, F.; Parise, R.; Dehaise, F.; Terrier, P. Evaluation of head-free eye tracking as an input device for air traffic control. *Ergonomics* **2013**, *2*, 246–255. [CrossRef] [PubMed]

49. Möhlenbrink, C.; Papenfuß, A. Eye-data metrics to characterize tower controllers' visual attention in a multiple remote tower exercise. In Proceedings of the ICRAT, Istanbul, Turkey, 26–30 May 2014.
50. Ohneiser, O.; Gürlük, H.; Jauer, M.-L.; Szöllősi, Á.; Balló, D. Please have a Look here: Successful Guidance of Air Traffic Controller's Attention. In Proceedings of the 9th SESAR Innovation Days, Athens, Greece, 2–5 December 2019.
51. Rataj, J.; Ohneiser, O.; Marin, G.; Postaru, R. Attention: Target and Actual–The Controller Focus. In Proceedings of the 32nd Congress of the International Council of the Aeronautical Sciences (ICAS), Shanghai, China, 6–10 September 2021.
52. Ohneiser, O.; Jauer, M.-L.; Gürlük, H.; Springborn, H. Attention Guidance Prototype for a Sectorless Air Traffic Management Controller Working Position. In Proceedings of the German Aerospace Congress DLRK, Friedrichshafen, Germany, 4–6 September 2018.
53. Di Flumeri, G.; De Crescenzio, F.; Berberian, B.; Ohneiser, O.; Kraemer, J.; Aricò, P.; Borghini, G.; Babiloni, F.; Bagassi, S.; Piastra, S. Brain-Computer Interface-Based Adaptive Automation to Prevent Out-Of-The-Loop Phenomenon in Air Traffic Controllers Dealing with Highly Automated Systems. *Front. Hum. Neurosci.* **2019**, *13*, 1–17. [CrossRef]
54. Hurter, C.; Lesbordes, R.; Letondal, C.; Vinot, J.L.; Conversy, S. Strip'TIC: Exploring augmented paper strips for air traffic controllers. In Proceedings of the International Working Conference on Advanced Visual Interfaces, Capri Island, Italy, 22–26 May 2012; ACM: New York, NY, USA, 2012; pp. 225–232.
55. Rheem, H.; Verma, V.; Becker, D.V. Use of Mouse-tracking Method to Measure Cognitive Load. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Philadelphia, PA, USA, 1–5 October 2018; Volume 62, pp. 1982–1986.
56. Huang, J.; White, R.; Buscher, G. User see, user point: Gaze and cursor alignment in web search. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12), Austin, TX, USA, 5–10 May 2012; ACM: New York, NY, USA, 2012; pp. 1341–1350.
57. Zgonnikov, A.; Aleni, A.; Piiroinen, P.T.; O'Hora, D.; di Bernardo, M. Decision landscapes: Visualizing mouse-tracking data. *R. Soc. Open Sci.* **2017**, *4*, 170482. [CrossRef] [PubMed]
58. Calcagnì, A.; Lombardi, L.; Sulpizio, S. Analyzing spatial data from mouse tracker methodology: An entropic approach. *Behav. Res.* **2017**, *49*, 2012–2030. [CrossRef] [PubMed]
59. Maldonado, M.; Dunbar, E.; Chemla, E. Mouse tracking as a window into decision making. *Behav. Res.* **2019**, *51*, 1085–1101. [CrossRef] [PubMed]
60. Krassanakis, V.; Kesidis, A.L. MatMouse: A Mouse Movements Tracking and Analysis Toolbox for Visual Search Experiments. *Multimodal Technol. Interact.* **2020**, *4*, 83. [CrossRef]
61. Claypool, M.; Le, P.; Wased, M.; Brown, D. Implicit interest indicators. In Proceedings of the 6th International Conference on Intelligent User Interfaces (IUI'01), Santa Fe, NM, USA, 14–17 January 2001; ACM: New York, NY, USA, 2001; pp. 33–40.
62. Rodden, K.; Fu, X.; Aula, A.; Spiro, I. Eye-mouse coordination patterns on web search results pages. In Proceedings of the CHI '08 Extended Abstracts on Human Factors in Computing Systems, Florence, Italy, 5–10 April 2008.
63. Chen, M.C.; Anderson, J.R.; Sohn, M.H. What can a mouse cursor tell us more? Correlation of eye/mouse movements on web browsing. In Proceedings of the CHI '01 Extended Abstracts on Human Factors in Computing Systems (CHI EA '01), Seattle, WA, USA, 31 March–5 April 2001; ACM: New York, NY, USA, 2001; pp. 281–282.
64. Cooke, N.; Shen, A.; Russell, M. Exploiting a 'gaze-Lombard effect' to improve ASR performance in acoustically noisy settings. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 1754–1758.
65. Cooke, N.; Russell, M. Gaze-contingent ASR for spontaneous, conversational speech: An evaluation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 4433–4436.
66. Shen, A. The Selective Use of Gaze in Automatic Speech Recognition. Ph.D. Thesis, College of Engineering and Physical Sciences, University of Birmingham, Birmingham, UK, 2013.
67. Alhargan, A.; Cooke, N.; Binjammaz, T. Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals. In Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17), Glasgow, UK, 13–17 November 2017; ACM: New York, NY, USA; pp. 479–486.
68. Rasmussen, M.; Tan, Z. Fusing eye-gaze and speech recognition for tracking in an automatic reading tutor-A step in the right direction? In Proceedings of the Speech and Language Technology in Education (SLaTE), Grenoble, France, 30 August–1 September 2013.
69. DLR Institute of Flight Guidance. TriControl–Multimodal ATC Interaction. Available online: http://www.dlr.de/fl/Portaldata/14/Resources/dokumente/veroeffentlichungen/TriControl_web.pdf (accessed on 19 August 2021).
70. Ohneiser, O.; Jauer, M.-L.; Gürlük, H.; Uebbing-Rumke, M. TriControl–A Multimodal Air Traffic Controller Working Position. In Proceedings of the 6th SESAR Innovation Days, Delft, The Netherlands, 8–10 November 2016.
71. Ohneiser, O.; Biella, M.; Schmugler, A.; Wallace, M. Operational Feasibility Analysis of the Multimodal Controller Working Position "TriControl". *Aerospace* **2020**, *7*, 15. [CrossRef]
72. Bernsen, N. Multimodality Theory. In *Multimodal User Interfaces. Signals and Communication Technologies*; Tzovaras, D., Ed.; Springer: Berlin/Heidelberg, Germany, 2008.
73. Nigay, L.; Coutaz, J. A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion. In Proceedings of the INTERCHI'93 Conference on Human Factors in Computing Systems, Amsterdam, The Netherlands, 24–29 April 1993; pp. 172–178.

74. Bourguet, M.L. Designing and Prototyping Multimodal Commands. In Proceedings of the Human-Computer Interaction INTERACT'03, Zurich, Switzerland, 1–5 September 2003; pp. 717–720.
75. Oviatt, S.L. Breaking the Robustness Barrier: Recent Progress on the Design of Robust Multimodal Systems. *Adv. Comput.* **2002**, *56*, 305–341.
76. Oviatt, S.L. Multimodal interactive maps: Designing for human performance. *Hum. Comput. Interact.* **1997**, *12*, 93–129.
77. Cohen, P.R.; McGee, D.R. Tangible multimodal interfaces for safety-critical applications. *Commun. ACM* **2004**, *1*, 1–46. [CrossRef]
78. Seifert, K. Evaluation of Multimodal Computer Systems in Early Development Phases, Original German Title: Evaluation Multimodaler Computer-Systeme in Frühen Entwicklungsphasen. Ph.D. Thesis, Technische Universität Berlin, Berlin, Germany, 2002.
79. Oviatt, S. User-centered modeling for spoken language and multimodal interfaces. *IEEE Multimed.* **1996**, *4*, 26–35. [CrossRef]
80. Den Os, E.; Boves, L. User behaviour in multimodal interaction. In Proceedings of the HCI International, Las Vegas, NV, USA, 22–27 July 2005.
81. Manawadu, E.U.; Kamezaki, M.; Ishikawa, M.; Kawano, T.; Sugano, S. A Multimodal Human-Machine Interface Enabling Situation-Adaptive Control Inputs for Highly Automated Vehicles. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; pp. 1195–1200.
82. Quek, F.; McNeill, D.; Bryll, R.; Kirbas, C.; Arslan, H.; McCullough, K.E.; Furuyama, N.; Ansari, R. Gesture, speech, and gaze cues for discourse segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662), Hilton Head Island, SC, USA, 15 June 2000; Volume 2, pp. 247–254.
83. Shi, Y.; Taib, R.; Ruiz, N.; Choi, E.; Chen, F. Multimodal Human-Machine Interface and User Cognitive Load Measurement. *Proc. Int. Fed. Autom. Control* **2007**, *40*, 200–205. [CrossRef]
84. Pentland, A. Perceptual Intelligence. *Commun. ACM* **2000**, *4*, 35–44. [CrossRef]
85. Oviatt, S.L. Ten myths of multimodal interaction. *Commun. ACM* **1999**, *11*, 74–81. [CrossRef]
86. Oviatt, S.L. Mutual disambiguation of recognition errors in a multimodal architecture. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems, Pittsburgh, PA, USA, 15–20 May 1999; pp. 576–583.
87. Oviatt, S.L.; Coulston, R.; Lunsford, R. When do we interact multimodally? Cognitive load and multimodal communication patterns. In Proceedings of the 6th International Conference on Multimodal interfaces, State College, PA, USA, 13–15 October 2004; pp. 129–136.
88. Neßelrath, R.; Moniri, M.M.; Feld, M. Combining Speech, Gaze, and Micro-gestures for the Multimodal Control of In-Car Functions. In Proceedings of the 12th International Conference on Intelligent Environments (IE), London, UK, 14–16 September 2016; pp. 190–193.
89. Jauer, M.-L. Multimodal Controller Working Position, Integration of Automatic Speech Recognition and Multi-Touch Technology, Original German Title: Multimodaler Fluglotsenarbeitsplatz, Integration von Automatischer Spracherkennung und Multi-Touch-Technologie. Bachelor's Thesis, Technische Universität Braunschweig, Braunschweig, Germany, 2014.
90. Seelmann, P.-E. Evaluation of an Eye Tracking and Multi-Touch Based Operational Concept for a Future Multimodal Approach Controller Working Position, Original German Title: Evaluierung Eines Eyetracking und Multi-Touch Basierten Bedienkonzeptes für Einen Zukünftigen Multimodalen Anfluglotsenarbeitsplatz. Bachelor's Thesis, Technische Universität Braunschweig, Braunschweig, Germany, 2015.
91. SESAR Joint Undertaking. *European ATM Master Plan–Digitalising Europe's Aviation Infrastructure*; SESAR Joint Undertaking: Brussels, Belgium; Luxembourg, 2020.
92. SESAR2020-Industrial Research Solution PJ.16-04. Controller Working Position/Human Machine Interface–CWP/HMI. Available online: https://www.sesarju.eu/projects/cwphmi (accessed on 19 August 2021).
93. Ohneiser, O. *RadarVision-Manual for Controllers, Original German Title: RadarVision–Benutzerhandbuch für Lotsen*; Internal Report 112-2010/54; German Aerospace Center (DLR), Institute of Flight Guidance: Braunschweig, Germany, 2010.
94. Salomea, I.-T. Integration of Eye-Tracking and Assistant Based Speech Recognition for the Interaction at the Controller Working Position. Bachelor's Thesis, "Politehnica" University of Bucharest, Bucharest, Romania, 2021.
95. Wickens, C.D.; McCarley, J.S. *Applied Attention Theory*; CRC Press Taylor & Francis Group: Boca Raton, FL, USA, 2008.
96. Adamala, J. Integration of Eye Tracker and Assistant Based Speech Recognition at Controller Working Position. Master's Thesis, Technische Universität Chemnitz, Chemnitz, Germany, 2021.
97. Ribeiro, M.; Ellerbroek, J.; Hoekstra, J. Review of Conflict Resolution Methods for Manned and Unmanned Aviation. *Aerospace* **2020**, *7*, 79. [CrossRef]
98. Roscoe, A.H. Assessing pilot workload in flight. In Proceedings of the AGARD Conference Proceedings Flight Test Techniques, Lisbon, Portugal, 2–5 April 1984.
99. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*; Hancock, P.A., Meshkati, N., Eds.; North Holland Press: Amsterdam, The Netherlands, 1988; Volume 198.
100. Hart, S.G. Nasa-Task Load Index (NASA-TLX); 20 years later. In Proceedings of the Human Factors and Ergonomics Society, San Francisco, CA, USA, 16–20 October 2006; Volume 50, pp. 904–908.
101. Brooke, J. SUS-A Quick and Dirty Usability Scale. In *Usability Evaluation in Industry*; Jordan, P.W., Thomas, B., McClelland, I.L., Weerdmeester, B.A., Eds.; Taylor and Francis: London, UK, 1996; pp. 189–194.
102. Bangor, A.; Kortum, P.T.; Miller, J.T. An empirical evaluation of the system usability scale. *Int. J. Hum.–Comput. Interact.* **2008**, *24*, 574–594. [CrossRef]

TriControl – A Multimodal Air Traffic Controller Working Position

Oliver Ohneiser, Malte-Levin Jauer, Hejar Gürlük, Maria Uebbing-Rumke
Institute of Flight Guidance, German Aerospace Center (DLR)
Lilienthalplatz 7, 38108 Braunschweig, Germany
{Oliver.Ohneiser;Malte-Levin.Jauer;Hejar.Guerluek;Maria.Uebbing}@DLR.de

***Abstract*—The TriControl multimodal controller working position (CWP) demonstrates a novel concept for natural human-computer interaction in Air Traffic Control (ATC) by integrating speech recognition, eye tracking and multi-touch sensing. All three parts of a controller command – aircraft identifier, command type and value – are inserted by the controllers via different modalities in parallel. The combination of natural gazes at aircraft radar labels, simple multi-touch gestures, and utterances of equivalent values are sufficient to initiate commands to be sent to pilots. This reduces both controller workload and the time needed to initiate controller commands. The concept promises easy, well-adjusted, and intuitive human-computer interaction.**

Keywords—Air Traffic Controller; Human Machine Interaction; Multimodality; Eye Tracking; Automatic Speech Recognition; Multi-touch Gestures; Controller Command; Workload Reduction

I. INTRODUCTION

Current human machine interfaces (HMI) of air traffic controllers mainly focus on the "speech" modality when communicating with pilots. Data link-based communication, wherever available, is generally initiated by mouse or pen input. Controllers usually use mouse and keyboard as interaction devices for keeping system information up-to-date. Multimodal HMIs emphasize the use of richer and more natural ways of interaction by combining different modalities, such as speech, gestures, and gaze. Therefore, they need to interpret information from various sensors and communication channels.

Multimodal systems have the potential to enhance human-computer interaction (HCI) in a number of ways by:

- adapting to a wider range of users, tasks, and situations,
- providing alternative methods for user interaction,
- conveying information via the appropriate communication channel,
- accommodating differences between individual operators by permitting flexible use of input modes,
- improving error avoidance, and
- supporting improved efficiency through faster task completion, especially when working with graphical information.

When people communicate with each other in person they have eye contact, use their hands for gestures and emphasis, and voice for content regarding "facts". Multimodal HMIs represent a new class of user-machine interfaces, applying the same principles from human interaction to human-computer interaction. It is anticipated that they will offer faster, easier, more natural and intuitive methods for data entry.

This capability is a prerequisite for advancing human-machine systems to the point where computers and humans can truly act as a team. Furthermore, air traffic research and development programs like SESAR (Single European Sky ATM (Air Traffic Management) Research Programme) require use and integration of new technologies such as touch- and speech applications for an enhanced controller-system interaction [1]. An efficient way of entering data into the system is also required to enable a beneficial data link application.

The primary goal of TriControl, the DLR demonstrator for a multimodal CWP in the ATC approach area, is to ensure that a human operator can enter data e.g. controller commands more quickly and intuitively. Based upon empirical findings and subjective evaluations we assessed the suitability of different input modes in relation to specific command elements.

To outline the scientific context, chapter II of this paper presents related work on different interaction modalities. Chapter III includes the concept and implementation of the TriControl prototype comprising eye tracking (ET), speech recognition (SR), and multi-touch (MT) gestures. A preliminary evaluation of the implemented system and results of that evaluation are outlined and discussed in chapter IV. Finally, chapter V draws conclusions and identifies future work.

II. RELATED WORK ON MULTIMODAL HUMAN MACHINE INTERACTION

Implementation of multimodal human-computer interaction concepts is still at an early stage in ATC. Nevertheless, different prototypes using modern interaction technologies as single interaction modalities have been developed.

In fact, any interaction modalities that can be digitally recognized by a computer are conceivable for interaction with the system. Within the SESAR work package 10.10.02 technology screening was carried out in order to assess the suitability of current interaction technologies for controller working positions. The multi-touch, eye tracking and handwriting recognition technologies were investigated [2] on the basis of the screening results. Within this research, the technologies were analyzed and prototypes were evaluated. Consolidated assessments were carried out, particularly for multi-touch and eye tracking. Speech recognition has been substantially developed and evaluated in the AcListant® [3] project.

The most promising interaction technologies currently assumed as being suitable for input are multi-touch (haptic modality), eye tracking (visual modality) and speech (auditive modality). DLR has already successfully evaluated implementations in the field of eye tracking [4], multi-touch [5], and speech recognition [6].

A. Eye tracking (ET)

Eye tracking technology offers at least two different opportunities for use in ATC. Firstly, it has been used to assess mental workload [7] and fatigue of controllers. Secondly, there are a number of other reasons for incorporating eye tracking as an input device for controllers [8]: It allows hand-free interaction and facilitates the manipulation of radar labels or electronic flight strips. Another argument in favor of eye tracking is that eye movements are fast and natural. For instance faster selection times were reported with eye–gaze interaction than with other input devices such as the mouse [9]. According to [8] there is empirical evidence that eye trackers can become an efficient pointing device that can be used instead of the mouse [10] or the keyboard [11].

B. Multi-touch (MT)

By scanning the use of multi-touch technology in the ATC area a prototypic implementation of a workstation – announced as “Indra advanced controller working position” – can be found on Indra’s website [12]. Besides stating that “Multi-touch technology is used routinely as a means of interaction” in this CWP, it does not provide a more specific description of what information is gained by MT input or indicate how it is subsequently used.

A master thesis [13] supervised by the German air navigation service provider DFS (DFS Deutsche Flugsicherung GmbH) contains a concept and first application of multi-touch for command input, later to be translated using text-to-speech technology and then sent to the pilots.

This was evaluated using DFS controllers and generated positive feedback on MT usability. The thesis also outlines expectations on deployment in CWPs in the near future.

DLR and DFS collaborated in the SESAR 1 Work package 10.10.02, dealing with ergonomics, hardware and design of the controller working position. Effort was expended in investigating the usability of multi-touch technology at TMA (Terminal Manoeuvring Area) and ACC (Area Control Center) CWPs. The DLR demonstrator with multi-touch interaction was evaluated against a comparable CWP with a mouse interaction concept [5]. In this study fourteen DFS air traffic controllers, aged from 23 to the fifties, were asked to guide approach traffic in a realistic scenario using both the multi-touch and mouse CWP.

Usability (see Figure 1) and workload were assessed. The results revealed higher usability scores for multi-touch technology. Mental effort and task effort were perceived as less of a strain.

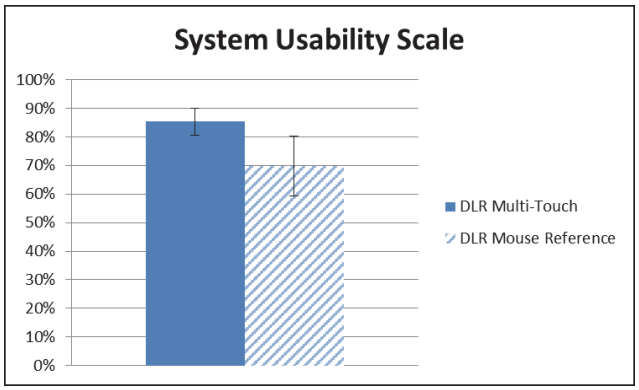


Figure 1. Overall system usability scale (SUS) [14] score multi-touch and mouse reference

The overall investigation indicated that it is likely to be worthwhile to continue developing controller working positions with multi-touch interaction philosophy. The use of multi-touch technology in an experimental context was found:

- not to be a show-stopper due to safety issues,
- to be conceivable at the working position,
- to be error tolerant,
- to be fast and efficient, and
- not to greatly influence controller performance.

The participants therefore encouraged the developers to continue developing the demonstrator.

C. Speech Recognition (SR)

Automatic speech recognition algorithms are capable of converting spoken words or numbers to text. Popular consumer products are for example Siri® [15] or Google’s search by voice [16]. The first steps in integrating speech recognition in ATM systems, including ATC training, took place as much as a quarter century ago [17].

SR may also be used to replace pseudo-pilots in ATC simulation environments [18]. The readback of simulated pilots communicating with controllers can be fulfilled by recognizing controllers' utterances (speech-to-text) and repeating the command in correct phraseology again (text-to-speech). Context knowledge of what utterances are most likely in the current air traffic situation makes it possible to improve speech recognition quality [19].

SR can also detect complete controller commands from ATC vocabulary with acceptable reliability. The knowledge of the spoken commands can be used to support different tasks of controllers (e.g. aircraft radar label maintenance of approach controllers via direct automatic input of controllers' uttered clearances) at their working positions [20]. An approach controller is responsible for merging several streams of air traffic into a single final sequence for specific runways. Highly automated decision support tools such as arrival or departure managers have been developed to support human operators in this challenging task. These systems need to adapt to the controller's intentions by providing support for next recommended clearances. Hence, these systems require knowledge of – and input from – their human operators such as given clearances.

Normally, manual input from controllers is necessary. SR can perform the input task automatically by analyzing the radio telephony channel between controller and pilot. The controller only has to check the correctness. This kind of procedure leads to less workload [20].

D. Multimodality

Although the definitions of multimodality differ greatly in literature, there is general consensus that multimodal systems involve multiple senses of the operating human or multiple sensors of the corresponding machine. For example, the European Telecommunications Standards Institute (ETSI) defines the term "multimodal" as an "adjective that indicates that at least one of the directions of a two-way communication uses two sensory modalities (vision, touch, hearing, olfaction, speech, gestures, etc.)" [21].

A multimodal Thales demonstrator called "Shape" already includes eye tracking, a multi-touch device, and voice recognition [22], [23]. However, for example, speech recognition is only used to detect flight numbers. Only controller commands given through the tactile surface as a whole seem to be uplinked to the pilot.

Within a bachelor thesis [24] at DLR the first approach for a multimodal ATC demonstrator was undertaken to integrate the three modalities eye tracking, multi-touch and speech recognition. The main aim of this thesis was to implement and evaluate eye tracking as an input modality for a multi-modal controller working position. For this purpose an existing concept consisting of multi-touch and speech recognition [25] was enhanced by integrating eye tracking in order to enable natural and fast selection of aircraft.

The findings gained from the investigations carried out for that thesis showed that eye tracking is a useful and well accepted input modality when it is accompanied by other intuitive modalities.

However, those modalities should go beyond just serving as a "pointing device" for elements on a screen like many eye tracking applications.

III. CONCEPT OF MULTIMODAL AIR TRAFFIC CONTROLLER INTERACTION

The motivation for building a prototypic multimodal CWP for approach controllers is based on presumed advantages of multimodal interaction (see chapter I) and promising research results gained from the previously developed unimodal and multimodal prototypes (see chapter II).

TriControl focuses on integrating the three most promising interaction technologies: speech recognition (SR), sensing of multi-touch gestures (MT), and eye tracking (ET) (see [26]). However, these modalities can be combined in a number of ways with respect to the three basic elements of a controller command (aircraft identifier (A), command type (T), and command value (V)). Furthermore, in former investigations some modalities were found to be more suitable for certain standardized command parts than others (Figure 2).

Modality	Aircraft	Type	Value
Speech Recognition	medium	medium	good
Multi-Touch	medium	good	medium
Eye Tracking	good	poor	poor

Figure 2. Matrix with suitability assessment of input modes (SR, MT, ET) with respect to controller command elements aircraft (A), command type (T) and value (V)

Figure 2 shows the favored assignment between input modality and command element. To identify the aircraft (A) that will receive the next command (e.g. DLH123) three possible ways are explained: uttering the callsign (A-SR), touching on its radar target/label on the situation representation or an auxiliary touch display (A-MT), or looking at the radar target/label (A-ET). The command might be transferred to pilots by data link or a text-to-speech interface.

Speech recognition rates of callsigns are good, but it takes some time to utter the whole callsign. Although in previous investigations the direct touch on an aircraft representation was assessed as easy and intuitive, the hand covers the radar screen and hence the traffic situation below. To guarantee a good overview of the whole traffic situation use of a second screen could solve the problem but would create a new issue in that the active gaze has to switch from one screen to the other and back.

However, the controller normally looks at the intended aircraft anyway. Hence, eye tracking seems to be the most convenient option for selecting the first controller command part, just as naturally as one usually makes eye contact in a face-to-face conversation.

Analogue to the aircraft identifier, the three input modalities are also discussed for the command type. SR (T-SR) would recognize International Civil Aviation Organization (ICAO) phraseology conform command types (T) quite well due to the limited search space of different types (e.g. *reduce*, *descend*, etc.). Selecting a type by eye movement (T-ET) – for example from a menu – will be tiring for the human operator as it requires unnatural and active control of gaze. In a human conversation, hands are also used to describe a general direction via gestures. Similarly, a multi-touch device can be used to draw a simple one- or more-finger gesture that is recognized very accurately to code a command type (T-MT).

Three different modalities also exist for entering command values. Selecting exact command values (V) (e.g. 210) with swiping gestures (V-MT), for example, on a visual scale can be difficult. Looking at values in certain menus (V-ET) is as exhausting as selecting command types with one's eyes. However, just uttering the short values works fast and is intuitive (V-SR).

From former investigations we derived a classification of the input mode suitability (poor-medium-good) in terms of a color-coded matrix (see Figure 2). This matrix depicts an initial point for implementation of TriControl (A-ET, T-MT, V-SR).

The chosen combination of modalities for TriControl enables the input of the three most common elements of a controller command (aircraft identifier, command type and value). The number "3" is spoken as "tri" (pronounced as "tree" in English) in radiotelephony to improve the understanding of digits even in bad speech quality, hence the name TriControl was used for the interaction design.

To generate a controller command, the operator has to focus on an aircraft on the radar situation display with his eyes, make a specific gesture on the multi-touch device, and utter a corresponding value for this type of command (see Figure 3 for the setup of modalities and Figure 4 for an example command).

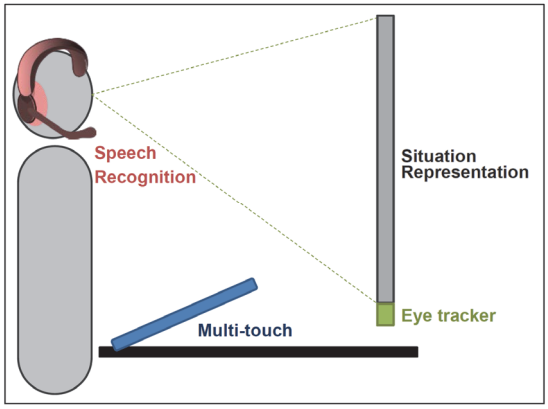


Figure 3. Interaction modalities of TriControl CWP

The information processed by the TriControl CWP is put together as a clearance shown in the far right box of Figure 4.

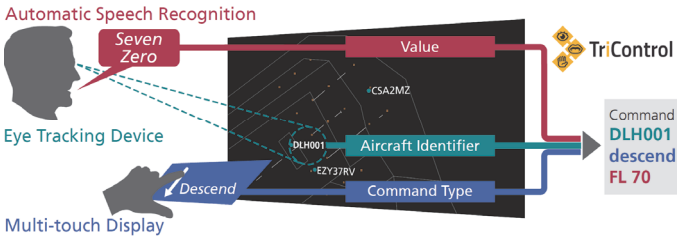


Figure 4. Schematic view of the communication modes and processed information

Our instantiated controller working position may be used by a feeder or pickup approach controller. One of the main tasks of approach controllers is monitoring the radar situation display. Within TriControl it is assumed that the aircraft radar label being looked at by the controller is the focus of attention. Eye gaze measurement is used to continuously calculate the position of the air traffic controller's eye gaze and correlate it with aircraft label positions on the display.

For our demonstrator we used DLR's radar screen software RadarVision [27] showing Düsseldorf airspace. It provides data about the position of aircraft icons, labels, and significant points (runway threshold, initial approach fixes, and waypoints). In TriControl, eye tracking enables aircraft radar labels to be selected as the first part of a command without losing focus of the traffic situation.

In our demonstrator we use a contact-free Tobii [28] infrared eye tracking device (Tobii EyeX Controller) which is mounted at the bottom of a monitor. Calibration is necessary prior to adapting eye tracking quality to people with contact lenses, glasses, or without corrected vision. The position of the user's pupils is followed regarding a standardized screen position set and connected to the display size by the manufacturer software.

Using the resulting display coordinates of the spot being looked at by the user in front of the display, we determine whether an aircraft icon or radar label is displayed. A dwell time of nearly one second was defined as the threshold for highlighting the currently focused aircraft label with a white frame. Otherwise, the controller could be distracted if the highlighting frame jumps around the whole screen, thereby indicating non-intended gazes, particularly while scanning the traffic situation.

Although, radar labels do hardly overlap in the TMA due to lateral aircraft and therefore label separation, manual or automatic deconflicting is possible to select intended aircraft safely via eye tracking. As a safety feature, the controller might fall back to selecting the aircraft callsign from a list on a multi-touch device.

In combination with two-dimensional gestures on a multi-touch display, the controller can add the type of a command to the selected aircraft to start insertion of a clearance.

The controller selects the type using a set of four single- and dual-touch gestures on a tablet – altitude, speed, or heading of the aircraft for example. The direction of the gestures indicates whether the aircraft should, for example, accelerate or decelerate.

Specifically to avoid head-down times the gestures and tablet usage are designed simply and intuitively. Furthermore, the user may perform all gestures at any location on the multi-touch screen while still focusing on the situation representation. We used a standard Wacom [29] multi-touch tablet in our demonstrator.

For the design of specific gestures typical natural gestures and well-known gestures from smartphone use were analyzed and assessed for the use in ATC. So, a one-finger swipe from left to right is recognized as *increase*, the opposite direction is recognized as *reduce*. A one-finger swipe from top to bottom indicates a *descend*, the opposite direction a *climb*. One finger held for more than one second pressed on any point of the multi-touch device is interpreted as a *direct-to* gesture. This gesture is also used for ILS clearance and handover to the next following responsible controller position, but requires different speech input compared to waypoints. Drawing a sector of a circle to the left or right with a two-finger gesture – either using two fingers of either hand – initiates a *heading* command. The multi-touch software evaluates the controller's gesture that results in "*reduce / increase [or-more / or-less], descend / climb [or-above / or-below], turn-right-heading / turn-left-heading, direct-to, handover, cleared-ILS, intercept-localizer*". Thus, the controller inserts the second part of his command – the type – via the haptic modality. If the multi touch device failed, a redundant method was implemented for safety reasons. The commands can also be entered by pressing device hardware respectively software buttons.

For the third and last part of the command – the value – the auditive modality is used. TriControl incorporates specific algorithms to detect spoken values such as numbers. By pressing a foot switch, recording of the subsequent utterance is started. The streamed audio file is the input for an automatic speech recognizer developed by Saarland University (UdS) and DLR [30]. For TriControl this speech recognizer is configured to analyze only command values without value units. There is a broad range of valid value types. The controller is allowed to speak between one and three consecutive digits ("zero, one, two, tree, four, five, six, seven, eight, niner"). For full multiples of ten, hundred or thousand, double numbers ("ten, twenty, thirty,…"), triple numbers (e.g. "two hundred") or a quadruple number (e.g. "four tousand") can be spoken. The system also recognizes special speed phrases ("own discretion, no restriction, minimum clean, final approach"). The speech recognizer accepts keywords for other clearances, e.g. inserting a handover by saying "tower", ILS clearance with the runway name e.g. "two tree right", or a *direct-to* command by a waypoint name in the Düsseldorf airspace ("Bottrop, Metma, Regno, Delta Lima 454 and so on"). Alternatively, values might also be selected from a software menu on the multi-touch device in cases of failure.

The value should of course correspond to a reasonable type to complete all three command parts. When all three modalities have been used, the TriControl system merges SR, ET, and MT data and displays it on the RadarVision screen. The whole command is presented, then to be validated by the controllers. For this visualization, five grey cells have been added to all aircraft radar labels (Figure 5). These cells represent five different command types. Cell one includes flight levels and altitudes, whereas all speeds in knots or Mach are presented in the second cell. The third label line contains the remaining three display areas for other current clearances. Headings, relative turns, waypoints, or transitions are shown there. Cell four includes rates of descent or climb. Cell five contains miscellaneous entries such as handover, ILS and localizer clearance, or holding.

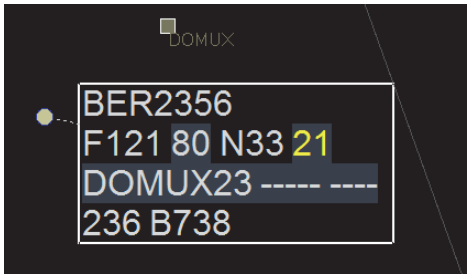


Figure 5. Interactive aircraft radar label cells in RadarVision

The completed command is shown as exactly one yellow value (command part three), in one of the five grey type cells (command part two) at one specific aircraft radar label (command part one). If the data is correct, the controller has to validate the command with a finger tap on the green-only area of the multi-touch device. The yellow value then becomes white. If controllers do not want to confirm the yellow value, they can cancel the entries using a hardware button on the tablet at any time during the process.

If the second or third part of the controller command, i.e. the type or a value, is selected after focusing on an aircraft label the eye tracking feature is locked. The purpose of this feature is to reduce unintentionally assigning command parts to other aircraft. Nevertheless, it is always possible to overwrite command type and value until the whole command has been confirmed or rejected. Furthermore, the obligatory multimodal activation (all three modalities are needed) enables the controller to freely look around without entering commands accidentally into the system.

After the command has been confirmed it will be sent to the aircraft. To yield the desired benefits in communication efficiency this may be done via reliable and fast data link connection. Even though data link technology with CPDLC (controller–pilot data link communications) protocol is now operational at many CWPs, most information exchanges between air traffic controllers and pilots still use voice communication owing to insufficient reliability in data link transfer speed. TriControl is designed to enable use of this digital connection by eliminating the speed bottleneck of human data input.

However, for reasons of compatibility with older aircraft equipment and the migration process from traditional communication, the concatenated command can also be sent via text-to-speech over the conventional radiotelephony channel. The pilot would then only experience a change to a more artificial and standardized voice that always sounds the same.

The following example explains how to insert the three parts of a command similarly to Figure 6.

Figure 6. Multimodal prototypic CWP exhibit TriControl

Figure 6 shows TriControl in a state where the three modalities have been used to insert data into the system: eye tracking (gaze on aircraft radar label of BER8411), multi-touch (two-finger circle-sector gesture indicating command type *heading*), automatic speech recognition (utterance of "two hundred"), situation data display (Düsseldorf approach area), and the resulting yellow input value (200 in grey "direction" cell) in aircraft label before validation of the controller command.

To reach this state the controller firstly looks at the BER8411 label on the radar situation display for nearly one second (see Figure 7).

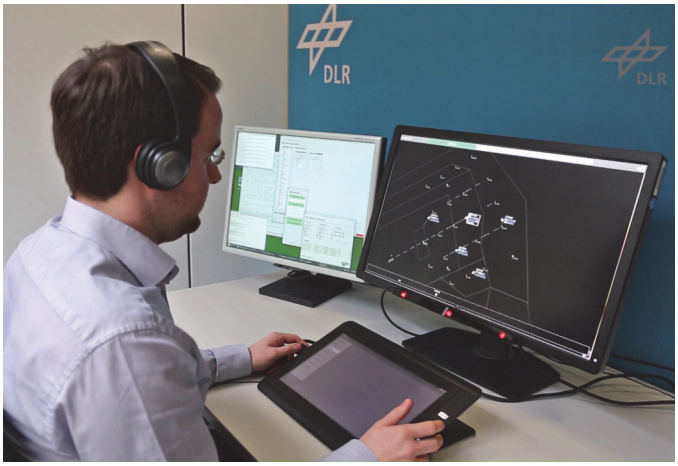


Figure 7. TriControl setup with radar display attached eye tracker, headset, and multi-touch device

This may be achieved very naturally by the controller merely checking the label of the aircraft that is to be addressed. In this way the aircraft with the given callsign is selected as the aircraft which is to receive a new command.

Secondly, the controller touches the multi-touch device with two fingers, rotates them on the screen, and lifts his fingers again. This is understood as a *heading* gesture.

Thirdly, the controller presses the foot-switch and says "two hundred" using his headset. SR evaluates the speech and will deliver "200" as a result. All three parts of the command are concatenated to "BER8411 *heading* 200", which means that flight Air Berlin 8411 must turn its heading to 200 degrees.

As the three interaction modes can be used simultaneously, the air traffic controller's intention is entered into the ATC system fast. In our opinion, the controller will roughly need only one third of the time needed to utter the whole command with its three parts "air berlin eight four one one turn heading two hundred". The time needed to utter "two hundred" is simultaneous to the heading gesture and looking at the aircraft radar label BER8411.

In addition, the unilateral workload for verbal communication will be reduced and balanced with other modalities. It greatly relieves the strain on the voice from talking. The reduction in the total time needed to issue one command frees up the controller's cognitive resources.

This may even result in higher mental capacity and more efficient work if controllers can manage more aircraft at a time through reduced communication contact times. This could then also increase air traffic operational capacity.

IV. Evaluation of Multimodal CWP Demonstrator *TriControl*

For preliminary evaluation of the multimodal system usability we gathered structured feedback data of fair guests at DLR's World ATM Congress 2016 booth in Madrid who 'worked' intensively on our exhibit and agreed to participate in the inquiry. The survey comprised ten items from the System Usability Scale questionnaire (SUS) [14], three additional questions – one on each modality, and one summarizing item on the complete system.

Participants had to rate 14 statements on a Likert scale [31] from 0 to 4 meaning from "strongly disagree" to "strongly agree". The questionnaire items consisted of seven positively/negatively formulated statements. Twelve people (many of them air traffic controllers) took part in the survey.

A SUS score between 0 and 60 indicates poor usability; good usability starts at just over 75, becoming excellent the closer the score gets to 100.

The average SUS score in our survey was 79 and no single participant score was below 60. Two participants even rated usability with a SUS score of 90. Hence, usability of the whole multimodal CWP can be assumed as good.

The worst single item rating (2.9) was obtained for the SUS question on "Frequent Use", with the best (3.3) being obtained for questions on "Simplicity of Use" and "Using without Training" (see Figure 8). Black bars indicate the standard error as the quotient between variance and square root of sample size.

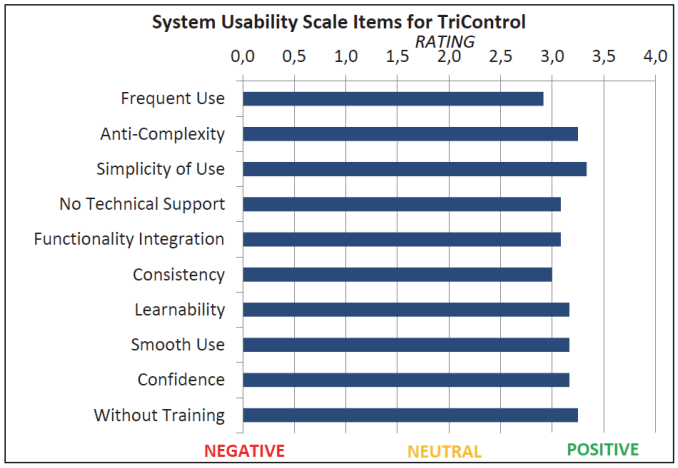


Figure 8. Participants ratings on system usability questionnaire items

A few minutes of exhibit use are not sufficient to contemplate a steady and more restricted use compared to current CWPs. However, simplicity was rated best (3.3). This demonstrates the clarity and intuitiveness of the multimodal concept.

All other item ratings lay between 3.0 and 3.3 (inversion of negatively formulated statement ratings for better comparability) showing good usability for different aspects of the prototypic multimodal CWP concept.

Multi-touch gesture recognition was rated best of the additional questions (3.3) (see Figure 9). Hence, after a quick training phase, the four different gesture types proved to be easy to remember and apply. Speech recognition of command values (3.1) worked very well for most participants. However, in a handful of speakers accents led to slightly lower recognition rates and demand for adjustments.

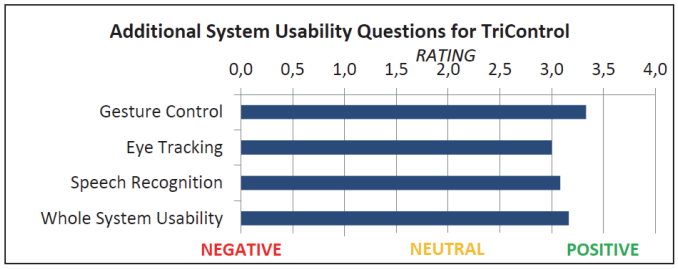


Figure 9. Participant ratings on additional system usability questions

Eye tracking was the most interesting and surprising modality being integrated. It worked fairly well for people without vision correction, with contact lenses, or glasses. Recalibration for each individual participant improved the eye tracking feature for aircraft labels.

Nevertheless, after changing seat settings or head position, the feature needed clearer gazes to react and was rated at 3.0, which is still within the good usability range. The great majority of participants were positively impressed by the overall performance of all three integrated modalities as reflected by the rating 3.2.

V. SUMMARY AND OUTLOOK

TriControl was the first ATC interaction demonstrator to combine eye tracking, multi-touch gestures, and speech recognition to generate full-featured controller commands. Dozens of air traffic controllers from roughly twenty different countries and all continents tested the TriControl exhibit in addition to those who took part in the survey. The feedback was broadly unanimous: training is needed, especially for simultaneously use of all modalities, but thereafter interaction is intuitive, fast, and straightforward.

The training need includes "handling" different devices simultaneously without looking at them. The training effect is similar to the difference between new and experienced drivers. Drivers have to manage different foot pedals, gearshift, steering wheel, indicator lights and other on-board equipment, whilst constantly watching the traffic, other drivers, pedestrians, road signs, etc. Thus, there is a general consensus that it should be fairly easy to acquire multimodal ATC interaction skills.

ATC experts also encourage further investigations into the advantages and drawbacks of multimodal interaction for controllers. Hence, other combinations of interaction modalities should be tested and compared. As ATC must satisfy stringent safety standards, the reliability and accuracy of these input modes must be very high.

In order to accommodate individual differences and preferences, the next phase anticipates allowing users to choose the modalities that they wish to use to interact with the system. Different extracts from the complete three times three matrix, comprising interaction modalities and controller command parts (see Figure 2), will be implemented in an enhanced version of TriControl. The speed gain for command input and interaction with TriControl should be measured against conventional systems. With rapid development of other innovative input technologies by the consumer industry, the mentioned matrix may grow further. When expanding this matrix to a tensor including parameters like personal preferences or variations over user workload even more combinations could be investigated.

Furthermore, each of the devices for interaction may be changed. The low-cost eye tracker could be replaced by a camera system, tracking head and eye position to improve accuracy and allow for greater freedom of body positions while working (see Figure 10).

A number of following studies shall prove and improve various aspects of TriControl. First, operational feasibility and suitability to controllers' requirements will be investigated.

Afterwards, experiments concerning user acceptance, usability, and related operational improvements will be performed and evaluated. Finally, capacity and safety will be analyzed. To this end, we will identify conditions for better and safer use of certain modalities.

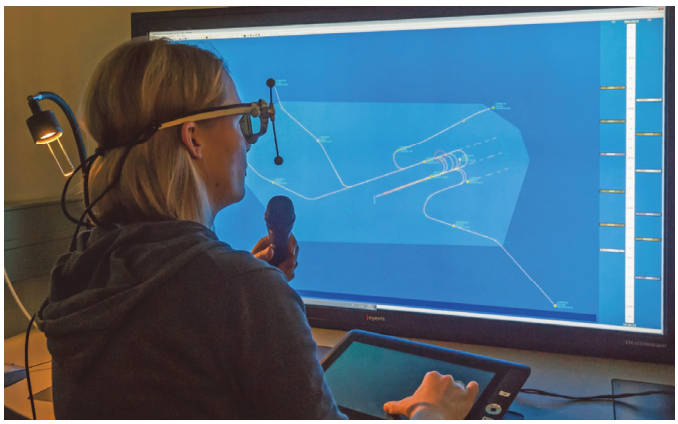


Figure 10. Advanced eye tracking device at CWP

With DLR's knowledge in CWP design and its validation infrastructure for executing realistic high-quality simulations, initial results and empirical evidence on the usefulness of multimodality for air traffic control will be gained in a continuative development phase to find out the best ways of achieving multimodal interaction.

REFERENCES

[1] SESAR, "The roadmap for delivering high performing aviation for Europe – European ATM Master Plan," Brussels, 2015.

[2] A. Labreuil, D. Bellopede, M. Poiger, K. Hagemann, M. Uebbing-Rumke, H. Gürlük, M.-L. Jauer, V. Sánchez, and F. Cuenca, "SESAR 10.10.02 D93, Innovation Analysis Report 2013," April 2014.

[3] AcListant® homepage: www.AcListant.de/wp.

[4] C. Möhlenbrink and A. Papenfuß, "Eye-data metrics to characterize tower controllers' visual attention in a multiple remote tower exercise," ICRAT, Istanbul, 2014.

[5] M. Uebbing-Rumke, H. Gürlük, M.-L. Jauer, K. Hagemann, and A. Udovic, "Usability evaluation of multi-touch displays for TMA controller working positions," 4th SESAR Innovation Days, Madrid, 2014.

[6] H. Gürlük, H. Helmke, M. Wies, H. Ehr, M. Kleinert, T. Mühlhausen, K. Muth, and O. Ohneiser, "Assistant based speech recognition – another pair of eyes for the Arrival Manager," 34th DASC, Prague, 2015.

[7] U. Ahlstrom and F. J. Friedman-Berg, "Using Eye Movement Activity as a Correlate of Cognitive Workload," in International Journal of Industrial Ergonomics 36: 2006, pp. 623–636.

[8] R. Alonso, M. Causse, F. Vachon, P. Robert, D. Frédéric, and P. Terrier, "Evaluation of head-free eye tracking as an input device for air traffic control," Taylor & Francis Group, Toulouse, France; Québec, Canada, 2012.

[9] L. E. Sibert and R. J. K. Jacob, "Evaluation of Eye Gaze Interaction," Proceedings of the CHI '00 Conference on Human Factors, in Computing Systems, New York, NY: ACM, 2000, pp. 281-288.

[10] P. Majaranta, I. S. MacKenzie, A. Aula, and K. J. Räihä, "Effects of Feedback and Dwell Time on Eye Typing Speed and Accuracy," in Universal Access in the Information Society 5: 2006, pp. 199–208.

[11] K. Kotani, Y. Yamaguchi, T. Asao, and K. Horii, "Design of Eye-typing Interface Using Saccadic Latency of Eye Movement," in International Journal of Human–Computer Interaction 26: 2010, pp. 361–376.

[12] Indra, "Advanced Controller Working Position," http://www.indracompany.com/sites/default/files/indra-indra_advanced_controller_working_position.pdf.

[13] D. Wald, "Implementation and evaluation of touch based air traffic control", original German title: "Programmierung und Evaluierung einer Touch basierten Flugverkehrskontrolle," Master Thesis, Langen, 2011.

[14] J. Brooke, "SUS - A quick and dirty usability scale," in Usability Evaluation in Industry, P. W. Jordan, B. Thomas, I. L. McClelland, and B. A. Weerdmeester, Eds. London, Taylor and Francis, 1996, pp. 189–194.

[15] SRI International, "Siri-based virtual personal assisstant technology," http://www.sri.com/engage/ventures/siri.

[16] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Google search by voice: A case study," in Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, Springer, 2010, pp. 61–90.

[17] C. Hamel, D. Kotick, and M. Layton, "Microcomputer System Integration for Air Control Training," Special Report SR89-01, Naval Training Systems Center, Orlando, FL, USA, 1989.

[18] D. Schäfer, "Context-sensitive speech recognition in the air traffic control simulation," Eurocontrol EEC Note No. 02/2001 and PhD Thesis of the University of Armed Forces, Munich, 2001.

[19] K. Dunkelberger and R. Eckert, "Magnavox Intent Monitoring System for ATC Applications," Magnavox, 1995.

[20] H. Helmke, O. Ohneiser, T. Mühlhausen, and M. Wies, "Reducing Controller Workload with Automatic Speech Recognition," 35th DASC, Sacramento, 2016.

[21] ETSI, EG, "Human Factors (HF); Multimodal interaction, communication and navigation guidelines," ETSI EG 202 191 V1.1.1, 2003, p. 7.

[22] Thales, "SHAPE - Innovative and Immersive ATC working position," https://onlineexhibitormanual.com/60atcaAnnual/PDF/Brochure_exhiReg477326_Shape.pdf, 2014.

[23] A. Schofield, "Thales' radical concept for controlling air traffic," in Things With Wings, http://aviationweek.com/blog/thales-radical-concept-controlling-air-traffic, 2014.

[24] P.-E. Seelmann, "Evaluation of an eye tracking and multi-touch based operational concept for a future multimodal approach controller working position", original German title: "Evaluierung eines Eyetracking und Multi-Touch basierten Bedienkonzeptes für einen zukünftigen multimodalen Anfluglotsenarbeitsplatz", Bachelor Thesis, Braunschweig, 2015.

[25] M.-L. Jauer, "Multimodal Controller Working Position, Integration of Automatic Speech Recognition and Multi-Touch Technology", Bachelor Thesis, Braunschweig, 2014.

[26] DLR Institute of Flight Guidance, "TriControl – multimodal ATC interaction", 2016, http://www.dlr.de/fl/Portaldata/14/Resources/dokumente/veroeffentlichungen/TriControl_web.pdf.

[27] O. Ohneiser, "RadarVision - Manual for Controllers", original German title: "RadarVision - Benutzerhandbuch für Lotsen", German Aerospace Center, Institute of Flight Guidance, Internal Report 112-2010/54, Braunschweig, 2010.

[28] Tobii homepage: www.tobii.com.

[29] Wacom homepage: www.wacom.com.

[30] H. Helmke, J. Rataj, T. Mühlhausen, O. Ohneiser, H. Ehr, M. Kleinert, Y. Oualil, and M. Schulder, "Assistant-Based Speech Recognition for ATM Applications", in 11th USA/Europe Air Traffic Management Research and Development Seminar (ATM2015), Lisbon, Portugal, 2015.

[31] R. A. Likert, "Technique for the Measurement of Attitudes," in Archives of Psychology 22, No. 140, 1932, pp. 5–55.

aerospace

MDPI

Article

Faster Command Input Using the Multimodal Controller Working Position "TriControl"

Oliver Ohneiser [1,*], Malte Jauer [1], Jonathan R. Rein [2] and Matt Wallace [3]

[1] German Aerospace Center (DLR), Institute of Flight Guidance, Lilienthalplatz 7, 38108 Braunschweig, Germany; Malte-Levin.Jauer@DLR.de
[2] Federal Aviation Administration (FAA), William J. Hughes Technical Center, Atlantic City International Airport, Egg Harbor Township, NJ 08405, USA; Jonathan.Rein@FAA.gov
[3] Deutsche Flugsicherung GmbH, Academy, Am DFS-Campus, 63225 Langen, Germany; Matthew.Wallace@DFS.de
* Correspondence: Oliver.Ohneiser@DLR.de; Tel.: +49-531-295-2566

Received: 6 April 2018; Accepted: 4 May 2018; Published: 8 May 2018

Abstract: TriControl is a controller working position (CWP) prototype developed by German Aerospace Center (DLR) to enable more natural, efficient, and faster command inputs. The prototype integrates three input modalities: speech recognition, eye tracking, and multi-touch sensing. Air traffic controllers may use all three modalities simultaneously to build commands that will be forwarded to the pilot and to the air traffic management (ATM) system. This paper evaluates possible speed improvements of TriControl compared to conventional systems involving voice transmission and manual data entry. 26 air traffic controllers participated in one of two air traffic control simulation sub-studies, one with each input system. Results show potential of a 15% speed gain for multimodal controller command input in contrast to conventional inputs. Thus, the use and combination of modern human machine interface (HMI) technologies at the CWP can increase controller productivity.

Keywords: air traffic controller; human machine interaction; human computer interaction; multimodality; eye tracking; automatic speech recognition; multi-touch gestures; controller command; speed gain

1. Introduction

Multimodal human-computer interaction (HCI) may enable more efficient [1,2] and especially natural "communication" because "natural conversation" is a complex interaction of different modalities [3]. It can also be seen as a "future HCI paradigm" [4]. The term "multimodal" can be defined as an "adjective that indicates that at least one of the directions of a two-way communication uses two sensory modalities (vision, touch, hearing, olfaction, speech, gestures, etc.)" [5]. Multimodal interaction (MMI) is interpreted as combining "natural input modes such as speech, pen, touch, hand gestures, eye gaze, and head and body movements" [6]. In addition, different types of cooperation between modalities in such systems can be used e.g., inputs from different channels might be redundant or need to be merged [7].

One advantage of multimodal systems is the flexibility due to alternative input modes, which also avoids overexertion and reduces errors [6]. Another benefit lies in the support of different types of users and tasks [6]. MMI also promise to be easy to learn and use, as well as being more transparent than unimodal interaction [8]. Furthermore, the operator load as a whole can be shared across all individual modalities [9]. Thus, the fusion of data with their origin in different input modalities is one essential factor for an effective multimodal system [10].

Even if some of the different modalities that are combined to a multimodal system are error-prone, the multimodal system normally is more robust against recognition errors [6]. In Oviatt's study, this is due to the users' intelligent selection of the best input mode for the current situation [6].

For many domains, the typical modalities to be used multimodally are speech recognition, eye gaze detection, and gestures [11], which are also applicable for the air traffic management (ATM) domain. In the course of SESAR (Single European Sky ATM Research Programme) it is necessary to integrate new technologies such as speech-, gaze-, and touch-inputs for an improved interaction between controllers and their system [12]. The possible speed and efficiency gain with multimodal interaction working with our CWP prototype TriControl is the key topic of this paper.

Section 2 outlines related work on multimodal systems. Our multimodal CWP prototype is described in Section 3. The study setup, methods, and participant data are presented in Section 4. The results of our usability study are shown in Section 5 and discussed in Section 6. Finally, Section 7 summarizes, draws conclusions, and sketches future work.

2. Related Work on Multimodal Human Computer Interaction

Different domains investigated the benefits and drawbacks of multimodal systems in the past. This section outlines multimodal prototypes and some important results on human performance when working multimodally.

2.1. Examples of Multimodal Interaction Prototypes

In recent years a variety of multimodal interfaces have been developed. MMI can support education for disabled people using gestures and sound [13]. Using MMI in a car, the driver may choose his/her preferred modality from speech, gaze, and gestures, and can combine the respective system input with different modalities [14]. Another MMI system connected to the steering wheel of a car enables input via speech and gestures [15]. A further example is the multimodal combination of gestures and voice to place items on a screen [16].

There are even a lot of examples in the air traffic domain. In a part-task flight simulator, the MMI allows for function control via eye gazes in combination with speech recognition [17]. Another air traffic control multimodal prototype incorporates pen and touch interaction as well as physical paper for flight strips [18]. The users of this system were able to get along with the MMI very quickly and did not feel overstrained. Eye tracking in combination with a mouse can be used for modification of air traffic control (ATC) radar display settings by the user as well [19,20].

In previous DLR developments, the use of eye tracking as an input device was also conceptually enhanced with speech recognition and multi-touch sensing for use in ATC [21,22]. DLR successfully evaluated the underlying unimodal prototypes for speech recognition [23], multi-touch [24], and eye tracking [25]. More MMIs related to air traffic management such as flight strip manipulation or an en-route interface can be found in [9].

2.2. Findings of Earlier Multimodal Interaction Studies

In findings of Oviatt, users were more likely to work multimodally if commands dealt with numbers or orientation of objects [26]. These aspects are to a certain extent also true when controlling aircraft on a radar display. However, the amount of multimodal overlap between manual input by hand and spoken utterances does vary heavily depending on the individual [6].

In a study on driving, the use of a buttons-only system was faster than the combination of gestures and speech interaction, but less visual demand and a comparable performance were reported for the multimodal system [15]. In a study concerning interactive maps, roughly 95% of users working with spatial data preferred interacting multimodally and a speed gain of 10% for system inputs was shown [27]. A preferred multimodal interaction was also found for a map-based military task, with a 3.5-fold improvement in error handling times [28].

3. Description of TriControl Prototype

In everyday communication with each other, we use multiple ways of transferring information, as for example speech, eye contact, and manual gestures. The same principle is the underlying idea of the DLR multimodal CWP prototype "TriControl". Therefore, it combines speech recognition, eye tracking, and multi-touch sensing, with the goal of enabling a faster and more intuitive interaction especially for approach controllers. The following section will outline the multimodal interaction philosophy of the prototype. A more detailed description including technical details can be found in [29].

The main task of the target user is the transfer of controller commands to the pilot and the read-back check [30,31]. Currently, these commands are mostly transmitted via radiotelephony (R/T) using standard phraseologies according to ICAO (International Civil Aviation Organization, Montreal, QC, Canada) specifications [32]. Furthermore, controllers are required to log the given commands into the aircraft radar labels respectively electronic flight strips of the CWP. This input is normally performed manually using keyboard and mouse. Regarding the structure of the commands, they are usually composed of "callsign–command type–value", e.g., "BER167–descend–FL60". Using TriControl, these three components are distributed over three input modalities, with each component assigned to a modality based on its suitability of transporting the corresponding piece of information.

As a result, the current two unimodal communication channels are replaced by one multimodal interaction with the CWP as visible in Figure 1. Each of the used input modalities of TriControl is presented in the following. When the multimodal input is completed, the combined controller command is logged by the system (no additional flight strip compiling required) and will then be transmitted from the CWP to the pilot via data-link. In case of failure in the submission process, the R/T communication may be used as a fallback solution. Although the TriControl concept is designed to enable a fast and efficient input into the CWP (intended for data-link transmission), the usage of text-to-speech could be a viable alternative for the transmission of commands entered into the system to aircraft in the traffic mix without data-link capability. As the concept focuses on the input method of the controller commands, future investigations are required regarding the integration of the pilots' read-back into the concept. Exemplarily, a digital acknowledge of the pilot sent via data-link could be visualized on the radar screen, or the usual read-back via R/T could be applied.

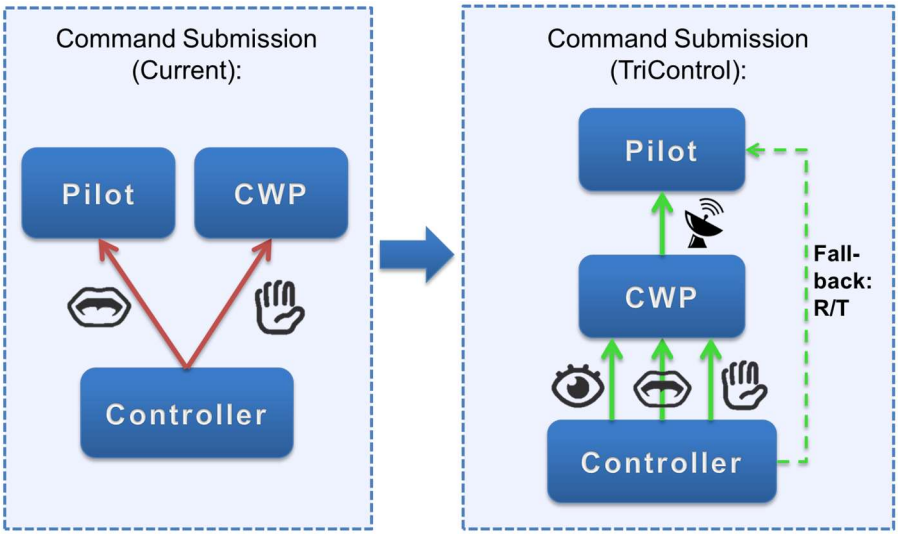


Figure 1. Difference between current controller command submission and TriControl command submission.

3.1. Eye-Tracking

TriControl's integrated eye tracking device determines the user's visual focus in order to detect the aircraft targeted by a controller command. In the same way as we address our conversational

partners by eye contact in daily life, TriControl users can address aircraft on the radar screen just by looking at the radar labels or head symbols.

Therefore, the first part of the controller command—the callsign—is selected using the visual modality. To emphasize the selection of an aircraft, the corresponding label is highlighted by a white box around it as shown on the radar display in Figure 2 [33].

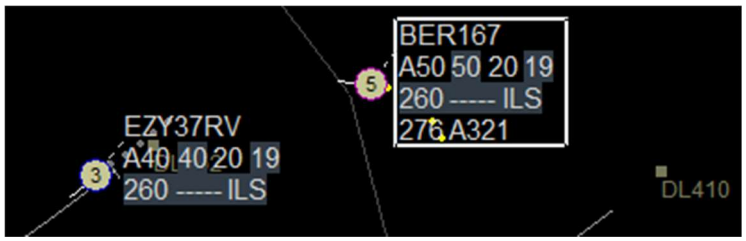


Figure 2. White box around the currently selected aircraft as shown in the radar display.

3.2. Gesture Recognition

If someone asks us how to navigate to a location, we automatically use hand gestures to describe directions. These coarse-grained pieces of information (e.g., "left", "right", "straight ahead", etc.) resemble the command types of a controller command like "climb", "descend" and so on. Thus, TriControl uses gesture recognition on a multi-touch device for the insertion of the command type. In the current prototype, the following gestures are implemented to insert the corresponding command types: swipe left/right for reduce/increase of speed; swipe up/down for a climb/descend in altitude; rotate two fingers for a heading command, and lastly long-press one finger for a direct/handover/cleared-ILS command, where the final type depends on the value of the complete command (e.g., a value of "two three right" would correspond to a runway, therefore the type would be interpreted as cleared-ILS).

Additionally, we implemented a few convenience functions for human machine interface (HMI) manipulations. Thus, the controller is able to change the zoom factor and visible area of the radar display by a 5-finger gesture: movement of all fingers moves the visible area and a spreading/contraction of the fingers zooms in/out the map section. TriControl also offers the display of distances in nautical miles between two aircraft that are selectable by multi-touch, i.e., by moving two fingers on the multi-touch screen to position two cursors on the main screen at one aircraft each.

3.3. Speech Recognition

Regarding discrete values, a natural choice to transmit this kind of information is by voice. Normally, when trying to insert specific values using different modalities, e.g., mouse, gestures, or eye gaze, the solution would most likely require some kind of menus that can be slow to search through. Automatic speech recognition is capable of converting spoken text or numbers into digital values, especially when the search space is limited—in this case limited to relevant values in the ICAO standard phraseology (e.g., flight levels, speed values, degrees, or waypoint names). TriControl therefore accepts spoken values from recorded utterances initiated by using a push-to-talk foot switch to complete the third element of a controller command in a natural and convenient way.

3.4. Controller Command Insertion

As an example, the insertion of the command "DLH271 reduce speed 180 knots" into the system is shown in Figure 3.

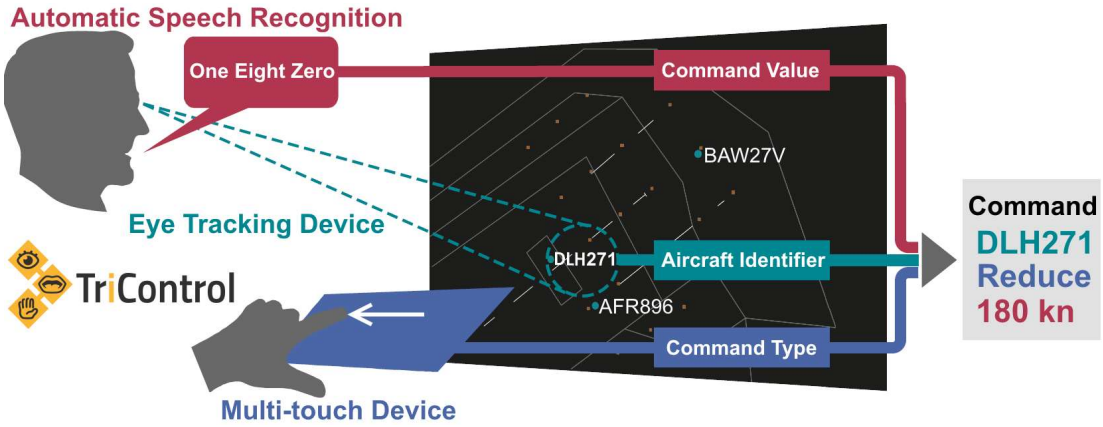


Figure 3. Interaction concept of TriControl with fusion of three modalities adapted from [34].

For this task, the controller firstly focused the corresponding label with the eyes. Then, the command type was inserted via swipe-left gesture and the value "one eight zero" was spoken. The order of uttering and swiping did not matter. It was even more efficient to perform it in parallel. When the controller began gesturing or speaking, the selected callsign locked so that the controller was already free to look somewhere else other than the label while completing the command with the other modalities. When all parts of the controller command were inserted into the system, a single tap on the touch screen confirmed the new clearance, which was also highlighted in the label. If command type or value was recognized or inserted incorrectly, it could be overridden by a new insertion before the confirmation. Alternatively, the whole command could be refused at any time using a cancel button on the multi-touch device as a measure of safety against unwanted insertion of clearances.

In contrast to the traditional method of inserting all command parts via voice when using radiotelephony or speech recognition of the whole command [23], this splitting of command parts to three modalities was envisaged to enable a timely overlap of the separate inputs. Because of this potential for concurrent input of the controller command parts and the marginal amount of time needed for the callsign selection (as the controller would look at the label anyway), the input method was designed to enable a faster, more natural and efficient insertion of commands into the system, while each modality was well suited for input of the respective type. A quantitative analysis of how much speed can be gained will be presented in the next section.

4. Usability Study with Controllers at ANSP Site

For a quantitative evaluation on how controllers use TriControl, we prepared the generation process of log files from different parts of the system. The log files capture:

- Positions of aircraft icons, aircraft radar labels, and waypoints as shown on the radar display;
- Eye gaze data with timeticks and fixation positions;
- Begin, update, and end of a touch gesture, with timeticks including confirmation gesture or rejection button press and release times;
- Press and release times of the push-to-talk foot switch (related to length of controller utterance);
- Timeticks for appearance of callsign, command type, and command value on the radar display.

4.1. Recorded Timeticks during TriControl Interaction

Eleven points in time (timeticks) were really relevant to judge speed differences during interaction: E1/E2/E3 (Eye Fixation), G1/G2/G3 (Gesture), S1/S2/S3 (Speech), and C1/C2 (Confirmation) as shown in Figure 4.

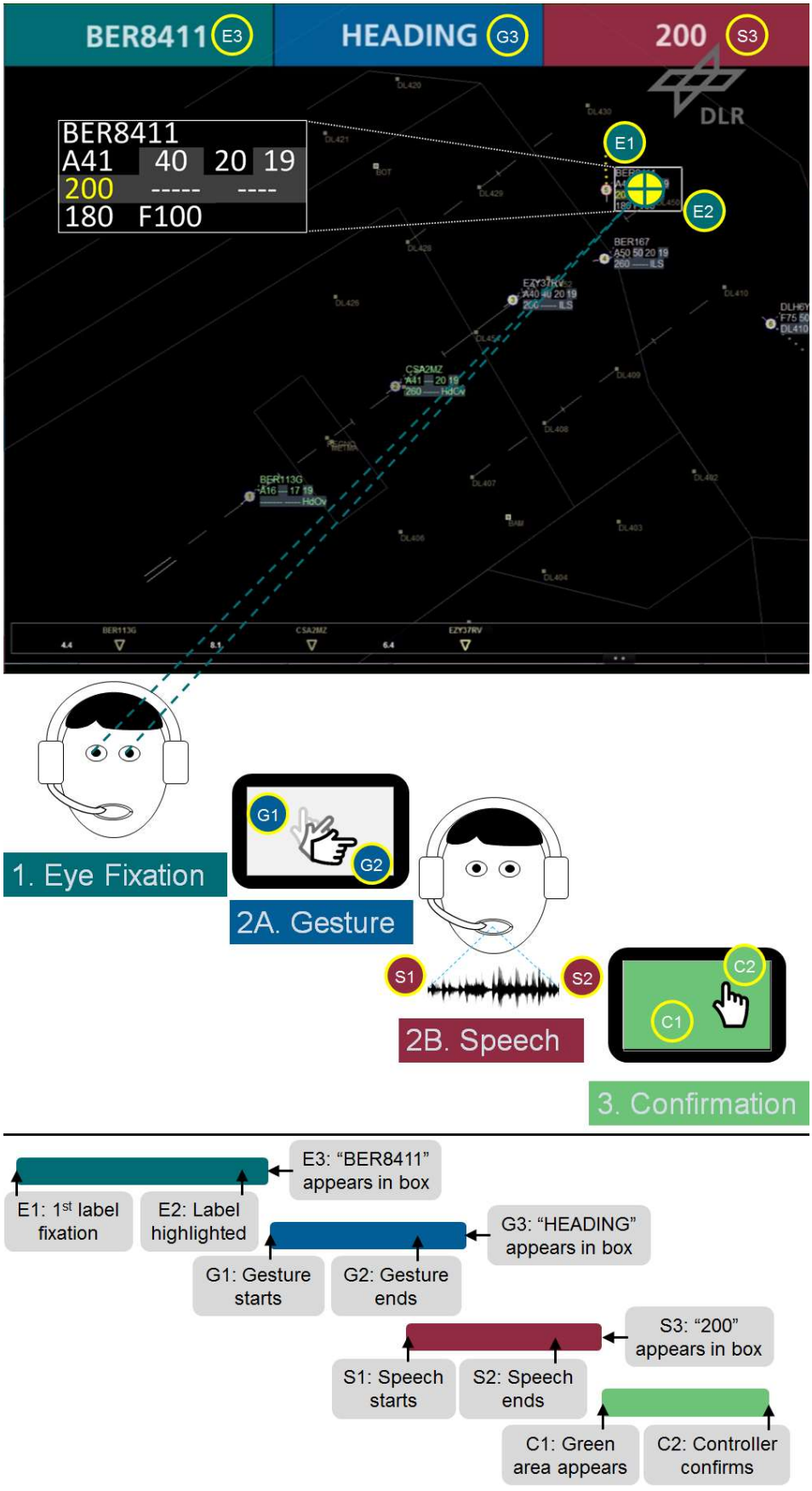


Figure 4. Measuring modality interaction and confirmation times when using multimodal controller working position (CWP) prototype TriControl.

The command input starts with the first eye fixation (1) of an aircraft icon or radar label. In the example of Figure 4, the label of "BER8411" was fixed by the participant's gaze (E1). If the captured gaze points are located within an area of around 0.2 percent of the screen for at least 20 ms, the system highlights the label with a white frame (E2) and presents the callsign in the dark green upper left box (E3).

Afterwards, the gesture phase (2A), the speech phase (2B), or both phases in parallel may start. In our example, the participant touched the multi-touch device (G1), performed a rotating "HEADING" gesture, and lifted his fingers again (G2). The gesture-recognizer module evaluated the gesture type and visualized it in the blue upper middle box (G3). However, the gestures may have already been recognized during the gesture update phase if further finger movements did not change the gesture type anymore. Hence, G3 could also happen before G2. The later timetick was valid for our calculations.

In our example, the participant pressed the foot switch to activate the speech recognition (S1) before ending up the gesture. After having spoken the value "Two Hundred" or "Two Zero Zero" (the system supports ICAO phraseology conform inputs as well as occasionally used variations), the controller had to release the foot switch (S2). The duration between S1 and S2 was also compared to the automatically evaluated length of the corresponding audio files in milliseconds. In the current prototype version, the utterance "200" was only analyzed after the foot switch release so that the recognition result appearance time (S3) in the red right upper box was always after S2.

As soon as all three command elements (callsign, command type, and command value) were available and presented in the three colored top boxes, two other elements appeared. First, a yellow "200" was displayed in the direction cell of the corresponding aircraft radar label (first cell in line 3). Second, the whole touchable area of the multi-touch device turned green to ask for confirmation of the generated command (C1). As soon as the single touch to confirm the command was recognized (C2), the command was entered into the system and the yellow "200" in the radar label turned white. The controller always had the opportunity to cancel his inputs by pressing a hardware button on the right side of the multi-touch device. This interaction was not considered as a completed command. After the confirmation event C2, the command insertion was considered as finished.

4.2. Multimodal Interaction Study Setup

From 4–6 April 2017 we conducted an interaction study using the multimodal CWP prototype TriControl at the German Air Navigation Service Provider (ANSP) DFS Deutsche Flugsicherung GmbH in Langen (Germany). 14 air traffic controllers from DFS took part in the study. The average age was 47 years (standard deviation = 10 years; with a range from 29 to 62 years). The ATC experience after finishing apprenticeship was an average of 21 years (standard deviation = 12 years; with a range from 7 to 42 years). Some of the older controllers were already retired. The current controller positions were Approach (7xAPP), Area Control Center (5xACC), Upper Area Center (2xUAC), Tower (4xTWR), and one generic instructor. Several controllers had experience in multiple positions. Eight participants wore glasses, two contact lenses, and four took part without vision correction. Two of those fourteen controllers were left-handed. Depending on the modality, some participants had previous experience with eye tracking (5), gesture-based interaction (3), or speech recognition (10). Controllers' native languages were German (9), English (3), Hindi (1), and Hungarian (1).

The simulation setup comprised a Düsseldorf Approach scenario using only runway 23R. There were 38 aircraft in a one-hour scenario without departures. Seven of them belonged to the weight category "Heavy", all others to "Medium". Each participant had to work as a "Complete Approach" controller (means combined pickup/feeder controller in Europe and combined feeder/final controller in the US, respectively). After approximately 15 min of training using a training run with less traffic density, a 30-min human-in-the-loop simulation run was conducted (see Figure 5).

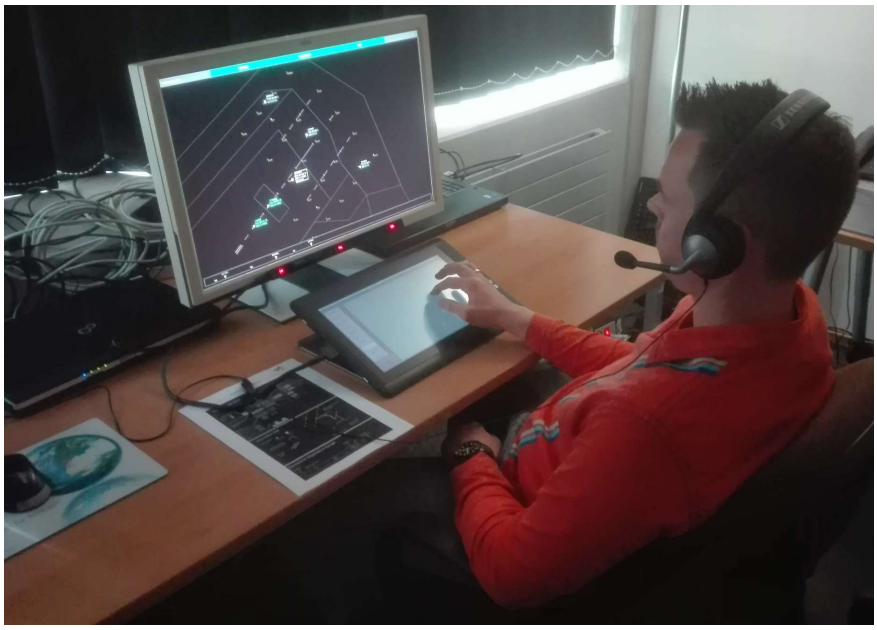


Figure 5. Participant during TriControl trials before command confirmation and after uttering a value, performing a multi-touch gesture. His gaze is being recognized at an aircraft almost on final (white radar label box).

The TriControl multimodal interaction results were compared against Baseline data gathered during a study with a conventional input system in November–December 2015 and January 2017. During these simulations, controllers had to give commands via voice to simulation pilots and had to enter those commands manually via mouse into the CWP. Both of these necessary controller tasks could be performed sequentially or in parallel. Twelve radar approach controllers took part in this earlier study. Four controllers were sent from DFS, eight controllers from COOPANS (consortium of air navigation service providers of Austria (Austro Control, Vienna, Austria, four controllers), Croatia (Croatia Control, Velika Gorica, Croatia, one controller), Denmark (Naviair, Copenhagen, Denmark, one controller), Ireland (Irish Aviation Authority, Dublin, Ireland, one controller), and Sweden (LFV, Norrköping, Sweden, one controller)).

The average age was 39 years (standard deviation = 11 years; within a range from 22 to 56), their professional work experience 17 years (standard deviation = 11 years; within a range from 1 to 34). The number of aircraft and weight category mix in the traffic scenario were the same as for the TriControl study. However, training and simulation runs lasted some minutes longer in this study (at least 45 min each). This aspect should not affect the duration of single controller commands during the run time except for some training effects.

5. Results Regarding Controller Command Input Duration and Efficiency

One of the primary goals of TriControl is to enable faster controller commands, as compared to Baseline. We defined the TriControl command input duration as the amount of time between the controller's initiation of the multi-touch gesture (or the pressing of the voice transmission pedal, whichever came first) and the confirmation of the command on the multi-touch display. In Figure 4, this is the difference between G1 and C2. For Baseline, we defined the command duration as the amount of time between the controller's pressing of the voice transmission pedal (or entering a value in the data label, whichever came first) and the confirmation of the command in the data label (or the release of the voice pedal, whichever came later).

The time for visually acquiring the radar label information and consideration (roughly E3–E1 in Figure 4 for TriControl) should be similar to the "time to think" in Baseline. Therefore, the "time to think" is not part of the command input duration time in both conditions. Response times are generally not normally distributed, due to the occasional outlier response that is significantly longer than average. Therefore, for both measures (command input duration of TriControl and Baseline), we computed the median duration for each controller, which is a much more stable estimate of individual controllers' response times. We compute the mean of those median response times and conduct *t*-tests comparing the means of Baseline and TriControl, reporting the *t*-statistic and *p*-values of individual comparisons. Although raw response times are not normally distributed, the controller medians are consistent with the normality assumptions of the *t*-test. When we presented data for individual command types (e.g., altitude, speed), we only included data from controllers with at least five of those commands.

5.1. Command Input Duration

Figure 6 shows the command duration results for the full set of commands and for the most common command types. Overall, TriControl commands (mean = 5.4 s) were slower than Baseline (mean = 4.7 s), though this was only marginally statistically significant: $t(24) = 1.77$; $p = 0.09$.

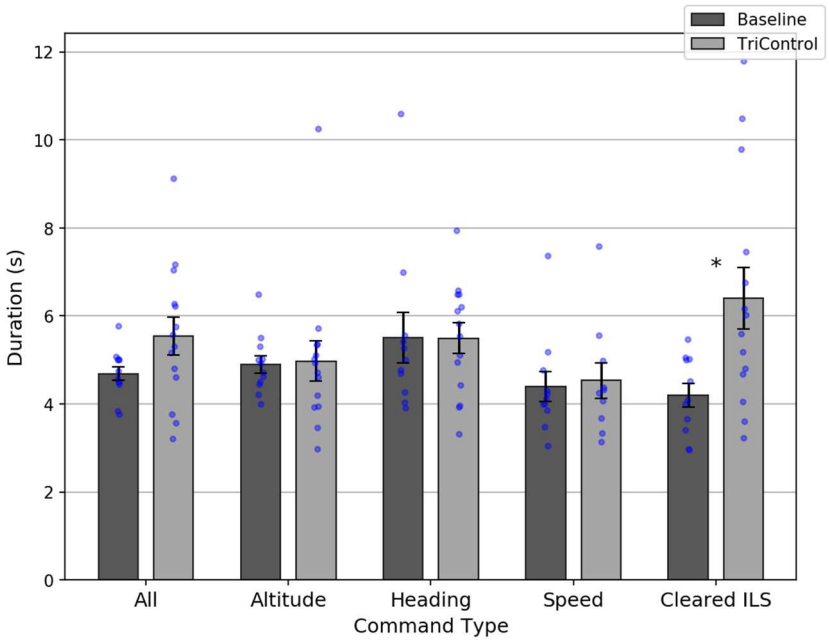


Figure 6. Baseline vs. TriControl command duration. Error bars represent ±1 standard error of the mean. Blue dots represent individual controllers.

Although TriControl was slower overall, this likely does not reflect the system's true performance potential. With only 15 min training, participants were still learning the system during the main scenario. Comparing the first 15 min to the second, there was an 11% drop in the number of gesture or voice corrections per command (0.62 to 0.55). There was also a 0.4 s decrease in the command duration for commands that did not require a correction (4.8 to 4.4 s).

In addition to training, there are also known limitations with TriControl that we plan to address in the future. At present, the speech recognition module is tuned to English language with a German accent. The German-speaking controllers made 50% fewer voice corrections and their command durations were 1.7 s shorter. Also, TriControl does not presently support chained commands (e.g., an altitude and speed command in a single transmission). In the Baseline condition, unchained command durations were 0.3 s longer than chained commands, on a per-command basis (mean = 4.7 s).

A best-case comparison would reflect the faster, low-error performance that we would expect from additional training and use, speech recognition that is trained on more accents, and an implementation of chained commands. With the current data set, we compared the Baseline unchained commands to the TriControl commands issued by native German speakers in the second half of the simulation that did not require corrections.

This comparison was reasonable, as controllers in the Baseline were very skilled and trained due to the similarity of the test system with current operational systems. Therefore, all following analyses encompass only the second half of the exercise. Results of the best-case comparison are shown in Figure 7.

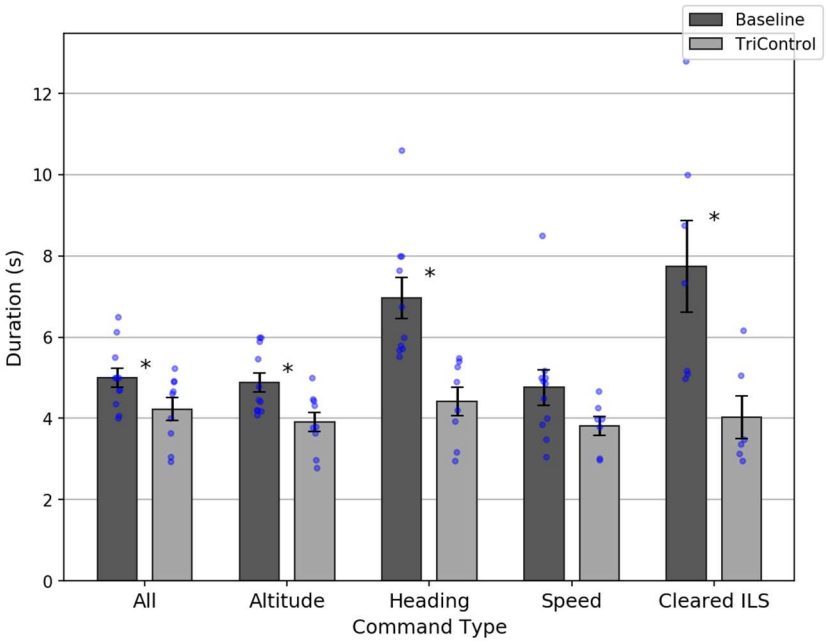


Figure 7. Baseline unchained commands vs. TriControl second half uncorrected German-speaker command duration. Error bars represent ±1 standard error of the mean. Blue dots represent individual controllers.

Overall, TriControl command durations (mean = 4.2 s) were shorter than Baseline (mean = 5.0 s), which is statistically significant: $t(18) = 2.11$; $p = 0.049$. This is a drop that exceeds 15%. There was also a significant speed improvement for altitude (3.9 vs. 4.9 s, $t(18) = 2.82$; $p = 0.01$), heading (4.4 vs. 7.0 s, $t(16) = 3.9$; $p = 0.001$), and ILS (4.0 vs. 7.7 s, $t(11) = 2.82$; $p = 0.02$) commands.

In addition to shorter median command durations in this best-case comparison, TriControl also has a smaller number of long-duration commands, excluding those that required input corrections. Figure 8 shows the cumulative distributions of command durations for Baseline and the uncorrected commands issued during the second 15 min of the TriControl simulation. Although the shortest 50% of commands have similar durations, the 75th percentile TriControl command is approximately one second shorter than Baseline, and the 90th percentile is two seconds shorter.

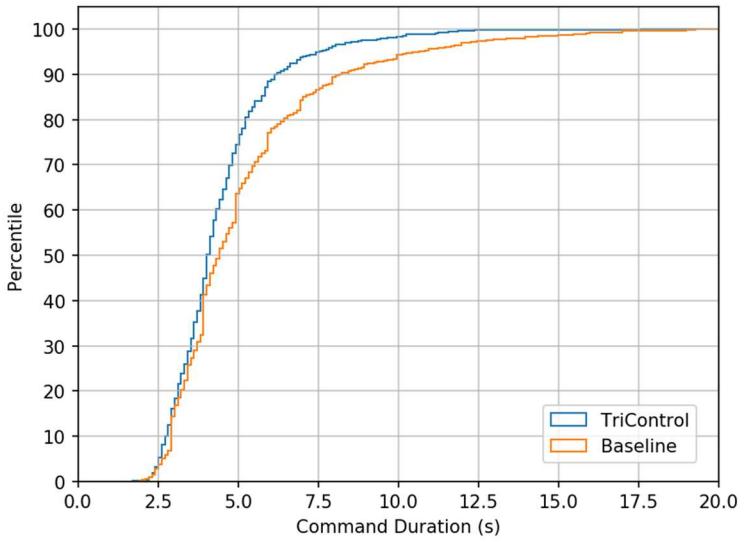


Figure 8. Cumulative distribution of Baseline vs. TriControl second half uncorrected command durations.

The 75th percentile comparison is also shown in Figure 9. Command durations were shorter for TriControl overall ($t(24) = 2.69$; $p = 0.01$), and for altitude ($t(23) = 5.51$; $p < 0.001$), heading ($t(21) = 2.41$; $p = 0.03$), and speed ($t(24) = 2.56$; $p = 0.02$) commands.

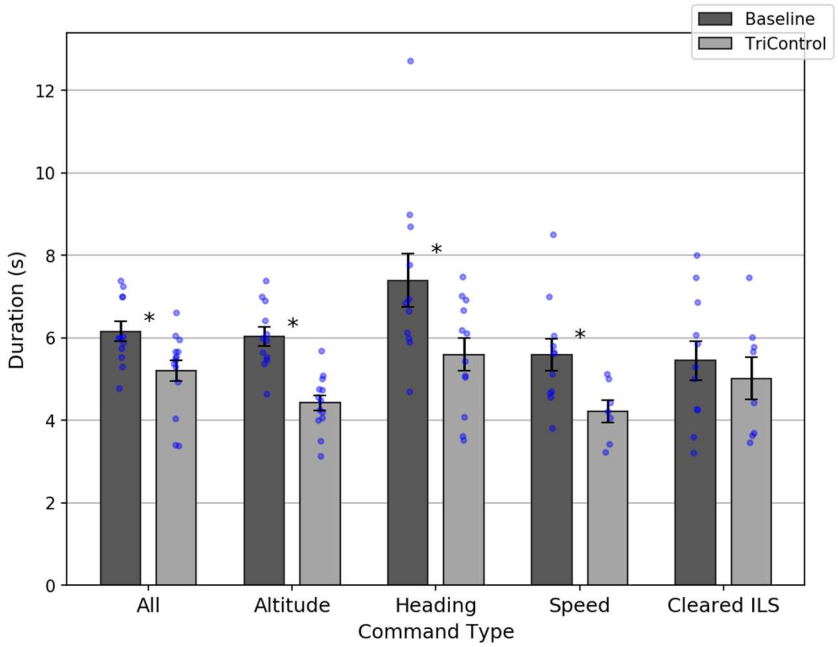


Figure 9. Baseline vs. TriControl second half uncorrected 75th percentile command duration. Error bars represent ±1 standard error of the mean. Blue dots represent individual controllers.

5.2. Command Input Efficiency

Finally, we investigated how efficiently controllers used the voice and manual modalities with each system. In particular, we looked at the portions of the command that can be parallelized. In the Baseline condition, this is the voice transmission of the command and the manual data entry into the radar label. In TriControl, this is the command type gesture and the command value vocalization. We defined the modality consecutiveness of the command as the ratio of (1) the total time to complete the voice and manual steps to (2) the larger of the two steps' durations. For example, consider a command

with a manual duration of one second and a voicing duration of two seconds. If the controller starts and completes the manual component while they are performing the voicing component—completely in parallel—then that command has a modality consecutiveness of 1.0. If the steps are done sequentially with no delay, that is a modality consecutiveness of 1.5. A delay of half a second would correspond to a consecutiveness of 2.0.

Figure 10 shows a scatterplot of each controller's median command duration and mean command modality consecutiveness of Baseline and for the uncorrected commands in the second half of TriControl. The data points for Baseline were fairly clustered, with modality consecutiveness ranging from 1.09 to 1.47, and most durations falling between 3.75 to 5 s. In contrast, the TriControl modality consecutiveness ranged from 1.0 to 2.04, and duration ranged from 2.9 to 5.2 s. A third of TriControl controllers were issuing their commands with more parallelism than the most parallel controller in the Baseline condition. This suggests that TriControl does support more parallelism than Baseline, and that as controllers become more comfortable performing the command components in parallel, they will produce commands more efficiently.

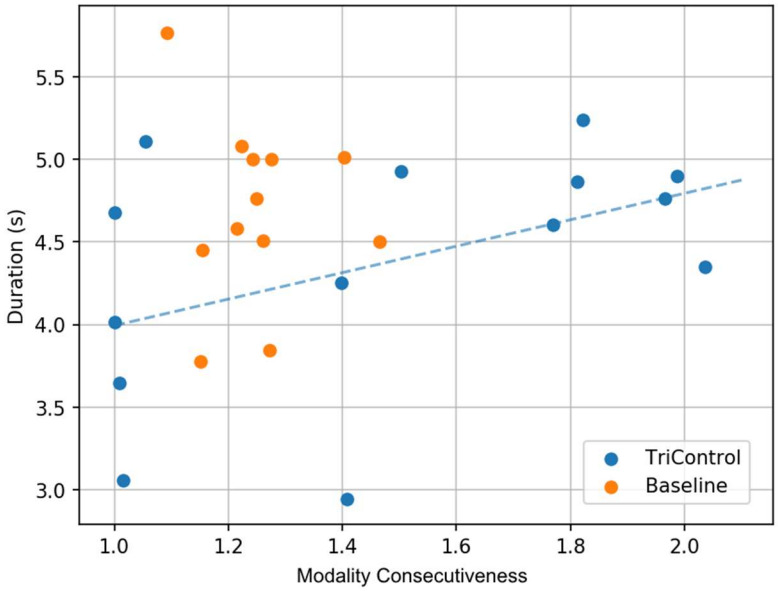


Figure 10. Command duration by modality consecutiveness for Baseline and TriControl second half uncorrected.

6. Discussion of Study Results and the TriControl Prototype

This section discusses the presented results with respect to performance, outlines some qualitative feedback and then considers safety aspects of the TriControl system.

6.1. Performance Considerations Derived from Results

On the one hand, for the majority of command types in Figure 6, the best performance was achieved from TriControl controllers. On the other hand, this is also true for the worst performance. However, this indicated a lack of training. More training could help the worst TriControl performers to also achieve results comparable to the best TriControl performers and thus better than the Baseline average. This assumption is already supported by the analyzed data subset of Figure 7.

Some TriControl enhancements, especially for the value recognition of different accents and more training, might lead to faster command inputs of up to two seconds per command, depending on the command type. The misrecognition of the uttered word "tower" of some native respectively non-German accent speakers to initiate an aircraft handover to the next controller position is one main aspect that needs to be improved. There was of course, only a limited set of modeled accents for this

first study. Enhancements are also necessary regarding a number of misrecognized confirmation touch taps on the multi-touch device.

The fastest commands of TriControl and Baseline were comparable (Figure 8). However, TriControl enabled faster medium command duration times that might again even improve with more training. This can result in roughly one second saved per command input over all analyzed command types (Figure 9).

The assumed improvement due to more than 15 min of training (and 30 min simulation run) with the multimodal system can also be seen in Figure 10. When analyzing the best participants working with TriControl (near *x*- and *y*-axis in Figure 10) they were more efficient and faster than the best Baseline participants. As the Baseline environment is quite equal to today's controller working positions, it can be assumed that the potential for improvement of performance (orange dots in Figure 10) is low. This is also supported by the clustering effect of those orange dots for Baseline participants. As there might not be anticipated a performance decrease of the best TriControl participants (lower left blue dots), the bad performers will probably improve with training and reach or overcome the Baseline average (blue dots in the upper right area of the plot might move more to the lower left). The familiarization to the new parallel multimodal input was reported as a main aspect of training. Many subjects already improved in the short amount of time during the study. This is also interesting from a pedagogical perspective: an increase in performance over the Baseline with only a short 15 min period of training raises the prospect of what results could be expected with the inclusion of a more comprehensive training and preparation program for the test subjects.

6.2. Qualitative Information about Test Subjects' Opinion

The feedback of test subjects gathered in a debriefing session after the simulation run concerned various positive aspects and suggestions for improvements. A lot of statements were related to the learnability such as:

- "I found it, surprisingly, simple after a period of practice.",
- " . . . easy to spot—and after a short time as well easy to correct.",
- "I had no problems at all.",
- "It's quite interesting on a conceptual level and easy to use.",
- "The system is easy to learn.".

However, there were some flaws that need to be taken into account in the further development according to:

- "Conditional Clearances and Combined Calls are needed for daily work; at this time the clearance given is too simple to reflect the demands.",
- "The system often reacts too slow; e.g., with respect to the eye tracking dwell time; too long 'staring' necessary to 'activate' the square.",
- "Instead of 'watching the traffic' I needed to 'watch if the system complies'.",
- "TriControl focuses on just one aircraft. That might be a reason for attentional/change blindness.",
- "Uttering the whole command might be better for the situational awareness.",
- "There is potential for confusion and errors.",
- " . . . for use in ATM systems many more variables need to be incorporated and tested for safety".

Nevertheless there were many encouraging comments to further follow the multimodal approach:

- "A useful tool.",
- "This leads to less misunderstandings.",
- "I think it's worth to think about systems like TriControl, but it is really hard to state now if I would really work live-traffic with it . . . ; the use is fun!",
- "I would prefer TriControl over mouse or screen inputs.",

- "As an On-the-job Training Instructor (OJTI) teaching controllers to be instructors, this would be a good system to easily see what the controller is thinking/doing. I liked the system. Naturally, it would require practice and exercise to improve the skills required. However, once done, I believe it would be a good aid to the controller.",
- "After adequate training I expect significantly more performance.".

The level of comments from the participants broadly corresponded to the age of the participants and their success with the system—the younger participants often had greater success (best case scenarios) with making command submissions/inputs simultaneously, and as a result, enjoyed a higher success rate with the evaluation. This lead to more positive anecdotal comments regarding the potential of the system compared to many of the older test subjects. Older subjects found the multi-modality more challenging to coordinate and commented that they were frustrated by the system. The success rate in ATM training generally has been recognized to be better with younger trainees within a certain age range; it is assumed that within this range they can develop cognitive skills more readily.

To sum comments and observations up, the TriControl prototype in the current stage has—as expected—still far to go until it can be considered for operationalization. However, the underlying concept seems to have great potential for benefits, and for being used in future ATC environments with controllers who are "native" with modern HMI technologies.

6.3. Safety Considerations

In terms of safety, the general ATC tasks and procedures of the controllers did not change using TriControl compared to the traditional operation. Although the current state of the TriControl prototype has a limited amount of available commands, the radiotelephony channel is always intended to serve as a safe fallback solution for exceptional cases. Also today there is the potential for a misunderstanding between controller and pilot in the voice communication. This safety factor is to some extent comparable to an accidental insertion of erroneous commands into the system. To prevent this issue, the confirmation step for the command inserted via the three modalities was introduced. However, during the trials, a number of commands mistakenly input into the system were recognized. As stated earlier, this is most probably a result of the voice recognition not modeled for the respective accent, and misrecognized confirmations. Those issues are expected to decrease due to two reasons. First, increased training of the controllers with the system will lead to less erroneous inputs. Second, further development of the prototype foresees an elaborated plausibility check with respect to all three parts of a command and the command as a whole with respect to the current air traffic situation.

Hence, this plausibility check even goes beyond the two-way command-read-back-check of controllers and pilots today. With TriControl it will not be possible to assign commands to aircraft callsigns that do not exist in the airspace as sometimes happens nowadays. Furthermore, the TriControl system might immediately execute a what-if-probing and warn the controller before confirmation and thus issuing of the command, in case of potential negative implications to the air traffic situation. Besides, the ambiguity of human verbal communication can be reduced as TriControl uses a clearly defined set of commands to be issued.

Nevertheless, it is obvious that low error rates regarding eye tracking, speech, and gesture recognition should exist. Those rates are achievable e.g., using Assistance Based Speech Recognition (ABSR) as already partly used by TriControl. When analyzing complete commands with all three command parts uttered as usual, Command Error Rates of 1.7% can be achieved [23]. At the same time, wrong and forgotten inputs into the aircraft radar labels can be reduced with the help of an electronic support system compared to just using the mouse for all manual label inputs [23]. An intelligent aircraft radar label deconflicting is important to avoid selecting wrong aircraft via eye tracking if some aircraft are near each other. Malfunction of gesture recognition to generate unintended command types have hardly being pointed out by controllers during the study. As low error rates with respect to

the "automatic recognition" of generated commands can be expected in the future TriControl version and the additional confirmation step exists, no basic ATC safety showstopper is currently expected.

The aspect of fatigue can also be related to safety. It needs to be analyzed if controllers might experience fatigue earlier using the three modalities as implemented in the current stage of the TriControl prototype compared to just using voice communication and mouse inputs. However, if a controller is able to—at a later stage of TriControl—choose the modality that is the most convenient for the current situation, this should not lead to earlier fatigue than in traditional CWP HMIs.

As this study was intended to get early insights on the potential speed and efficiency benefit of the novel interaction concept, future versions of the prototype will incorporate measures to maintain or increase the safety introduced by TriControl compared to traditional CWPs.

7. Summary and Outlook

TriControl is DLR's multimodal CWP prototype to enable faster and more natural interaction for air traffic controllers. The comparison of TriControl (controllers used this setup for the first time ever) and Baseline (controllers were very familiar with this kind of setup) study results primarily reveals the potential to improve speed and efficiency of air traffic controller command input under certain conditions. The short training duration during the TriControl study and thus the familiarization with the multimodal system most probably even worsened the TriControl in contrast to Baseline results.

Hence, in the pure analysis of the data as a whole, TriControl does not seem to enable faster and more efficient interaction. However, a closer look into specific parts of the data unveils relevant benefits. When investigating the second 15 min of each TriControl simulation result, the data of unchained and uncorrected commands and of participants with a German accent, for which TriControl was designed in its current version, a speed gain of 15% can be seen. An improvement in speed and efficiency might lead to less controller workload, or could in the long run help to increase the number of aircraft handled per controller.

The multimodal TriControl system is of interest to research and educational specialists, for example within the DFS, because it combines several developing technologies, which are either being researched, in operational use in ATC systems, or already used in training. Touch input devices are a mature and widely used ATC interface. The ability to 'email' clearances to aircraft via Controller Pilot Data Link (CPDLC) rather than by voice is used extensively in ATC communication systems. Research is being undertaken in eye tracking and its' use in controller support tools. Voice recognition and response (VRR) systems are used in various aspects of training, and the technology is improving.

Trials of VRR in ATC simulation within the DFS Academy have suggested similar results to those in the TriControl trial: general VRR recognition results are often overshadowed by higher misrecognition rates of the verbal instructions of a small number of users, sometimes because of an unusual accent for which the VRR has not been tuned. This limits its use as a training tool, and can also lead to low expectations and poor acceptance levels by users. Improving the system's recognition ability can remove this potential roadblock to its implementation.

Many aspects like controllers' experience with the involved interaction technologies, the parallelism of input, their own controller working position design and responsibility area as well as age might have influenced the TriControl results. Nevertheless, some controllers performed really well with the multimodal system that has only been planned for future CWPs. Hence, there will probably be a new generation of trainable controllers in the future to benefit from the potential speed gain in command input. There is a list of reasonable enhancements for this early version of a multimodal CWP prototype that we gained from the evaluations.

In a next step, TriControl will for example, use a context-based fusion of input modality data for plausibility checking of information, as DLR has already shown that it is reliable in speech recognition [35]. This step aims to reduce recognition errors by using the information from all three modalities to cooperatively construct a reasonable controller command. This could also increase

the safety as the pre-check of complementary command parts will be performed before sending the information to a pilot via data-link or standardized text-to-speech.

Other differences to current interaction are the lack of chain commands, conditional clearances, or traffic information. Therefore, we hope to find reasonable ways to include possibilities to cover a great amount of the controllers' interaction spectrum by incorporating their feedback as early as possible. Accordingly, the effects on safety using the TriControl system have to be analyzed as well, when the maturity of the prototype has increased. The limited time available to each participant did not allow familiarity with the system to be fully developed before the trial took place. With a longer period of time in a second trial, the individual tasks could be evaluated, as well as combined tasks, prior to the exercise. This data evaluation would be of interest, but would involve more intensive evaluation that was not possible in the first trial. Training could then be tailored to assist and overcome the main problems identified in a qualitative error analysis. A quantitative error analysis was not reasonable for the first trial due to the described different sources of errors and as it was not the scope of the first trial.

To furthermore investigate the effects of users' free choice of modalities, we will extend each of the involved modalities to enable inputs for as many command parts as possible. On the one hand, users will be able to choose the modality for a piece of information as it fits their needs or the situation. On the other hand, quantitative evaluations will be performed to assess the performance of different interaction modality combinations. This will also enable an assessment of the performance using the modality combination presented in this work compared to combinations more similar to the current work of the controller.

The multimodal interaction CWP prototype TriControl showed the possibility of fast and efficient input of most approach controller command types into the digital system already now. This will be essential in the context of digital ATC information distribution and future ATC tasks.

Very similar benefits are as well anticipated by the major European ANSPs and ATM system providers in the course of SESAR2020's solution "PJ.16-04 CWP HMI" investigating the effects of automatic speech recognition, multi-touch inputs and eye-tracking at the CWP on controller productivity. Thus, future activities will probably also use and combine the advantages of those innovative HMI technologies.

Author Contributions: O.O. and M.J. were responsible for development of the TriControl system. They also conceived and designed the experiments with great organizational support of M.W., O.O. was the main author of this article being supported by all three co-authors. J.R.R. was responsible for the preparation and conduction of the data analysis (mainly represented in results section).

Acknowledgments: We like to thank all air traffic controllers that participated in our human-in-the-loop studies during the last two years. Besides, we are grateful for the support of Konrad Hagemann (DFS Planning and Innovation) and Hartmut Helmke (DLR) in preparing the simulation trials in Langen (DFS) respectively Braunschweig (DLR). Thanks also to Axel Schmugler (Technical University Dresden) for assisting during the TriControl trials and conducting a feasibility analysis questionnaire (results to be published soon).

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Oviatt, S. Multimodal Interfaces. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*; L. Erlbaum Associates Inc.: Hillsdale, NJ, USA, 2003; pp. 286–304.
2. Sharma, R.; Pavlovic, V.I.; Huang, T.S. Toward multimodal human-computer interface. *Proc. IEEE* **1998**, *86*, 853–869. [CrossRef]
3. Quek, F.; McNeill, D.; Bryll, R.; Kirbas, C.; Arslan, H.; McCullough, K.E.; Furuyama, N.; Ansari, R. Gesture, speech, and gaze cues for discourse segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2000 (Cat. No. PR00662), Hilton Head Island, SC, USA, 15 June 2000; Volume 2, pp. 247–254.

4. Caschera, M.; D'Ulizia, A.; Ferri, F.; Grifoni, P. Towards Evolutionary Multimodal Interaction. In *On the Move to Meaningful Internet Systems: OTM 2012 Workshops: Confederated International Workshops: OTM Academy, Industry Case Studies Program*; Herrero, P., Panetto, H., Meersman, R., Dillon, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 608–616.
5. ETSI. *Human Factors (HF); Multimodal Interaction, Communication and Navigation Guidelines*; ETSI EG 202 191 V1.1.1; ETSI: Valbonne, France, 2003; p. 7.
6. Oviatt, S. Ten myths of multimodal interaction. *Commun. ACM* **1999**, *42*, 74–81. [CrossRef]
7. Martin, J.C. Towards intelligent cooperation between modalities. The example of a system enabling multimodal interaction with a map. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'97) Workshop on Intelligent Multimodal Systems, Nagoya, Japan, 24 August 1997.
8. Oviatt, S. Multimodal interfaces. In *Handbook of Human-Computer Interaction*; Jacko, J., Sears, A., Eds.; Lawrence Erlbaum: Mahwah, NJ, USA, 2002.
9. Tavanti, M. *Multimodal Interfaces: A Brief Literature Review*; EEC Note No. 01/07; EUROCONTROL: Brussels, Belgium, 2007.
10. Dumas, B.; Lalanne, D.; Oviatt, S. Multimodal Interfaces: A Survey of Principles, Models and Frameworks. In *Human Machine Interaction: Research Results of the MMI Program*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 3–26.
11. Koons, D.B.; Sparrell, C.J.; Thorisson, K.R. Integrating simultaneous input from speech, gaze, and hand gestures. In *Intelligent Multimedia Interfaces*; Maybury, M.T., Ed.; American Association for Artificial Intelligence: Menlo Park, CA, USA, 1993; pp. 257–276.
12. SESAR. *The Roadmap for Delivering High Performing Aviation for Europe—European ATM Master Plan*; EU Publications: Brussels, Belgium, 2015.
13. Czyzewski, A. New applications of multimodal human-computer interfaces. In Proceedings of the 2012 Joint Conference New Trends in Audio & Video and Signal Processing: Algorithms, Architectures, Arrangements and Applications (NTAV/SPA), Lodz, Poland, 27–29 September 2012; pp. 19–24.
14. Neßelrath, R.; Moniri, M.M.; Feld, M. Combining Speech, Gaze, and Micro-gestures for the Multimodal Control of In-Car Functions. In Proceedings of the 12th International Conference on Intelligent Environments (IE), London, UK, 14–16 September 2016; pp. 190–193.
15. Pfleging, B.; Schneegass, S.; Schmidt, A. Multimodal interaction in the car: combining speech and gestures on the steering wheel. In Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'12), Portsmouth, NH, USA, 17–19 October 2012; ACM: New York, NY, USA, 2012; pp. 155–162.
16. Bolt, R.A. 'Put-that-there': Voice and gesture at the graphics interface. In Proceedings of the 7th annual Conference on Computer Graphics and Interactive Techniques, Seattle, WA, USA, 14–18 July 1980; pp. 262–270.
17. Merchant, S.; Schnell, T. Applying Eye Tracking as an Alternative Approach for Activation of Controls and Functions in Aircraft. In Proceedings of the 19th DASC, Philadelphia, PA, USA, 7–13 October 2000; pp. 5.A.5-1–5.A.5-9.
18. Savery, C.; Hurter, C.; Lesbordes, R.; Cordeil, M.; Graham, T. When Paper Meets Multi-touch: A Study of Multi-modal Interactions in Air Traffic Control. In Proceedings of the 14th International Conference on Human-Computer Interaction (INTERACT), Cape Town, South Africa, 2–6 September 2013; Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M., Eds.; Lecture Notes in Computer Science, LNCS-8119 (Part III), Human-Computer Interaction. Springer: Berlin/Heidelberg, Germany, 2013; pp. 196–213.
19. Traoré, M.; Hurter, C. Exploratory study with eye tracking devices to build interactive systems for air traffic controllers. In Proceedings of the International Conference on Human-Computer Interaction in Aerospace (HCI-Aero'16), Paris, France, 14–16 September 2016; ACM: New York, NY, USA, 2016. Article 6.
20. Alonso, R.; Causse, M.; Vachon, F.; Robert, P.; Frédéric, D.; Terrier, P. Evaluation of Head-Free Eye Tracking as an Input Device for Air Traffic Control. In *Ergonomics*; Taylor & Francis Group: Abingdon, UK, 2012; Volume 56, pp. 246–255.
21. Seelmann, P.-E. Evaluation of an Eye Tracking and Multi-Touch Based Operational Concept for a Future Multimodal Approach Controller Working Position (Original German Title: Evaluierung eines Eyetracking und Multi-Touch basierten Bedienkonzeptes für einen zukünftigen multimodalen Anfluglotsenarbeitsplatz). Bachelor's Thesis, DLR-Interner Bericht, Braunschweig, Germany, 2015.

22. Jauer, M.-L. Multimodal Controller Working Position, Integration of Automatic Speech Recognition and Multi-Touch Technology (Original German Title: Multimodaler Fluglotsenarbeitsplatz, Integration von automatischer Spracherkennung und Multi-Touch-Technologie). Bachelor's Thesis, Duale Hochschule Baden-Württemberg Mannheim in Cooperation with DLR, Braunschweig, Germany, 2014.
23. Helmke, H.; Ohneiser, O.; Mühlhausen, T.; Wies, M. Reducing Controller Workload with Automatic Speech Recognition. In Proceedings of the 35th DASC, Sacramento, CA, USA, 25–29 September 2016.
24. Uebbing-Rumke, M.; Gürlük, H.; Jauer, M.-L.; Hagemann, K.; Udovic, A. Usability evaluation of multi-touch displays for TMA controller working positions. In Proceedings of the 4th SESAR Innovation Days, Madrid, Spain, 25–27 November 2014.
25. Möhlenbrink, C.; Papenfuß, A. *Eye-Data Metrics to Characterize Tower Controllers' Visual Attention in a Multiple Remote Tower Exercise*; ICRAT: Istanbul, Turkey, 2014.
26. Oviatt, S.; DeAngeli, A.; Kuhn, K. Integration and synchronization of input modes during multimodal human-computer interaction. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'97), Atlanta, GA, USA, 22–27 March 1997; ACM: New York, NY, USA, 1997; pp. 415–422.
27. Oviatt, S. Multimodal interactive maps: Designing for human performance. *Hum. Comput. Interact.* **1997**, *12*, 93–129.
28. Cohen, P.; McGee, D.; Clow, J. The efficiency of multimodal interaction for a map-based task. In Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLC'00), Association for Computational Linguistics, Seattle, WA, USA, 29 April–4 May 2000; pp. 331–338.
29. Ohneiser, O.; Jauer, M.-L.; Gürlük, H.; Uebbing-Rumke, M. TriControl—A Multimodal Air Traffic Controller Working Position. In Proceedings of the Sixth SESAR Innovation Days, Delft, The Netherlands, 8–10 November 2016.
30. McMillan, D. Miscommunications in Air Traffic Control. Master's Thesis, Queensland University of Technology, Brisbane, Australia, 1999.
31. Cardosi, K.M.; Brett, B.; Han, S. *An Analysis of TRACON (Terminal Radar Approach Control) Controller-Pilot Voice Communications*; (DOT/FAA/AR-96/66); DOT FAA: Washington, DC, USA, 1996.
32. ICAO. *ATM (Air Traffic Management): Procedures for Air Navigation Services*; DOC 4444 ATM/501; International Civil Aviation Organization (ICAO): Montréal, QC, Canada, 2007.
33. Ohneiser, O. *RadarVision—Manual for Controllers (Original German Title: RadarVision—Benutzerhandbuch für Lotsen)*; Internal Report 112-2010/54; German Aerospace Center, Institute of Flight Guidance: Braunschweig, Germany, 2010.
34. DLR Institute of Flight Guidance, TriControl—Multimodal ATC Interaction. 2016. Available online: http://www.dlr.de/fl/Portaldata/14/Resources/dokumente/veroeffentlichungen/TriControl_web.pdf (accessed on 6 April 2018).
35. Helmke, H.; Rataj, J.; Mühlhausen, T.; Ohneiser, O.; Ehr, H.; Kleinert, M.; Oualil, Y.; Schulder, M. Assistant-Based Speech Recognition for ATM Applications. In Proceedings of the Eleventh USA/Europe Air Traffic Management Research and Development Seminar (ATM2015), Lisbon, Portugal, 23–26 June 2015.

aerospace

MDPI

Article

Operational Feasibility Analysis of the Multimodal Controller Working Position "TriControl"

Oliver Ohneiser [1,*], Marcus Biella [1], Axel Schmugler [2] and Matt Wallace [3]

[1] German Aerospace Center (DLR), Institute of Flight Guidance, Lilienthalplatz 7, 38108 Braunschweig, Germany; Marcus.Biella@DLR.de
[2] Technische Universität Dresden, School of Science, Faculty of Psychology, 01062 Dresden, Germany; Axel.Schmugler@Mailbox.TU-Dresden.de
[3] Deutsche Flugsicherung GmbH, Academy, Am DFS-Campus, 63225 Langen, Germany; Matthew.Wallace@DFS.de
* Correspondence: Oliver.Ohneiser@DLR.de

Received: 14 January 2020; Accepted: 9 February 2020; Published: 20 February 2020

Abstract: Current Air Traffic Controller working positions (CWPs) are reaching their capacity owing to increasing levels of air traffic. The multimodal CWP prototype TriControl combines automatic speech recognition, multitouch gestures, and eye-tracking, aiming for more natural and improved human interaction with air traffic control systems. However, the prototype has not yet undergone systematic evaluation with respect to feasibility. This paper evaluates the operational feasibility, focusing on the system usability of the approach CWP TriControl and its fulfillment of operational requirements. Fourteen controllers took part in a simulation study to evaluate the TriControl concept. The active approach controllers among the group of participants served as the main core target subgroup. The ratings of all controllers in the TriControl assessment were, on average, generally in slight agreement, with just a few showing statistical significance. However, the active approach controllers performed better and rated the system much more positively. The active approach controllers were strongly positive regarding the system usability and acceptance of this early-stage prototype. Particularly, ease of use, user-friendliness, and learnability were perceived very positively. Overall, they were also satisfied with the command input procedure, and would use it for their daily work. Thus, the participating controllers encourage further enhancements to be made to TriControl.

Keywords: air traffic controller; human–machine interaction; usability; multimodality; eye-tracking; automatic speech recognition; multitouch gestures; controller command; feasibility analysis

1. Introduction

Facing the growing levels of air traffic and increased automation, conventional working methods and workstations are no longer adequate for air traffic controllers (ATCOs). Hence, there is a need for a faster, more efficient, and, ideally, more natural way of working, considering that a huge amount of ATCO work includes communication with aircraft pilots to issue commands. The communication between ATCOs and pilots is still mostly based on radio telephony. Hence, current human–machine interfaces (HMIs) in air traffic control (ATC) support the single voice interaction mode. However, this principle contradicts natural, intuitive, and efficient human communication. For instance, when explaining the route and distance to the airport to someone, verbal instruction is most often simultaneously accompanied by gestures for the direction and eye contact. In order to comply with these everyday interactions, the modalities should also be available at the controller working position (CWP), which mainly comprises an interactive air traffic situation data display and the communication infrastructure.

First, controllers observe and analyze the air traffic at their situation data display. Thus, it is only a small step to utilizing eye-tracking technology to select the currently observed aircraft as the one to receive the next command. Secondly, verbal commands are an everyday concept in ATC. However, utterances can be reduced to only articulating command values. Thirdly, performing simple and fast multitouch gestures for command types—as are widely used nowadays on electronic consumer products—complements an easy and natural way of creating commands, which the controller finally confirms.

Multimodal HMIs combine different interaction modalities, aiming to support a natural [1] and efficient way of human communication [2,3]. Recent research has revealed that reasonable interaction technologies [4] for a CWP should recognize touch, speech, and gaze [5–8]. In accordance with these findings, the German Aerospace Center (DLR) has developed the multimodal CWP TriControl concept, which combines automatic speech recognition, multitouch gestures with one or multiple fingers on a touch input device, and eye-tracking via infrared sensors located at the bottom of the monitor. These modalities can be used to input the three basic ATC command parts, i.e., aircraft callsign, command type, and value, into the ATC system [9]. Hence, conventional subsequent command parts that are uttered verbally are replaced by parallel ATC system input with different modalities [10].

First analyses with the multimodal CWP prototype TriControl showed an acceleration of command input by up to 15% [10]. Furthermore, the artificial voice broadcast or data-link transmission of commands resulting from combined command parts of the parallel ATC system input in TriControl can also reduce misunderstandings in verbal communication caused by various "foreign language" English accents [11], that might even lead to serious accidents [12].

However, the operational feasibility including the system usability and acceptability of TriControl have not yet been systematically evaluated with ATCOs in a realistic environment [9]. As ATCOs work in a highly safety-critical domain, they and the air navigation service providers are very cautious with respect to new technologies [13].

The goal of this paper is to evaluate the multimodal CWP prototype TriControl in practice, i.e., mainly based on questionnaires after simulation runs, to receive input from target users for future development [14]. The evaluation concentrates on operational feasibility in terms of system usability and analyzing the fulfillment of operational requirements.

In the next section, we present the relevant background on multimodal HMIs, the validation methodology, and the TriControl system. Section 3 is the method part, introducing the participants, setup, and contents of the TriControl feasibility analysis study. All analysis results on system usability, acceptability, and performance are presented in Section 4. The main results and further comments are briefly presented in Section 5. Finally, Section 6 summarizes and concludes this paper, and sketches out future work.

2. Background of Multimodal Interfaces, Feasibility Analysis, and the CWP prototype TriControl

Many studies have investigated the advantages and disadvantages of specific interaction technologies as well as of multimodal HMIs. The most important results regarding multimodal HMIs and their relationship with the ATC domain are outlined in the following section. Furthermore, a theoretical background regarding the main aspects of the feasibility analysis study is outlined, i.e., concerning the validation methodology, the concepts of usability and acceptability, as well as the user-centered design. Finally, the functionality of the multimodal CWP prototype TriControl that was evaluated in a feasibility analysis is explained.

2.1. Multimodal HMIs and Their Benefits for ATC

Human–machine systems comprise reciprocal interaction between system components such as hardware and software as well as humans to achieve specific goals [15]. The communication channel for information between human and machine is called the "modality" [16]. HMIs usually utilize visual-, audio-, and sensor-based modalities [17]. Hence, there can be unimodal or multimodal HMIs,

depending on the number of utilized modalities. The HMI serves as a means for information exchange, including input and output. Well-designed HMIs can lead to better operational performance, safety [18], and efficiency [19], in particular, for high risk systems such as air traffic control workstations [20].

ATCOs shall ensure safe and efficient air traffic flow, also avoiding fuel and noise pollution, [21] being supported by their CWP. The basic tasks of ATCOs follow a cognitive cycle that consists of checking external information, searching for conflicts, issuing commands, and updating their mental picture [22]. The radar screen of their CWP is a central means to receiving external information, to search for conflicts, and to updating the ATCOs' mental picture. Thus, the information flow from the machine to the ATCO is mostly visually based. For issuing of commands—and the check of pilots' readbacks—ATCOs use audio-based radio telephony in two-way communication with pilots [23] and other ATCOs. Furthermore, there are touch-based input devices at modern CWPs, which are used to update the information in the system. However, due to a lack of reasonable parallel usage of those modalities, current CWPs are more like a set of different unimodal HMIs instead of a multimodal HMI as sketched.

The limits of human cognitive resources for information processing compared to the machine especially are a disadvantage of unimodal HMIs [24]. If the speed of information input or output, natural interaction, error-proneness, or individualization is essential, unimodal HMIs are often not the best concepts [17]. Conversely, multimodal HMIs take into account that humans process information modality-dependent and potentially simultaneously via different modalities considering the cognitive load theory [25] and the working memory model [26]. Respective interfaces and their analysis started in the 1980s [27].

Nowadays, many versions of multimodal HMIs exist, but all of them contain the principle of "more than one modality" for a human to interact with a machine [16,28,29]. Focusing more on the technical aspects and with regard to TriControl, a definition of multimodal interfaces could be "two or more combined user input modes—such as speech, pen, touch, manual gestures, gaze, and head and body movements—in a coordinated manner with multimedia system output." [30]. Multimodal HMIs have several advantages compared to unimodal HMIs. They promise to be more intuitive [31], to be better fitting to human needs for system control [32], to be faster, safer, and more reliable [33], to be less error-prone [34,35], to be more robust [35], to offer the best-suiting modality for users' choice [36], to reduce cognitive workload of users [37], and to comprise briefer and less disfluent input [34,38].

In a tested use case, roughly 95% of target users would prefer the multimodal over unimodal interaction [34]. However, users often mix unimodal and multimodal interaction [39,40]. Particularly in case of low cognitive load, users may even prefer unimodal interaction [41]. If users established their individual way of using multiple modalities, this interaction style would hardly change [42]. Overloading one or multiple modalities with information can result in less trust in the HMI [43].

Further findings on the three interaction modalities (automatic speech recognition, eye-tracking, and multitouch inputs as used by TriControl) are outlined in the following. The use of speech recognition in ATC started decades ago [44]. As ATCOs and pilots are obliged to use standard phraseology, a set of rules and terms for verbal radio telephony communication defined by the International Civil Aviation Organization (ICAO), the number of words and word sequences is limited compared to natural language [45]. Transcribing controller utterances word-by-word is only a small step for further applications [46]. The interpretation of those words also needs to be annotated to understand the respective command parts [47] especially if they do not follow the ICAO phraseology.

Given the commands, it is possible to highlight aircraft labels, to assess ATCO workload [48], or to enable safety functions [49,50] in operational environments or for the support of ATC training and simulations [51]. However, low command error rates are crucial for those applications. Thus, assistant-based speech recognition has been introduced [52]. A command hypotheses generator provides the most probable and reasonable commands in a given situation to a speech recognition engine. This reduces the command error rate down to 1.7%. The respective radar label maintenance

task supported by speech recognition reduces ATCO workload for ATC system input by more than 30% [53].

Eye-tracking interaction has already been analyzed in the air traffic domain [54,55]. The freeing of hands to be used for other manual interaction is one central advantage of this interaction means [56]. However, the visual selection of elements after a gaze dwell time does not seem beneficial [57]. The use of gestures in the ATC context has also been investigated. Earlier prototypes mostly include multitouch surfaces for more complex gestures [58]. Further examples from ATC research prototypes combine gestures with eye-tracking [59] or speech recognition [60]. Another application uses visual gesture recognition of air marshallers [61,62].

Touch gesture recognition has been investigated in the context of multimodal CWP [13,63]. Multitouch based interaction was evaluated as natural and fast enough for ATC applications [13]. Furthermore, users were able to work with the tested modalities quickly and perceived easy interaction [13]. Speed gains of up to 14% could be achieved compared to mouse inputs [64]. As ATCOs work in a highly safety-critical environment, the acceptability and trust by ATCOs of their HMIs is essential.

The user-centered design process takes target users into account in each design step of the HMI. Hence, there are early opportunities to influence the HMI development to the needs of ATCOs also in low technology readiness levels respective to validation phases.

2.2. Validation Methodology for Feasibility Analysis with Usability, Acceptability, and User-Centred Design

The European Operational Concept Validation Methodology 3 (E-OCVM 3) developed by the European Organisation for the Safety of Air Navigation (EUROCONTROL) [65] provides a processual approach for the validation of air traffic management (ATM) operational concepts. The methodology shall include all relevant stakeholders and support the development process. The E-OCVM concept lifecycle model encompasses eight steps for maturing concepts based on iterative loops for design and evaluation. The steps for "validation phases" (V) are "ATM needs (V0)", "Scope (V1)", "Feasibility (V2)", "Pre-Industrial Development and Integration (V3)", "Industrialization (V4)", "Deployment (V5)", "Operations (V6)", and "Decommissioning (V7)" focusing on V1 to V3 for the concept validation methodology. Many of those phases are similar to the more popular "technology readiness levels" (TRL) 1 to 9 [66]. Hence, V1 corresponds to TRL2 "Technology concept and/or application formulated", V2 corresponds to TRL4 "Component validation in laboratory environment", and V3 corresponds to TRL6 "System/subsystem model or prototype demonstration in an operational environment" [67].

The TriControl CWP prototype is assumed to fulfill step V2 "feasibility" of E-OCVM 3 or TRL4 "Component validation in laboratory environment" respectively. In this step, the technological concept of a prototype in the ATC domain should be elaborated to be operationally feasible in normal and non-normal conditions, the latter e.g., comprising emergency flights or severe weather. An initial functional prototype should undergo a simulation for further analysis and revelation of further development needs. The aspect of feasibility itself is again subdivided into operability respectively usability, (system) acceptability, and performance.

Usability as one aspect of feasibility is defined as a construct with many facets including the aspects being "easy to learn, efficient to use, easy to remember, having a low error rate, and meeting user satisfaction" [68]. It can also be seen as the extent to which a system can be used regarding specified users and context to reach effectiveness, efficiency, and satisfaction [69]. Much background of the concept "usability" is given in [70]. The focus on users and environments next to just the tasks in the tool development is a central factor resulting from usability concerns [71,72]. If usability is taken into account, this can increase productivity, reduce training and support needs, improve users' acceptance [73], or even lead to higher efficiency [74]. Usability can be measured directly or indirectly [75,76] via questionnaires and interviews on perceived usability respectively via behavioral and interaction data from system experiments [77,78]. Therefore, a combination of evaluation methods improves the usability assessment [79]. System usability problems can be detected with small user sample sizes. In specific studies, five study subjects were able to find 80% of usability problems and

15 study subjects detected all usability problems [80,81], however there may be hierarchies in those problems so that fixed numbers of subjects might not make sense [82].

Acceptability as another aspect of feasibility can be defined as the perceived usefulness and ease of use of a system to fulfill a task [83–85]. This affects the attitude towards the system as well as the behavioral intention and actual use of this system following the Technology Acceptance Model (TAM) [86]. The TAM has broadly been applied and developed high reliability to become a valuable acceptability assessment model [87]. If acceptability is taken into account during concept and system development especially in complex environments, this can avoid user resistance [88] and avoid negative use of the system such as obstruction or under-utilization [89]. Acceptability can be measured via Likert attitude scales [90] in questionnaires. A widely used questionnaire with 12 items [91] has high reliability [92].

Those aspects of feasibility can be assessed early in the applied "user-centered design" process. User-centered design encompasses the involvement of all relevant stakeholders in iterative design steps having an appropriate view on the requirements of tasks, users, and task distribution [73]. If user-centered design is applied, this can result in better acceptance of users [93], higher satisfaction [94], improved usability [95], and less training needs [96]. The iterative design loop of DIN EN ISO 9241–210 includes an analysis of system usage context, followed by a deduction of requirements and the development of a design solution that is evaluated afterwards. Non-satisfied user requirements of the evaluation tests lead back to the beginning of the design loop. This methodology has also been applied to certain extents to other ATC interfaces [97] next to TriControl.

2.3. Multimodal CWP TriControl

Nowadays, an ATCO usually issues verbal clearances to pilots via radio telephony and enters the clearances' contents manually into the ATC system. The clearances include structured information about the necessary pilot actions. The central contents are an aircraft callsign, a command type, and a command value. Pilot actions after readback of the clearance can lead to trajectory changes of the aircraft, e.g., due to speed, altitude, and heading changes; or can be of organizational manner, e.g., to handover the controlling responsibility for an aircraft to an ATCO of the adjacent airspace sector.

When using the multimodal CWP TriControl—combining three input modalities—an ATCO is able to generate a clearance with an aircraft callsign via gaze, a command type via multitouch gesture, and a command value via verbal utterance (see Figure 1).

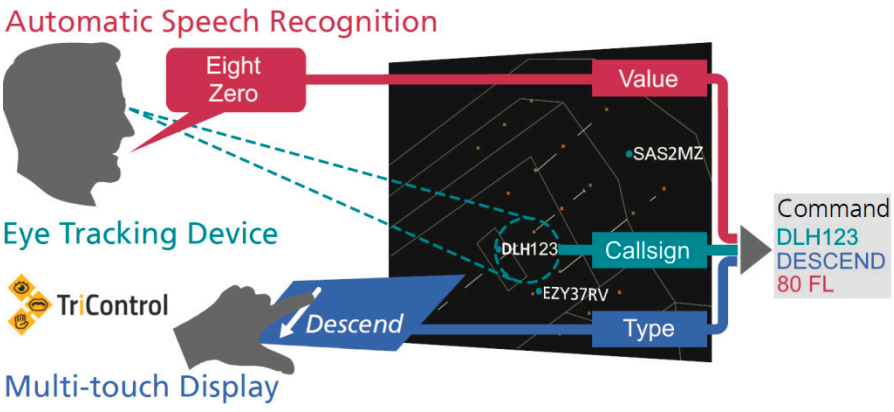


Figure 1. Multimodal interaction with TriControl combining the inputs from eye-tracking via gaze, automatic speech recognition via utterance, and multitouch display via gesture to a controller command (adapted from [98]).

The eye-tracking device detects the spot at the radar screen the ATCO is fixing for a certain dwell time. If there is an aircraft label or icon, the respective callsign is selected as the aircraft that will then receive the next command. Afterwards, the ATCO can input the command type and value in sequence or in parallel. The targeted aircraft is locked as soon as the system detects that the ATCO performs a

gesture or utters something, to avoid the clearance being sent to another aircraft that may be looked at thereafter.

The command type is entered into the system via two-dimensional gestures on the multitouch display. The one-finger gestures swipe down results in "descend", swipe up for "climb", swipe left for "reduce", swipe right for "increase", and long press for "cleared ILS/handover/direct to" depending on the value. When rotating semi-circle wise with two fingers, this will be recognized as a "heading" type. In an additional multitouch interaction mode that can be activated and deactivated with a button press on the multitouch device, some aspects of the graphical user interface of TriControl can be adjusted. This is done with two further multifinger gestures: Five fingers are used to zoom in and out on the radar map via spreading/contracting or they are just moved to pan the map; two fingers are needed to attach the distance measurement tool circles on two selectable aircraft.

The command value is spoken and recorded with a microphone. This value only consists of digits or is a waypoint, runway, or controller position name, respectively, i.e., "one five zero", "two hundred", "delta lima four five five", "two three right", "tower", etc. The command type gesture and command value utterance can be entered in parallel as Figure 2 demonstrates.

The callsign, type, and value are then combined to a controller command displayed to the ATCO. The uttered value is displayed in yellow in the type field of the corresponding aircraft label. For the validation trial, there was an additional top bar on the radar screen [99] with the three input elements next to a yellow value in the corresponding command type label field of the respective aircraft as shown in Figure 2. This visualization of the generated controller command before issuing it helps to detect mistakenly entered or falsely recognized command parts. Thus, the controller can either completely cancel the clearance or overwrite a wrong callsign as detected by gaze recognition, a wrong command type as analyzed by the multitouch device, or a wrong value as recognized by the speech recognition. Hence, there is even one more additional manual check for correctness with TriControl than just to listen to the pilot readback to determine if a conventional—completely verbal—clearance might contain an error.

If the ATCO acknowledges the completed command via a confirmation tap on the multitouch device, it is entered into the ATC system and could be further processed to influence the aircraft trajectory. Hence, the command could be sent to the aircraft via datalink or could be read by an artificial voice via the usual radio telephony channel. More details on the background of TriControl as well as functionalities especially with respect to aspects of command element input orders and timing can be found in [10].

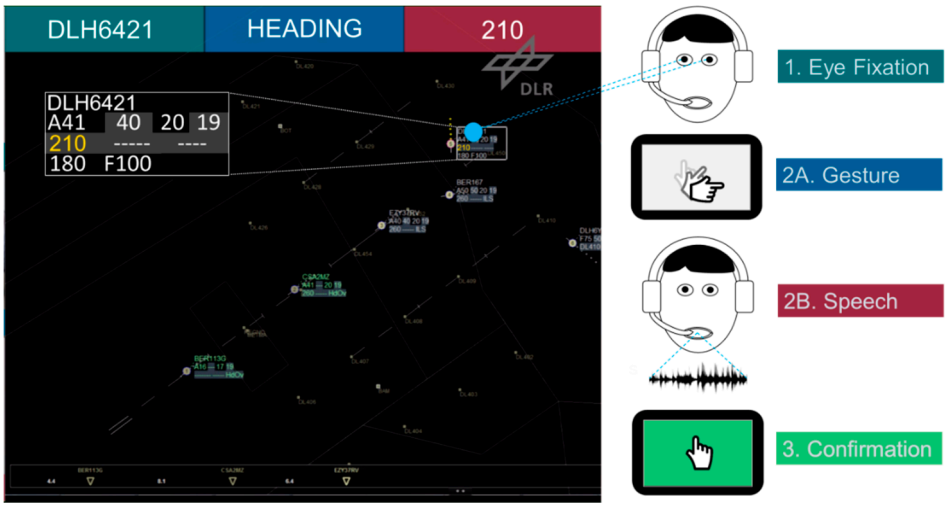


Figure 2. Input order of TriControl starting with eye fixation (on label of callsign DLH6421), followed by potentially parallel touch gesture (rotating two-finger semi-circle swipe) and speech utterance (two one zero) being recognized, terminated with a confirmation tap (short press) to finalize a command.

3. Multimodal CWP Feasibility Analysis

The participants, setup, and tasks of the feasibility analysis study as well as questionnaires and study hypotheses are explained in the following sections.

3.1. Evaluation Site and Study Participants' Characteristics

The TriControl human-in-the-loop feasibility analysis study took place at the German air navigation service provider DFS Deutsche Flugsicherung GmbH in Langen, Germany in April 2017. Fourteen DFS ATCOs with an average age of 47 years (standard deviation (SD): 10 years) participated as study subjects. They had an average professional ATC experience of 21 years (SD: 12 years). The ATCOs worked at different positions, such as approach (7xAPP), area control center (5xACC), upper area center (2xUAC), tower (4xTWR), and as a generic instructor, so multiple answers were possible, and multiple perspectives were obtained. Four ATCOs were identified as the core target group being active APP ATCOs. This is due to the fact that TriControl is an approach control CWP, and current controlling skills, i.e., not being retired, influences the performance during the simulation study.

Some characteristics and experiences of ATCOs were asked as this could be relevant for the efficient usage of the different input modalities. Five participants had previous experience with eye-tracking, 3 with gesture-based inputs, and 10 with automatic speech recognition. Four participants did not use vision correction devices, two participants wore contact lenses, and eight participants wore glasses. Twelve of 14 participants were right-handed. All participants had appropriate English language skills as needed for air traffic control. The participants' native languages were German (9), English (3), Hungarian (1), and Hindi (1).

3.2. Tasks during the Human-in-the-Loop Study for Feasibility Analysis

The complete study included four different phases: Introduction, training, simulation run, and evaluation with debriefing. The TriControl concept and functionalities were described during the 15 min introduction. This included a standardized presentation about project goals as well as the system handling with the three interaction modalities and the graphical user interface taught by the technical supervisor.

The training phase consisted of a practice human-in-the-loop simulation run and lasted roughly 15 min. The traffic scenario used Düsseldorf approach airspace and comprised less aircraft than the later evaluation trial. This gave study participants time for repetition and familiarization with multiple and very different command inputs. Furthermore, they could focus on gathering information using the new radar screen environment. In addition, the eye-tracking device was calibrated to the participants' physical requirements. This phase was accompanied by the technical and psychological supervisors who answered open questions and corrected possible mistakes. As soon as the study participants stated his/her comfort and confidence with the system, the practice run was finished.

The simulation run in which participants worked with the TriControl CWP lasted a bit more than 30 min. The hardware comprised commercial-off-the-shelf devices (laptops, monitor, touchpad, eye-tracker, headset, foot switch). The Düsseldorf approach area with only active runway 23R was used as simulation setup (see Figure 3).

The air traffic scenario comprised 38 arriving aircraft including seven of wake turbulence category "Heavy" and 31 "Medium". The scenario did not encompass departure traffic and was sufficient for one-hour maximum simulation time. Each participant's task was to work as a "Complete Approach" controller (meaning combined pickup/feeder ATCO in Europe and combined feeder/final ATCO in the US, respectively). The traffic scenario used standard arrival routes and there were no unusual traffic or weather conditions. The aircraft followed the issued command instructions directly after confirmation. Hence, the participants got an impression of TriControl's functional mechanisms in a standard ATC approach environment.

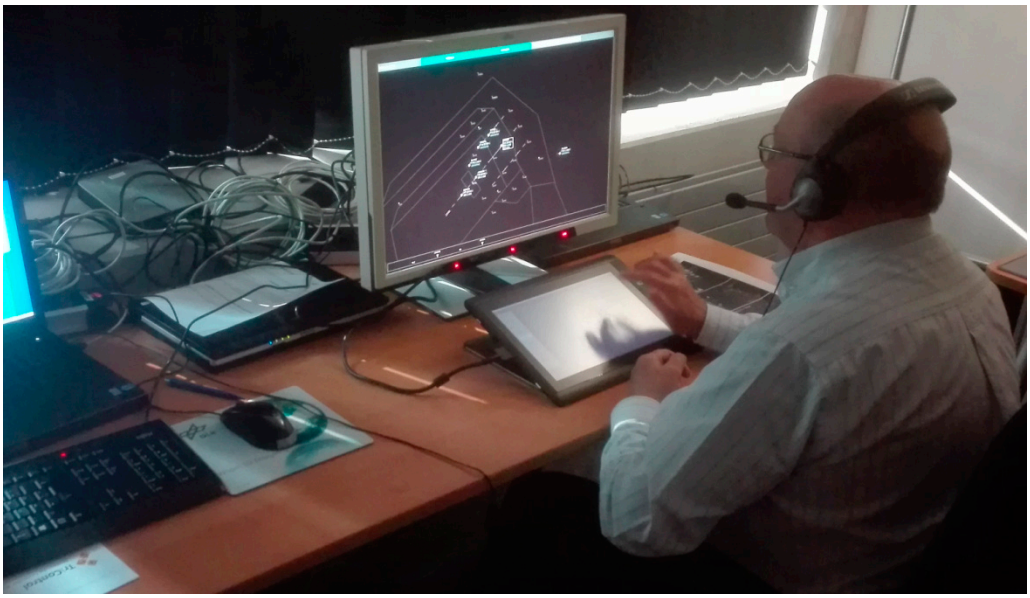


Figure 3. Setup of TriControl feasibility simulation study with a participant before uttering a command value, after performing a command type gesture and selecting an aircraft with his eyes.

During the final evaluation and debriefing phase, all study participants needed to fill out questionnaires regarding feasibility with usability and acceptability, demographics, and profession-related data in the presence of the psychological supervisor. More precisely, 10 questions about personal data as well as 146 statements comprising the system usability scale (SUS) as well as the topics TriControl concept (T), eye-tracking (E), clearances (C), gestures (G), speech recognition (S), input procedure (I), and radar screen (R) plus 30 lines for optional comments on certain elements needed to be handled. Examples for those statements contain the ability to guide air traffic with TriControl (topic T), usefulness of eye-tracking (topic E), ability to issue different command types (topic C), learnability of gestures (topic G), user-friendliness of speech recognition (topic S), satisfaction with command input procedure (topic I), and identification of radar information (topic R) (see Appendix A for all statements to be rated). Together with further 17 categories for notes taken by the psychological supervisor during the experiment, this sums up to 203 lines of raw data for each of the 14 participants.

Three classes of requirements have been defined for the feasibility analysis of TriControl: (1) Multimodal interface fitness for intended use in the "TriControl concept", (2) "information retrieval", and (3) "command issuing". The developed questionnaires apply the norm DIN EN ISO 9241–11 (2017) with the subcategories effectiveness, efficiency, and satisfaction, as well as acceptability. For class (1—TriControl concept), the category effectiveness consisted of controlling air traffic as the core task of an ATCO. The category efficiency orients on DIN EN ISO 9241–11 Dialogue Principles (2006) for HMIs. The general requirements of this main norm were used for the category satisfaction. The category acceptability is assessed with widely used items of system use aspects [86]. For class (2—information retrieval), category effectiveness took the retrieving of information into account, category efficiency bases on DIN EN ISO 9241-12 Presentation of Information (2000), categories satisfaction and acceptability used the same sources as for class (1). For class (3—command issuing), category effectiveness again took the core task of issuing commands into account, categories satisfaction and acceptability consider the E-OCVM 3 demands.

3.3. System Usability and Feasibility Analysis Questionnaire

To answer the research questions on basic feasibility and usability of the TriControl prototype in a quantitative and qualitative way, different assessment approaches were necessary. The quality of the operational concept and system usability of TriControl in general needed to be analyzed with

a globally comparable measure. Therefore, the System Usability Scale (SUS) [100,101]—a subjective system usability assessment tool—was chosen. The SUS questionnaire consists of 10 statements to be rated on a scale comprising five possible answers coded as 0 to 4 points. The statements alter their positive-negative formulation respectively to prevent bias [102]. All ten items are multiplied with 2.5 to span a range from 0 to 100 whereas a higher score indicates better perceived system usability. The SUS proved to be highly reliable with an α of 0.911 [103] and to represent an overall trend [104]. Furthermore, the SUS scale has been used to evaluate TriControl in an earlier phase [9] and thus allows for better comparability and continuing the system usability assessment.

For the current analysis, the SUS score should indicate a sufficient value to represent a system usability as "ok", i.e., be at least at 50.9 as investigated in the literature [104]. Thus, the formal hypotheses on system usability of TriControl are:

$$H_{01}: \overline{x} < 50.9 \quad (1)$$

$$H_{11}: \overline{x} \geq 50.9 \quad (2)$$

The SUS score represents an overall score of ten recorded items [100] that is used for a point estimation. The SUS score is analyzed for a confidence interval (condition $\alpha = 0.05$) for an interval estimation. It also investigates possible significant deviations from the critical cutoff value [79].

The feasibility was tested with a newly developed Likert-scale questionnaire based on user requirements. The self-based assertions aimed to evaluate the single elements of the TriControl system in a systematic manner [105]. The respective scale ranged from 1 (strongly disagree) to 6 (strongly agree) and included two further items (not important) and (not affected) [106]. The newly developed feasibility questionnaire should indicate at least a positive evaluation above the average score of 3.5 on the Likert-scale [90] ranging from 1 to 6 for all items. Hence, the formal hypotheses on feasibility of TriControl's operational concept are:

$$H_{02}: \overline{x} < 3.5 \quad (3)$$

$$H_{12}: \overline{x} \geq 3.5 \quad (4)$$

The non-parametric binomial test was used for the statistical significance analysis due to the small sample size of $N = 14$. However, taking into account the robust binomial distribution supporting the null hypothesis, results will less likely be significant with respect to the desired direction [107]. The binomial test for each item included the n answers actually given by ATCOs, a test ratio of 0.5, an α of 0.05, and an expected mean value of 3.5 as the answers lay within 1 to 6. The further qualitative analysis was structured content-wise to deduct recommendations for certain feasibility elements according to [108].

Additionally, verbal remarks by the study subjects on the human–machine interface during non-task-interfering times were noted in a similar version compared to the Thinking Aloud technique [109]. Furthermore, non-verbal mistakes when using the prototype were noted [110].

4. Results of the Feasibility Study

The questionnaire results on system usability and feasibility as well as the most important comments of the 14 ATCOs are reported in the following sections.

4.1. Score of System Usability Scale (SUS)

The average SUS of TriControl for all 14 ATCOs was 60.9 (SD = 21.9; lower and upper confidence interval limits: 48.3/73.5). Hence, the mean value indicates a system usability between "ok" (50.9) and "good" (71.7) [104]. However, the confidence interval overlapped the cutoff value. So, the mean value does not significantly deviate from the null hypothesis value of 50.9. It has to be rejected that the TriControl prototype offers a valid operational concept for ATC at the current stage. Though, when reducing the sample set to the core target group of active approach (APP) ATCOs ($N = 4$), the results

dramatically improved, i.e., the mean increased to 79.4 (SD = 9.7; lower and upper confidence interval limits: 73.8/85.0). This would indicate a system usability evaluation of TriControl between "good" and "excellent" [104] as shown in Table 1. The SUS score of 79 also equaled an older non-systematic pre-evaluation of TriControl [9]. Table 1 also lists the 10 single SUS items S01–S10 representing a similar result regarding the ratings of active APP ATCOs. They did not perceive TriControl as cumbersome (S08) or inconsistent (S06) but would even like to use the system frequently (S01) with 3.5 points or more on the scale up to 4 points. Furthermore, the four usability statements S11–S14 on the three different input modalities and the combination of it was rated above the scale mean and, again, better from active APP ATCOs.

Table 1. Scores for system usability and four extra statements for single items and total system usability scale (SUS) score.

	I . . .	*N* = 14 (All ATCOs)		*N* = 4 (Active APP ATCOs)	
No.	**System Usability Score Items** [1]	**M**	**SD**	**M**	**SD**
S01	think that I would like to use the system frequently.	2.1	1.5	3.5	0.6
S02	found the system unnecessarily complex. [2]	2.6	1.2	3.3	0.5
S03	thought the system was easy to use.	2.5	1.2	3.3	0.5
S04	think that I would need the support of a technical person to be able to use the system. [2]	2.7	1.3	3.3	1.0
S05	found the various functions in the system were well integrated.	2.3	1.1	3.0	0.8
S06	thought there was too much inconsistency in the system. [2]	2.4	1.2	3.5	0.6
S07	would imagine that most people would learn to use the system very quickly.	2.2	1.1	2.5	1.0
S08	found the system very cumbersome to use. [2]	2.5	1.6	4.0	0.0
S09	felt very confident using the system.	2.1	1.1	3.0	0.0
S10	needed to learn a lot of things before I could get going with the system. [2]	2.9	1.1	2.5	1.7
	Total SUS score	**60.9**	**21.9**	**79.4**	**9.7**
S11	found that TriControl multitouch gestures for command selection are intuitive and easy to learn.	2.8	1.2	3.5	0.6
S12	think that the use of eye-tracking feature for selecting aircraft is disturbing. [2]	2.3	1.4	2.5	1.0
S13	think that automatic speech recognition is a good way to enter values.	2.2	1.4	2.8	1.5
S14	found the use of multiple modalities (eye gaze, gestures, speech) is too demanding. [2]	2.6	1.2	3.0	1.2

[1] Rating per single item from 0 "worst rating" to 4 "best rating", multiplied by 2.5 for Total SUS score. M represents the mean, SD the standard deviation. [2] Statement rating has been "inverted" due to negative formulation, i.e., 0.5 points in the raw data are presented as 3.5 points here to enable better comparability of all items.

4.2. Feasibility Questionnaire Ratings

All 25 statements on the TriControl concept (T) are presented in Table A1. The 44 statements on command input in different categories (E/C/G/S/I) are shown in Table A2. Table A3 lists all 63 statements on the used prototypic radar screen (R). The tables include values for ratings' mean, standard deviation, number of answers, and number of positive answers. They also list the *p*-value of the binomial test for significance analysis, i.e., to assess if the mean value significantly deviates from the null hypothesis

value of 3.5. In roughly 85% of all 132 items of those three tables—especially except in the majority of items in category "R1.2 Coordination"—the rating of the active APP ATCOs was equal or better than the rating of all ATCOs. More than 55% of active APP ATCO ratings, on average, were equal or even above 5 points on the six-point scale. Some meaningful results per category are highlighted in the following.

4.2.1. Ratings on TriControl Concept (T)

The active APP ATCOs rated the statements on the TriControl concept with an average of 4.8 points (on a scale from 1 to 6, see Table A1). Except for the statement on "T6.1 need of suitability for individualization", the active APP ATCO ratings were better than of all ATCOs, i.e., almost one point higher.

ATCO (in particular active APP ATCOs) were able to guide aircraft to their destination in an efficient way following the common safety requirements with TriControl (Controlling T1.1–T1.3). The TriControl interface was rated as appropriate for the intended use. ATCOs—especially with the parallel command input—felt supported to quickly and effectively achieve their best performance (Task Adequacy T2.1–T2.3). All ATCOs were aware of TriControl command input states and knew which and how actions could be executed to perform their controlling tasks due to the average ratings (Self-Descriptiveness T3.1–T3.4). They were also able to intuitively interact with TriControl as it matched common CWP conventions (Expectation Conformity T4.1–T4.2).

Furthermore, the statement ratings on timing and issuing of commands were rated above the scale mean (Controllability T5.1–T5.3). Particularly, active APP ATCOs felt safe to issue commands with little time and mental extra effort in case of a mistake (Error Tolerance T6.1). Active APP ATCOs less likely wanted to be able to adapt TriControl's interface to personal preferences than all ATCOs on average, even though they preferred to have the settings options (Suitability for Individualization T7.1). The satisfaction, notably of active APP ATCOs with TriControl, was good (T8.6). There were high ratings for the ease of use, user-friendliness, and learnability (Satisfaction and Acceptability of TriControl T8.1–T8.8). Some even wished to use TriControl in their daily work if they had the option.

To sum it up, almost all ratings were in the positive half of the scale, indicating a feasible TriControl concept even if not being statistically significant in all cases. Some circumspection existed to state that TriControl is preferred over common ATC interfaces, even in its current prototypic stage.

4.2.2. Ratings on Command Input

Every single active APP ATCO rating on the command input statements was better than that of the group of all ATCOs, i.e., more than 0.7 points better in average on the six-point scale (see Table A2). Almost one-third of the statements even had a significantly positive rating.

Ratings on Eye-Tracking (E)

The eye-tracking modality worked fine for aircraft selection. ATCOs perceived the eye-tracking as useful, user-friendly, as well as easy to use and learn (Aircraft Selection E1.1–E1.2; Satisfaction and Acceptability of the Eye-Tracking Feature E2.1–E2.8). Ratings of active APP ATCOs were mostly in the positive scale range.

Ratings on Clearances (C)

According to the ratings, ATCOs were able to issue each type of clearance that TriControl offers. They also knew the command state they were in and could even simultaneously enter command type and value. Almost all statements were rated statistical significantly positive (Issuing Commands C2.1–C2.9).

Ratings on Gestures (G)

The multitouch gestures to input the command type were perceived as useful, user-friendly, as well as easy to use and learn (Satisfaction and Acceptability of the Gesture based Command type input G2.1–G2.8) and could be an option for ATCOs' daily life CWPs with respect to ratings.

Ratings on Speech Recognition (S)

ATCOs were also satisfied with the automatic speech recognition modality for command value input in average, especially true for the active APP ATCOs (Satisfaction and Acceptability of Speech-Recognition based command value input S2.1–S2.9). Speech recognition was rated as useful, user-friendly, as well as easy to use and learn. Furthermore, the majority did not have problems verbalizing only the value instead of a whole command.

Ratings on Input Procedure (I)

The ratings on the complete command input are very similar to those of single input modalities (Satisfaction and Acceptability of the complete command input procedure I2.1–I2.8). TriControl received positive ratings for usefulness, user-friendliness, as well as ease to use and learn. In particular, active APP ATCOs were satisfied with eye-tracking, multitouch gesture recognition, speech recognition, and command confirmation elements of TriControl.

If all ATCOs are considered, the ratings on daily work use of TriControl and preference over conventional ATC interfaces are only around the scale mean. It is noticeable that all statements on effectiveness of the single interaction modes and TriControl as a compound, respectively (E/G/S/I.2.2 and T8.2) were rated rather negatively below the scale mean of 3.5. As TriControl would replace or support an APP CWP, respectively, it is also not expected that it would be more effective. Controlling air traffic via commands is still possible and is still the valid method to actively guide the traffic. The term effectiveness does not say anything about the efficiency of TriControl. The potential for efficiency gains—comparing TriControl with pure speech commands and following manual ATC system input—has been reported in [10].

4.2.3. Ratings on Radar Screen (R)

The majority of ATCOs' ratings on the radar screen used for TriControl were in the positive scale range and even statistical significantly positive (Aircraft within and aircraft heading to ATCOs' sector, orientation aids, centerline separation range (CSR), information design, as well as satisfaction and acceptability R1.1.1–R6.8). Active APP ATCOs had some difficulties to obtain weight classes and alphanumeric distances at the linker line between two aircraft as the appearance was different from their usual radar screen. The basic radar screen appearance should represent a common state-of-the-art like radar display. This was confirmed by ATCOs as it was usable for monitoring, designed to be user-friendly, and was easy to learn (R6.3–R6.5). The active APP ATCO ratings on the radar screen (see Table A3) were slightly better than the ratings of all ATCOs, i.e., more than two-thirds of the statements had better scores. However, even more than 60% of the statements for all ATCOs also had a significantly positive rating.

For the TriControl concept itself, it was important that ATCOs rated the statement on discriminability between different command states within the aircraft label (inactive, active, received, confirmed) positively (T3.2).

4.3. Feasibility Questionnaire Comments of all ATCOs

On the one hand, there are hints for improvements. Some ATCOs recommended that speech recognition needs to recognize multiple accents better as it did not work reliably for some ATCOs during the simulation trials. One ATCO perceived the foot pedal for speech recording as not helpful. Thus, a button at the headset microphone would be preferred. The technical issue with the non-reliable

confirmation gesture recognition should of course be solved. One ATCO claimed that he disliked the two-finger selection for separation assessment and that the left hand is completely unused. Moreover, the additional graphical user interface mode of the multitouch device offers too few functionalities to deserve its own category. A number of ATCOs reported that the eye-tracking was too slow reacting for them. Furthermore, aircraft have been deselected when looking away during command issue phases. The precision of the mouse has been rated better than eye-tracking especially in cases of vertically separated aircraft with partially overlapping labels. Also, the label itself seems to present too much information. Furthermore, many requests for additional functionalities were formulated such as the option to input combined and conditional clearances, to enter vertical speeds, to differentiate between left and right turns, to see the pilot statements in the aircraft radar labels as well, and a possibility for rotating or moving labels. The labels themselves can be analyzed separately as there are other dedicated studies and developments of labels and label interactivity. The main focus of this paper is on the multimodality.

Some further comments were made with respect to the simulation capabilities and the radar screen layout, which have not been central aspects of the feasibility study. Therefore, ATCOs wanted to have a better trail history and heading needles that are not overlapped by radar labels. Some aircraft did not follow instructions and the descent rates were too slow. Some ATCOs wanted better highlighting of heavy aircraft and information about whether an aircraft is on his/her frequency. Furthermore, a traffic forecast for the next ten minutes would be helpful. For some ATCOs, it was uncommon to work with a dark background radar map and without minimum vector altitude and airspace maps.

One ATCO remarked that the focus would change from "watching the traffic" to "watching if the system complies". Thus, the system feels like an extra step. In addition, there would be a great potential for confusion and errors due to convoluted features. Command issuing via voice was perceived as easier by some of the ATCOs, because it allows better situational awareness. A fallback redundancy would be necessary next to a safety assessment during further system development.

On the other hand, there are many aspects liked by ATCOs. Some ATCOs would prefer TriControl over mouse and screen input. Another ATCO still liked the Paperless Strip System (PSS) better; however, the mouse menu in the labels was perceived to be worse than in TriControl. Other ATCOs remarked that the use of TriControl is fun, it is easy to learn, and that they liked the system. The idea of combining three input methods was appreciated. One ATCO experienced no problem at all. For another, the speech recognition worked fantastically. The eye-tracking input was interesting and worked well after short practice for a number of ATCOs. Hence, it has potential with more development. If the system input was successful, the response is much quicker than with common systems to overall save time due to other ATCOs. It could also lead to fewer misunderstandings. Further thoughts were on the helpfulness in ATC training. An on-the-job-training instructor, who teaches ATCOs to be instructors, noticed that TriControl would be a good system to easily see what the controller is thinking and doing. The centerline separation range support functionality was especially appreciated as it was helpful without needing deduction. It was also reported that the plausibility check of command elements is a good idea and better than the solutions of competitors. Many ATCOs reported that their performance improved with practice and would further improve, so TriControl would be a good aid to ATCOs. They also encouraged following up the project.

5. Discussion of TriControl Feasibility

The system usability of TriControl as rated by all ATCOs was in a range up to a "good" result. This system usability score increased to a range up to "excellent" if considering only active APP ATCOs. However, these values should not be taken as equivalent due to the small sample size of only four active APP ATCOs. It can be a bias indicator with respect to the working positions the ATCOs usually work with. The system usability results were also reflected in the same range by the single system usability statements and the additional items for the specific interaction modalities.

The 132 feasibility analysis statements on TriControl concept, command input, and radar screen were slightly positive, whereas active APP ATCOs again agreed far more positively in the majority of items. Particularly, ATCOs appreciated user-friendliness, usefulness, as well as ease to use and learn.

It is worth mentioning that there were great differences in the TriControl concept ratings depending on some personal and technical abilities of ATCOs during the simulation run. TriControl concept was rated much better by ATCOs—than by other ATCOs—if they:

- Were able to perform parallel input with different modalities,
- Hardly experienced any malfunction with the multitouch pad correspondence,
- Did not forget to perform the confirmation gesture after command completion,
- Did not perform wrong gestures,
- Did not experience some troubles with eye-tracking,
- Experienced more reliable speech recognition,
- Did not make other interaction mistakes, such as:
    - Too long press for confirmation and thus turning into a direct_to command,
    - Forgetting to toggle back from the multitouch device's graphical user interface mode,
    - Pressing the foot pedal for voice recording during complete command creation.

All of the above criteria do fit for the active APP ATCOs. However, it is not completely clear why the four active APP ATCOs performed much better and almost error-free compared to the other 10 ATCOs even if TriControl was designed as an APP CWP. The average age might be an indicator for that, i.e., the four active APP ATCOs were 37 years, the other ATCOs 52 years in average. Assuming that younger people are more familiar with modern interaction technologies from their daily life, this could explain the better ratings of active APP ATCOs. Furthermore, the simulation run time of 30 min might have been too short for ATCOs usually working on other positions to familiarize with the APP environment in addition to the new input modalities.

6. Summary, Conclusions, and Outlook

The feasibility of the multimodal CWP prototype TriControl—integrating eye-tracking, multitouch gesture, and speech recognition for command input—has been analyzed with 14 ATCOs in a human-in-the-loop study. Feasibility, system usability, and acceptability were judged slightly positive. The subgroup of active approach controllers agreed even more positively due to statistical analysis of questionnaire results. They were also motivated to further improve the TriControl system to bring it closer to operational needs as the achieved feasibility scores do not indicate significant showstoppers.

The SESAR2020 (Single European Sky ATM Research Programme) project PJ.16-04 CWP HMI "Workstation, Controller productivity" also dealt with automatic speech recognition, multitouch inputs, and eye-tracking in three different activities. However, those interaction technologies are not combined, but investigated stand-alone. Further research activities on the three interaction technologies will be continued in SESAR2020's wave 2 projects PJ.10-96 "HMI Interaction modes for ATC centre" and PJ.05-97 "HMI Interaction modes for Airport Tower". Hence, there is and will be research on modern interaction technologies in air traffic control. However, TriControl is one of the few concepts integrating multiple promising technologies to extract the benefits of each of them.

Recent iterations of ATC system development in general—and in particular, interface developments—have resulted in significant efficiency gains in the ability to process increased traffic levels, which soon rise in the real world to reach the new system limitations. A significant limitation in all further developments, however, seems to be the "bottleneck" of frequency congestion. The process of getting clearances clearly and safely submitted from the ATCO to aircraft and checking pilot readbacks for correctness is time consuming. The TriControl system seeks to address this with what could potentially be the effective removal of the existing bottleneck, allowing a greatly improved capacity increase.

According to the above study results, and to already revealed increased efficiency potential, further development of the early prototype TriControl will be performed to overcome the revealed malfunctions and integrate many suggestions for improvement. Afterwards, TriControl should be applied in different contexts also comprising non-nominal conditions such as weather influence, high-density air traffic, and emergency aircraft. Then, TriControl's operational concept can be compared with current systems including cognitive workload assessment. Overall, the feasibility analysis motivated to foster multimodal interaction for air traffic control.

7. Patents

TriControl can serve as an input means and usable environment for the command generator European patent application 17158692.8.

Author Contributions: O.O. was responsible for development of the TriControl system. He also conceived and designed the experiments with great organizational support of M.W. O.O. was the technological supervisor and main author of this article being supported by all three co-authors. The psychological supervisor A.S. was responsible for the preparation and conduction of the data analysis (mainly represented in results section) being closely supervised by M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We like to thank all ATCOs that participated in the human-in-the-loop study with TriControl. Besides, we are grateful for the support of Konrad Hagemann (DFS Planning and Innovation) in preparing the simulation trials in Langen. Sebastian Pannasch (Technische Universität Dresden) provided valuable input during initial reviews and for the master thesis contents of Axel Schmugler regarding the feasibility analysis. Thanks also to Malte Jauer (DLR) for assisting during the study and for performing earlier implementations.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix A

Table A1. Evaluation of statements on TriControl core concept in different categories.

				***N* = 14 (AllATCOs)**			***N* = 4 (Active APP ATCOs)**	
No.	**Statement [1]**	**M**	**SD**	**n**	**k**	**p**	**M**	**SD**
	Controlling							
T1.1	I was able to guide the aircraft to their destination.	4.8	1.1	14	11	0.03	5.8	0.5
T1.2	I was able to guide the aircraft following the common safety requirements.	4.5	1.2	14	10	0.09	5.3	1.0
T1.3	I was able to guide the aircraft in an efficient way.	3.4	1.7	14	8	0.40	5.3	0.5
	Task Adequacy							
T2.1	The interface sufficiently supported me achieving my best performance.	3.6	1.5	14	7	0.60	4.3	1.0
T2.2	The parallel command input enabled me to issue commands fast and effectively (type and value).	3.4	2.0	14	7	0.60	5.3	0.5
T2.3	All in all the command procedure was appropriate for its intended use (input and feedback).	4.4	1.4	14	11	0.03	5.0	0.8
	Self-Descriptiveness							
T3.1	I was always aware of the state of use I was currently operating in (monitoring, issuing commands).	3.9	1.1	14	9	0.21	5.0	0.0
T3.2	I was always aware of the state the command input was in (inactive, active, receiving, received, accepted).	3.6	0.9	14	8	0.40	4.0	0.8
T3.3	I always knew which actions I was able to execute at any given moment.	4.4	1.1	14	11	0.03	5.3	0.5

Table A1. *Cont.*

		N = 14 (AllATCOs)					*N* = 4 (Active APP ATCOs)	
No.	**Statement [1]**	**M**	**SD**	**n**	**k**	**p**	**M**	**SD**
	Expectation Conformity							
T4.1	I was always able to intuitively interact with the interface the way I needed to.	3.6	1.6	14	9	0.21	5.0	0.0
T4.2	TriControl matched common conventions of use (content, depictions, specificity of numeric information etc.)	3.6	1.6	14	8	0.40	4.8	0.5
	Controllability							
T5.1	I was able to start the command issuing exactly when I wanted to.	3.8	1.8	14	9	0.21	5.0	0.8
T5.2	I was able to control the command issuing the way I wanted to (proceed, cancel, or confirm).	4.1	1.5	14	11	0.03	5.3	0.5
T5.3	I was able to control the pace at which the commands were entered.	3.6	1.7	14	9	0.21	4.0	0.8
	Error Tolerance							
T6.1	In case of a mistake, a command could still be issued with little extra effort (time and mental effort).	3.8	1.7	13	9	0.13	5.0	0.0
	Suitability for Individualization							
T7.1	I would like to be able to adapt the interface to my personal preferences.	4.2	1.5	9	7	0.09	4.0	2.0
	Satisfaction and Acceptability of TriControl							
T8.1	TriControl is useful for managing routine approach air traffic.	3.6	1.6	14	8	0.40	4.8	1.0
T8.2	Working with TriControl is more effective than working with common interfaces.	2.8	1.4	14	3	0.99	3.3	0.5
T8.3	TriControl is easy to use.	4.3	1.6	14	9	0.21	5.5	1.0
T8.4	TriControl is user friendly.	4.4	1.6	14	10	0.09	5.8	0.5
T8.5	It is easy to learn to use TriControl.	4.9	1.0	14	13	0.00	5.8	0.5
T8.6	Overall, I am satisfied with TriControl.	4.2	1.4	14	10	0.09	5.3	0.5
T8.7	I would want to use TriControl for my daily work if I had the option.	3.3	1.8	12	5	0.81	4.0	2.2
T8.8	I would prefer TriControl over common ATC interfaces.	3.1	1.5	14	6	0.79	3.5	1.3

[1] Rating per single item from 1 "worst rating" to 6 "best rating", other/missing ratings are ignored. M represents the mean, SD is the standard deviation, n is the number of given valid ratings, k is the number of "successful" ratings above the scale mean (>=4), and p is the p-value of the binomial test (1-tailed) that is underlined if equal or below 0.05 to state significance.

Table A2. Evaluation of statements on TriControl command input in different categories.

		N = 14 (AllATCOs)					*N* = 4 (Active APP ATCOs)	
No.	**Statement [1]**	**M**	**SD**	**n**	**k**	**p**	**M**	**SD**
	Aircraft Selection							
E1.1	I was able to select every aircraft I wanted to.	3.8	1.3	14	10	0.09	4.5	0.6
E1.2	Only little effort was needed to select aircraft.	4.0	1.6	13	10	0.05	5.0	0.8

Table A2. *Cont.*

		N = 14 (AllATCOs)					*N* = 4 (Active APP ATCOs)	
No.	**Statement [1]**	**M**	**SD**	**n**	**k**	**p**	**M**	**SD**
	Satisfaction and Acceptability of the Eye-Tracking Feature							
E2.1	The eye-tracking method is useful for aircraft selection.	4.1	1.4	14	11	0.03	5.0	0.0
E2.2	The eye-tracking method works more effectively than conventional aircraft selection methods.	2.9	1.4	14	6	0.79	3.0	1.4
E2.3	The eye-tracking method is easy to use.	4.1	1.5	14	10	0.09	5.0	1.4
E2.4	The eye-tracking method is user-friendly.	3.9	1.4	14	10	0.09	4.8	1.3
E2.5	It is easy to learn to use the eye-tracking method.	4.9	1.0	14	12	0.01	5.5	0.6
E2.6	Overall, I am satisfied with the eye-tracking as a method of aircraft selection.	3.7	1.4	14	9	0.21	4.8	0.5
E2.7	I would want to use it for my daily work if I had the option.	3.3	1.9	14	7	0.60	3.8	2.2
E2.8	I would prefer it over conventional input methods.	3.1	1.8	14	7	0.60	3.3	1.7
	Issuing Commands							
C2.1	I was able to issue altitude clearance.	4.9	0.9	14	13	0.00	5.8	0.5
C2.2	I was able to issue speed clearance.	4.9	0.9	14	13	0.00	5.8	0.5
C2.3	I was able to issue heading clearance.	4.8	1.0	14	13	0.00	5.5	1.0
C2.4	I was able to command heading to a certain waypoint.	5.0	0.7	12	12	0.00	5.7	0.8
C2.5	I was able to command hand over to tower.	4.0	1.8	14	9	0.21	5.0	2.0
C2.6	I was able to identify when I was able to issue commands.	4.8	1.2	13	12	0.00	5.8	0.6
C2.7	I was able to identify when my commands were being received.	4.7	1.3	14	12	0.01	5.8	0.5
C2.8	I was able to identify when my commands were being accepted by the system.	4.8	1.1	14	13	0.00	5.3	1.0
C2.9	I was able to enter command type and command value simultaneously.	3.9	1.5	14	9	0.21	4.5	1.3
	Satisfaction and Acceptability of the Gesture based Command type input							
G2.1	The gesture-based command type input is useful for the input of command types.	4.6	1.2	14	11	0.03	5.5	0.6
G2.2	The gesture-based command type input is more effective than common approaches.	3.2	1.4	14	7	0.60	3.5	1.3
G2.3	The gesture-based command type input is easy to use.	4.4	1.4	14	10	0.09	5.3	0.5
G2.4	The gesture-based command type input method is user friendly.	4.1	1.5	14	9	0.21	5.0	0.8
G2.5	It is easy to learn the gestures.	5.1	0.5	14	14	0.00	5.5	0.6
G2.6	Overall, I am satisfied with the gesture-based command type input.	4.0	1.5	14	9	0.21	5.3	0.5
G2.7	I would want to use it for my daily work if I had the option.	3.4	1.4	14	7	0.60	4.0	1.2
G2.8	I would prefer it over common methods of command type input.	3.1	1.4	14	6	0.79	3.5	1.3
	Satisfaction and Acceptability of Speech-Recognition based command value input							
S2.1	Speech recognition is useful for the input of command values.	4.2	1.4	14	9	0.21	5.3	0.5

Table A2. *Cont.*

		N = 14 (AllATCOs)					*N* = 4 (Active APP ATCOs)	
No.	**Statement [1]**	**M**	**SD**	**n**	**k**	**p**	**M**	**SD**
S2.2	The speech recognition command value input is more effective than common approaches.	3.3	1.3	13	6	0.71	4.3	1.2
S2.3	The speech recognition is easy to use.	4.0	1.6	13	8	0.29	5.3	0.5
S2.4	The speech recognition-based command value input is user friendly.	4.1	1.5	14	9	0.21	5.3	0.5
S2.5	It is easy to learn to use the speech recognition.	4.4	1.4	14	10	0.09	5.3	0.5
S2.6	It was easy to get used to only verbalize the command value and not the whole command.	4.3	1.4	14	11	0.03	4.8	0.5
S2.7	Overall, I am satisfied with the speech recognition-based command value input.	3.8	1.3	14	7	0.60	4.8	0.5
S2.8	I would want to use it for my daily work if I had the option.	3.2	1.6	14	6	0.79	3.8	1.5
S2.9	I would prefer it over common methods of command value input.	3.0	1.2	14	6	0.79	3.3	1.0
	Satisfaction and Acceptability of the complete command input procedure							
I2.1	TriControl command input procedure is useful for issuing commands.	4.3	1.3	14	10	0.09	4.8	0.5
I2.2	The command input procedure is more effective than common approaches for command issuing.	2.9	1.3	14	5	0.91	3.0	1.4
I2.3	TriControl's command input procedure is easy to use.	4.2	1.6	14	10	0.09	5.0	0.8
I2.4	The combination of eye-tracking, gestures, speech recognition and confirmation is user friendly.	3.8	1.7	14	8	0.40	4.8	1.3
I2.5	It is easy to learn to use the command input procedure.	4.7	1.1	14	13	0.00	5.0	0.8
I2.6	Overall, I am satisfied with the command input procedure.	3.9	1.4	14	9	0.21	5.0	0.0
I2.7	I would want to use the command input procedure for my daily work if I had the option.	3.1	1.4	14	7	0.60	3.8	1.0
I2.8	I would prefer the command input procedure over common methods of command value input.	2.6	1.3	14	4	0.97	2.8	1.0

[1] Rating per single item from 1 "worst rating" to 6 "best rating", other/missing ratings are ignored. M represents the mean, SD is the standard deviation, n is the number of given valid ratings, k is the number of "successful" ratings above the scale mean (>=4), and p is the p-value of the binomial test (1-tailed) that is underlined if equal or below 0.05 to state significance.

Table A3. Evaluation of statements on radar screen prototypic design used for TriControl in different categories.

		N = 14 (AllATCOs)					*N* = 4 (Active APP ATCOs)	
No.	**Statement [1]**	**M**	**SD**	**n**	**k**	**p**	**M**	**SD**
	Aircraft within my sector: Identification							
R1.1.1	I was able to identify every aircraft's presence.	4.9	1.2	14	12	0.01	5.3	0.5
R1.1.2	I was able to identify every aircraft's location.	5.1	0.9	14	13	0.00	5.5	0.6
R1.1.3	I was able to identify every aircraft's call sign.	5.4	0.6	14	14	0.00	5.8	0.5
R1.1.4	I was able to identify every aircraft's weight class.	4.0	1.8	14	7	0.60	3.3	1.9

Table A3. *Cont.*

		N = 14 (AllATCOs)					**N = 4 (Active APP ATCOs)**	
	Aircraft within my sector: Coordination							
R1.2.1	I was able to obtain information regarding every aircraft's altitude.	4.7	1.1	14	12	0.01	4.5	1.7
R1.2.2	I was able to obtain information regarding every aircraft's cleared altitude.	4.6	1.1	14	12	0.01	4.5	1.7
R1.2.3	I was able to obtain information regarding every aircraft's speed.	4.7	1.1	14	12	0.01	4.5	1.7
R1.2.4	I was able to obtain information regarding every aircraft's cleared speed.	4.6	1.1	14	12	0.01	4.5	1.7
R1.2.5	I was able to obtain information regarding every aircraft's heading.	4.6	1.1	14	12	0.01	4.5	1.7
R1.2.6	I was able to obtain information regarding every aircraft's cleared heading.	4.6	1.1	14	12	0.01	4.5	1.7
R1.2.7	I was able to obtain information regarding every aircraft's next selected waypoint.	4.6	1.1	13	11	0.01	4.3	2.1
R1.2.8	I was able to obtain information regarding every aircraft's distance to another aircraft.	3.8	1.5	14	9	0.21	2.8	2.2
R1.2.9	I was able to obtain information regarding every aircraft's sequence number suggested by the AMAN.	4.3	1.3	11	8	0.11	5.0	0.0
R1.2.10	I was able to obtain information regarding every aircraft's miscellaneous information (Cleared ILS, Handover to Tower).	4.2	1.4	13	8	0.29	4.0	1.8
	Aircraft heading into my sector: Identification							
R2.1.1	I was able to obtain the information that aircraft were heading into my sector.	4.0	1.1	11	7	0.27	5.0	0.0
R2.1.2	I was able to obtain the information how many aircraft were heading into my sector.	3.8	1.2	12	6	0.61	4.3	1.2
R2.1.3	I was able to obtain the call sign of every aircraft heading into my sector.	4.8	0.8	13	12	0.00	5.3	0.5
	Aircraft heading into my sector: Coordination							
R2.2.1	I was able to obtain every aircraft's estimated time of arrival (ETA).	2.5	0.5	6	0	0.99	-	-
R2.2.2	I was able to obtain every aircraft's point of entry.	2.4	1.0	7	1	0.99	-	-
	Orientation Aids							
R3.1	I was able to obtain the runway location.	5.3	0.6	13	13	0.00	5.3	0.6
R3.2	I was able to obtain the runway orientation.	5.4	0.5	13	13	0.00	5.3	0.6
R3.3	I was able to obtain the extended runway centerline.	5.4	0.5	14	14	0.00	5.3	0.5
R3.4	I was able to obtain the standard arrival routes (STAR).	5.2	0.6	11	11	0.00	5.3	0.6
R3.5	I was able to obtain the borders of my airspace sector.	5.1	0.8	11	10	0.01	5.3	0.6
R3.6	I was able to obtain GPS waypoints.	5.3	0.5	14	14	0.00	5.3	0.5
	The Centerline Separation Range							
R4.1	I was able to obtain the location of aircraft in final descend.	4.8	1.0	13	11	0.01	5.3	0.5
R4.2	I was able to obtain the separation between aircraft and neighboring elements (runway, different aircraft).	4.8	1.0	13	11	0.01	5.3	0.5

Table A3. *Cont.*

		N = 14 (AllATCOs)					**N = 4 (Active APP ATCOs)**	
R4.3	I was able to obtain the weight class of aircraft in final descend.	4.3	1.6	13	9	0.13	3.5	2.4
	Information Design: Clarity							
R5.1.1	I was able to obtain all information quickly.	4.1	1.3	14	11	0.03	4.3	1.7
R5.1.2	All information is as specific as I need it to be.	4.0	1.1	14	10	0.09	4.3	1.0
	Information Design: Discriminability							
R5.2.1	I was able to discriminate between different radar screen elements in general.	4.7	0.9	14	12	0.01	5.0	0.8
	Information Design: Discriminability—Aircraft							
R5.2.1.1	I was able to easily discriminate between different aircraft within my sector.	4.8	1.1	14	13	0.00	5.5	0.6
R5.2.1.2	I was able to easily discriminate between different aircraft heading into my sector.	4.6	1.1	14	12	0.01	5.3	1.0
R5.2.1.3	I was able to easily discriminate between different information within the label.	4.5	1.2	14	11	0.03	5.3	1.0
R5.2.1.4	I was able to easily discriminate between different command states within the aircraft label (Inactive, active, received, confirmed).	4.1	1.5	14	9	0.21	5.0	0.8
R5.2.1.5	I was able to easily discriminate between different indicated weight classes.	4.1	1.2	13	9	0.13	3.8	1.7
R5.2.1.6	I was able to easily discriminate between different Arrival Manager order suggestions.	3.8	1.5	10	7	0.17	5.0	1.4
R5.2.1.7	I was able to easily discriminate between different heading directions.	3.9	1.4	14	10	0.09	4.3	1.7
	Information Design: Discriminability—Orientation Aids and Centerline Separation Range (CSR)							
R5.2.2.1	I was able to easily discriminate between different categories of orientation aids in general.	4.7	0.8	11	10	0.01	4.0	1.4
R5.2.2.2	I was able to easily discriminate between different GPS waypoints.	5.0	0.8	11	11	0.00	5.3	0.8
R5.2.2.3	I was able to easily discriminate between different runways.	5.2	0.6	10	10	0.00	5.3	0.8
R5.2.2.4	I was able to easily discriminate between different aircraft on the CSR.	5.3	0.7	8	8	0.00	5.5	0.7
R5.2.2.5	I was able to easily discriminate between different distances between aircraft on the CSR.	5.0	0.9	8	7	0.04	5.3	0.6
	Information Design: Consistency							
R5.3.1	The format of the information given was consistent with what I expected it to be.	4.4	1.0	14	13	0.00	4.8	1.0
	Information Design: Compactness							
R5.4.1	I obtained all the information I needed to monitor the area effectively.	4.0	1.5	14	10	0.09	5.0	0.8
R5.4.2	The radar screen didn't present any unnecessary information.	4.5	0.9	14	12	0.01	4.3	1.0
	Information Design: Detectability							
R5.5.1	I was able to direct my attention towards the currently necessary information.	3.9	1.4	14	10	0.09	4.8	0.5
R5.5.2	The radar screen didn't divert my attention towards currently unnecessary information.	4.4	1.2	14	11	0.03	5.3	1.0

Table A3. *Cont.*

		N = 14 (AllATCOs)					N = 4 (Active APP ATCOs)	
Information Design: Readability								
R5.6.1	I was able to easily read alphanumeric information concerning the aircraft.	4.9	1.0	14	13	0.00	5.5	0.6
R5.6.2	I was able to easily read alphanumeric information concerning the orientation aids.	5.1	0.7	14	13	0.00	5.5	0.6
R5.6.3	I was able to easily read alphanumeric information within the CSR.	5.2	0.6	11	11	0.00	5.5	0.7
Information Design: Comprehensibility of coded meaning								
R5.7.1	I was able to easily understand the coded information in general.	4.5	1.1	13	11	0.01	5.3	0.6
R5.7.2	I perceived the used coding of information as unambiguous.	3.9	1.4	12	8	0.19	4.0	1.7
R5.7.3	I was able to easily interpret all used codes.	4.1	1.3	13	9	0.13	5.3	0.6
R5.7.4	I found it easy to deduce the coded meaning of the given information.	4.2	1.3	12	9	0.07	5.3	0.6
Satisfaction and acceptability of the radar screen								
R6.1	The information design used in the radar screen is useful for sector monitoring.	4.1	0.8	14	12	0.01	4.0	0.8
R6.2	The radar screen depicts information more effectively than conventional models.	2.8	1.2	13	3	0.99	3.0	1.4
R6.3	The radar screen is easy to use for monitoring.	4.1	0.9	14	11	0.03	4.3	1.0
R6.4	The radar screen design is user friendly.	4.2	1.0	13	11	0.01	4.5	1.0
R6.5	It was easy to learn to use the radar screen.	4.4	1.0	14	13	0.00	4.5	1.7
R6.6	Overall, I am satisfied with the radar screen information design.	3.9	1.2	14	10	0.09	3.8	2.1
R6.7	I would want to use it for my daily work if I had the option.	3.0	1.3	13	6	0.71	3.0	1.4
R6.8	I would prefer it over conventional radar screen designs.	2.7	1.2	13	4	0.95	2.8	1.0

[1] Rating per single item from 1 "worst rating" to 6 "best rating", other/missing ratings are ignored. M represents the mean, SD is the standard deviation, n is the number of given valid ratings, k is the number of "successful" ratings above the scale mean (>=4), and p is the p-value of the binomial test (1-tailed) that is underlined if equal or below 0.05 to state significance.

References

1. Quek, F.; McNeill, D.; Bryll, R.; Kirbas, C.; Arslan, H.; McCullough, K.E.; Furuyama, N.; Ansari, R. Gesture, speech, and gaze cues for discourse segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2000 (Cat. No.PR00662), Hilton Head Island, SC, USA, 15 June 2000; Volume 2, pp. 247–254.
2. Oviatt, S.L. Multimodal interfaces. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*; CRC Press: Boca Raton, FL, USA, 2003; pp. 286–304.
3. Oviatt, S.L. Advances in Robust Multimodal Interface Design. *IEEE Comput. Graph. Appl.* **2003**, *23*, 62–68. [CrossRef]
4. Koons, D.B.; Sparrell, C.J.; Thorisson, K.R. Integrating simultaneous input from speech, gaze, and hand gestures. In *Intelligent Multimedia Interfaces*; Maybury, M.T., Ed.; American Association for Artificial Intelligence: Menlo Park, CA, USA, 1993; pp. 257–276.

5. Uebbing-Rumke, M.; Gürlük, H.; Jauer, M.-L.; Hagemann, K.; Udovic, A. Usability evaluation of multi-touch displays for TMA controller working positions. In Proceedings of the 4th SESAR Innovation Days, Madrid, Spain, 25–27 November 2014.
6. Gürlük, H.; Helmke, H.; Wies, M.; Ehr, H.; Kleinert, M.; Mühlhausen, T.; Muth, K.; Ohneiser, O. Assistant Based Speech Recognition—Another Pair of Eyes for the Arrival Manager. In Proceedings of the 34th Digital Avionics Systems Conference (DASC), Prague, Czech Republic, 13–17 September 2015.
7. Helmke, H.; Ohneiser, O.; Mühlhausen, T.; Wies, M. Reducing Controller Workload with Automatic Speech Recognition. In Proceedings of the 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 25–29 September 2016.
8. Möhlenbrink, C.; Papenfuß, A. Eye-data metrics to characterize tower controllers' visual attention in a multiple remote tower exercise. In Proceedings of the ICRAT, Istanbul, Turkey, 26–30 May 2014.
9. Ohneiser, O.; Jauer, M.-L.; Gürlük, H.; Uebbing-Rumke, M. TriControl—A Multimodal Air Traffic Controller Working Position. In Proceedings of the 6th SESAR Innovation Days, Delft, The Netherlands, 8–10 November 2016.
10. Ohneiser, O.; Jauer, M.-L.; Rein, J.R.; Wallace, M. Faster Command Input Using the Multimodal Controller Working Position "TriControl". *Aerospace* **2018**, *5*, 54. [CrossRef]
11. Tiewtrakul, T.; Fletcher, S.R. The challenge of regional accents for aviation English language proficiency standards: A study of difficulties in understanding in air traffic control-pilot communications. *Ergonomics* **2010**, *2*, 229–239. [CrossRef] [PubMed]
12. ICAO. The Second Meeting of the Regional Airspace Safety Monitoring Advisory Group (RASMAG/2). 2004. Available online: https://www.icao.int/Meetings/AMC/MA/2004/RASMAG2/ip03.pdf (accessed on 14 January 2020).
13. Chatty, S.; Lecoanet, P. Pen Computing for Air Traffic Control. In Proceedings of the CHI'96: SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, 13–18 April 1996; pp. 87–94.
14. Schmugler, A. Feasibility Analysis of the Multimodal Air Traffic Controller Working Position Prototype "TriControl". Master's Thesis, Technische Universität Dresden, Dresden, Germany, 2018.
15. Czaja, S.J.; Nair, S.N. Human Factors Engineering and Systems Design. In *Handbook of Human Factors and Ergonomics*; Salvendy, G., Ed.; John Wiley & Sons: Hoboken, NJ, USA, 2006. [CrossRef]
16. Bernsen, N. Multimodality Theory. In *Multimodal User Interfaces. Signals and Communication Technologies*; Tzovaras, D., Ed.; Springer: Berlin/Heidelberg, Germany, 2008. [CrossRef]
17. Adkar, P. Unimodal and Multimodal Human Computer Interaction: A Modern Overview. *Int. J. Comput. Sci. Inf. Eng. Technol.* **2013**, *2*, 8–15.
18. Norman, D.A. Design Rules Based on Analysis of Human Error. *Commun. ACM* **1983**, *4*, 254–258. [CrossRef]
19. Nachreiner, F.; Nickel, P.; Meyer, I. Human factors in process control systems: The design of human-machine interfaces. *Saf. Sci.* **2006**, *44*, 5–26. [CrossRef]
20. Sheridan, T.B. Humans and Automation. System Design and Research Issues. In *Wiley Series in System Engineering and Management: HFES Issues in Human Factors and Ergonomics Series*; Human Factors and Ergonomics Society, Ed.; John Wiley & Sons Inc.: Hoboken, NJ, USA, 2002; Volume 3.
21. EUROCONTROL. *Integrated Task and Job Analysis of Air Traffic Controllers—Phase 2: Task Analysis of En-route Controllers*; EUROCONTROL: Brussels, Belgium, 1999.
22. EUROCONTROL. *Integrated Task and Job Analysis of Air Traffic Controllers—Phase 3: Baseline Reference of Air Traffic Controller Task and Cognitive Processes in the ECAC Area*; EUROCONTROL: Brussels, Belgium, 2000.
23. Cardosi, K.M.; Brett, B.; Han, S. *An Analysis of TRACON (Terminal Radar Approach Control) Controller-Pilot Voice Communications*; DOT/FAA/AR-96/66; DOT FAA: Washington, DC, USA, 1996.
24. Proctor, R.W.; Vu, K.-P.L. Human Information Processing: An Overview for Human-Computer Interaction. In *Human Computer Interaction Fundamentals*; Sears, A., Jacko, J.A., Eds.; CRC Press: Boca Raton, FL, USA, 2009; pp. 19–38.
25. Oviatt, S.L. Human-centered design meets cognitive load theory: Designing interfaces that help people think. In Proceedings of the 14th Annual ACM international Conference on Multimedia, New York, NY, USA, 23–27 October 2006; pp. 871–880.
26. Baddeley, A.D. Working Memory. *Science* **1992**, *255*, 556–559. [CrossRef]
27. Bolt, R.A. Put-that-there: Voice and gesture at the graphics interface. *Comput. Graph.* **1980**, *3*, 262–270. [CrossRef]

28. Nigay, L.; Coutaz, J. A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion. In Proceedings of the INTERCHI'93 Conference on Human Factors in Computing Systems, Amsterdam, The Netherlands, 24–29 April 1993; pp. 172–178.
29. Bourguet, M.L. Designing and Prototyping Multimodal Commands. In Proceedings of the Human-Computer Interaction INTERACT'03, Zurich, Switzerland, 1–5 September 2003; pp. 717–720.
30. Oviatt, S.L. Breaking the Robustness Barrier: Recent Progress on the Design of Robust Multimodal Systems. *Adv. Comput.* **2002**, *56*, 305–341.
31. Manawadu, E.U.; Kamezaki, M.; Ishikawa, M.; Kawano, T.; Sugano, S. A Multimodal Human-Machine Interface Enabling Situation-Adaptive Control Inputs for Highly Automated Vehicles. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; pp. 1195–1200.
32. Pentland, A. Perceptual Intelligence. *Commun. ACM* **2000**, *4*, 35–44. [CrossRef]
33. Seifert, K. Evaluation of Multimodal Computer Systems in Early Development Phases, Original German Title: Evaluation Multimodaler Computer-Systeme in Frühen Entwicklungsphasen. Ph.D. Thesis, Technische Universität Berlin, Berlin, Germany, 2002. [CrossRef]
34. Oviatt, S.L. Multimodal interactive maps: Designing for human performance. *Hum. Comput. Interact.* **1997**, *12*, 93–129.
35. Cohen, P.R.; McGee, D.R. Tangible multimodal interfaces for safety-critical applications. *Commun. ACM* **2004**, *1*, 1–46. [CrossRef]
36. den Os, E.; Boves, L. User behaviour in multimodal interaction. In Proceedings of the HCI International, Las Vegas, NV, USA, 22–27 July 2005; Available online: http://lands.let.ru.nl/literature/boves.2005.2.pdf (accessed on 14 January 2020).
37. Shi, Y.; Taib, R.; Ruiz, N.; Choi, E.; Chen, F. Multimodal Human-Machine Interface and User Cognitive Load Measurement. *Proc. Int. Fed. Autom. Control* **2007**, *40*, 200–205. [CrossRef]
38. Oviatt, S. User-centered modeling for spoken language and multimodal interfaces. *IEEE Multimed.* **1996**, *4*, 26–35. [CrossRef]
39. Oviatt, S.L. Mutual disambiguation of recognition errors in a multimodal architecture. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems, Pittsburgh, PA, USA, 15–20 May 1999; pp. 576–583.
40. Oviatt, S.L. Ten myths of multimodal interaction. *Commun. ACM* **1999**, *11*, 74–81. [CrossRef]
41. Oviatt, S.L.; Coulston, R.; Lunsford, R. When do we interact multimodally? Cognitive load and multimodal communication patterns. In Proceedings of the 6th International Conference on Multimodal interfaces, State College, PA, USA, 13–15 October 2004; pp. 129–136.
42. Oviatt, S.L.; Coulston, R.; Tomko, S.; Xiao, B.; Lunsford, R.; Wesson, M.; Carmichael, L. Toward a theory of organized multimodal integration patterns during human-computer interaction. In Proceedings of the ICMI 5th International Conference on Multimodal Interfaces, Vancouver, BC, Canada, 5–7 November 2003; pp. 44–51.
43. Marusich, L.R.; Bakdash, J.Z.; Onal, E.; Yu, M.S.; Schaffer, J.; O'Donovan, J.; Höllerer, T.; Buchler, N.; Gonzalez, C. Effects of information availability on command-and-control decision making performance, trust, and situation awareness. *Hum. Factors* **2016**, *2*, 301–321. [CrossRef] [PubMed]
44. Connolly, D.W. *Voice Data Entry in Air Traffic Control*; Report N93-72621; National Aviation Facilities Experimental Center: Atlantic City, NJ, USA, 1977.
45. ICAO. *ATM (Air Traffic Management): Procedures for Air Navigation Services*; DOC 4444 ATM/501; International Civil Aviation Organization (ICAO): Montréal, QC, Canada, 2007.
46. Helmke, H.; Oualil, Y.; Schulder, M. Quantifying the Benefits of Speech Recognition for an Air Traffic Management Application. *Konferenz Elektronische Sprachsignalverarbeitung*. 2017, pp. 114–121. Available online: http://essv2017.coli.uni-saarland.de/pdfs/Helmke.pdf (accessed on 14 January 2020).
47. Helmke, H.; Slotty, M.; Poiger, M.; Herrer, D.F.; Ohneiser, O.; Vink, N.; Cerna, A.; Hartikainen, P.; Josefsson, B.; Langr, D.; et al. Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ.16-04. In Proceedings of the IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), London, UK, 23–27 September 2018.
48. Cordero, J.M.; Dorado, M.; de Pablo, J.M. Automated speech recognition in ATC environment. In Proceedings of the 2nd International Conference on Application and Theory of Automation in Command and Control Systems, London, UK, 29–31 May 2012; pp. 46–53.

49. Chen, S.; Kopald, H.D.; Elessawy, A.; Levonian, Z.; Tarakan, R.M. Speech inputs to surface safety logic systems. In Proceedings of the IEEE/AIAA 34th Digital Avionics Systems Conference (DASC), Prague, Czech Republic, 13–17 September 2015.
50. Chen, S.; Kopald, H.D.; Chong, R.; Wei, Y.; Levonian, Z. Read back error detection using automatic speech recognition. In Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 26–30 June 2017.
51. Updegrove, J.A.; Jafer, S. Optimization of Air Traffic Control Training at the Federal Aviation Administration Academy. *Aerospace* **2017**, *4*, 50. [CrossRef]
52. Helmke, H.; Ohneiser, O.; Buxbaum, J.; Kern, C. Increasing ATM Efficiency with Assistant Based Speech Recognition. In Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 26–30 June 2017.
53. Helmke, H.; Rataj, J.; Mühlhausen, T.; Ohneiser, O.; Ehr, H.; Kleinert, M.; Oualil, Y.; Schulder, M. Assistant-Based Speech Recognition for ATM Applications. In Proceedings of the 11th USA/Europe Air Traffic Management Research and Development Seminar (ATM2015), Lisbon, Portugal, 23–26 June 2015.
54. Traoré, M.; Hurter, C. Exploratory study with eye tracking devices to build interactive systems for air traffic controllers. In Proceedings of the International Conference on Human-Computer Interaction in Aerospace (HCI-Aero'16), Paris, France, 14–16 September 2016; ACM: New York, NY, USA, 2016.
55. Merchant, S.; Schnell, T. Applying Eye Tracking as an Alternative Approach for Activation of Controls and Functions in Aircraft. In Proceedings of the 19th Digital Avionics Systems Conference (DASC), Philadelphia, PA, USA, 7–13 October 2000.
56. Hurter, C.; Lesbordes, R.; Letondal, C.; Vinot, J.L.; Conversy, S. Strip'TIC: Exploring augmented paper strips for air traffic controllers. In Proceedings of the International Working Conference on Advanced Visual Interfaces, Capri Island, Italy, 22–26 May 2012; ACM: New York, NY, USA, 2012; pp. 225–232.
57. Alonso, R.; Causse, M.; Vachon, F.; Parise, R.; Dehaise, F.; Terrier, P. Evaluation of head-free eye tracking as an input device for air traffic control. *Ergonomics* **2013**, *2*, 246–255. [CrossRef]
58. Westerman, W.C. Hand Tracking, Finger Identification, and Chordic Manipulation on a Multi-Touch Surface. Ph.D. Thesis, University of Delaware, Newark, DE, USA, 1999. Available online: https://resenv.media.mit.edu/classarchive/MAS965/readings/Fingerwork.pdf (accessed on 14 January 2020).
59. Seelmann, P.-E. Evaluation of an eye tracking and multi-touch based operational concept for a future multimodal approach controller working position, original German title: Evaluierung eines Eyetracking und Multi-Touch basierten Bedienkonzeptes für einen zukünftigen multimodalen Anfluglotsenarbeitsplatz. Bachelor's Thesis, Technische Universität Braunschweig, Braunschweig, Germany, 2015.
60. Jauer, M.-L. Multimodal Controller Working Position, Integration of Automatic Speech Recognition and Multi-Touch Technology, original German title: Multimodaler Fluglotsenarbeitsplatz, Integration von automatischer Spracherkennung und Multi-Touch-Technologie. Bachelor's Thesis, Technische Universität Braunschweig, Braunschweig, Germany, 2014.
61. Prakash, A.; Swathi, R.; Kumar, S.; Ashwin, T.S.; Reddy, G.R.M. Kinect Based Real Time Gesture Recognition Tool for Air Marshallers and Traffic Policemen. In Proceedings of the 2016 IEEE 8th International Conference on Technology for Education (T4E), Mumbai, India, 2–4 December 2016; pp. 34–37.
62. Singh, M.; Mandal, M.; Basu, A. Visual gesture recognition for ground air traffic control using the Radon transform. In Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, AB, Canada, 2–6 August 2005; pp. 2586–2591.
63. Savery, C.; Hurter, C.; Lesbordes, R.; Cordeil, M.; Graham, T.C.N. When Paper Meets Multi-touch: A Study of Multi-modal Interactions in Air Traffic Control. In *Human-Computer Interaction—INTERACT 2013*; Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2013; Volume 8119, pp. 196–213.
64. Mertz, C.; Chatty, S.; Vinot, J.-L. Pushing the limits of ATC user interface design beyond S&M interaction: The DigiStrips experience. In Proceedings of the 3rd USA/Europe Air Traffic Management Research and Development Seminar (ATM2000), Naples, Italy, 3–6 June 2000.
65. EUROCONTROL. *E-OCVM Version 3.0 Volume I—European Operational Concept Validation Methodology*; EUROCONTROL: Brussels, Belgium, 2010.
66. NASA. Technology Readiness Level Definitions. n.d. Available online: https://www.nasa.gov/pdf/458490main_TRL_Definitions.pdf (accessed on 14 January 2020).

67. SESAR Joint Undertaking. Introduction to the SESAR 2020 Programme Execution. 2015. Available online: https://ec.europa.eu/research/participants/data/ref/h2020/other/guides_for_applicants/jtis/h2020-pr-exec-intro-er-sesar-ju_en.pdf (accessed on 14 January 2020).
68. Nielsen, J. *Usability Engineering*; Academic Press: Boston, MA, USA, 1993.
69. DIN EN ISO 9241-11:2016. *Ergonomics of Human-System-Interaction—Part 11: Usability: Definitions and Concepts*; ISO: Geneva, Switzerland, 2017.
70. Chen, Y.-H.; Germain, C.A.; Rorissa, A. An Analysis of Formally Published Usability and Web Usability Definitions. *Proc. Am. Soc. Inf. Sci. Technol.* **2009**, *46*, 1–18. [CrossRef]
71. Shackel, B. The concept of usability. In *Visual Display Terminals: Usability Issues and Health Concerns*; Ennet, J.L.B., Arver, D.C., Andelin, J.S., Smith, M., Eds.; Prentice-Hall: Englewood Cliffs, NJ, USA, 1984; pp. 45–88.
72. Shackel, B. Usability—Context, Framework, Definition, Design and Evaluation. In *Human Factors for Informatics Usability*; Shackel, B., Richardson, S., Eds.; Cambridge University Press: Cambridge, UK, 1991; pp. 21–38.
73. Maguire, M. Methods to support human-centred design. *Int. J. Hum.-Comput. Stud.* **2001**, *55*, 587–634. [CrossRef]
74. Weinschenk, S. Usability: A Business Case, Human Factors International. White Paper. 2005. Available online: https://humanfactors.com/downloads/whitepapers/business-case.pdf (accessed on 14 January 2020).
75. Seebode, J.; Schaffer, S.; Wechsung, I.; Metze, F. Influence of training on direct and indirect measures for the evaluation of multimodal systems. In Proceedings of the Tenth Annual Conference of the International Speech Communication Association (INTERSPEECH2009), Brighton, UK, 6–10 September 2009.
76. Nielsen, J.; Levy, J. Measuring usability: Preference vs. performance. *Commun. ACM* **1994**, *4*, 66–75. [CrossRef]
77. Xu, Y.; Mease, D. Evaluating web search using task completion time. In Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09), Boston, MA, USA, 19–23 July 2009; ACM: New York, NY, USA, 2009; pp. 676–677.
78. Wechsung, I. What Are Multimodal Systems? Why Do They Need Evaluation? Theoretical Background. In *An Evaluation Framework for Multimodal Interaction*; T-Labs Series in Telecommunication Services; Springer: Cham, Switzerland, 2014; pp. 7–22. [CrossRef]
79. Landauer, T.K. Research methods in human-computer interaction. In *Handbook of Human-Computer Interactio*; Elsevier: Amsterdam, The Netherlands, 1988; pp. 905–928.
80. Virzi, R.A. Refining the Test Phase of Usability Evaluation: How Many Subjects is Enough? *Hum. Factors* **1992**, *4*, 457–468. [CrossRef]
81. Nielsen, J. Why You Only Need to Test with 5 Users. 2000. Available online: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ (accessed on 14 January 2020).
82. Schmettow, M. Sample size in usability studies. *Commun. ACM* **2012**, *4*, 64–70. [CrossRef]
83. Ajzen, I.; Fishbein, M. *Understanding Attitudes and Predicting Social Behavior*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1980.
84. Davis, F.D. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q.* **1989**, *3*, 319–340. [CrossRef]
85. Davis, F.D.; Bagozzi, R.P.; Warshaw, P.R. User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Manag. Sci.* **1989**, *8*, 982–1003. [CrossRef]
86. Davis, F.D.; Venkatesh, V. A critical assessment of potential measurement biases in the technology acceptance model: Three experiments. *Int. J. Hum.-Comput. Stud.* **1996**, *45*, 19–45. [CrossRef]
87. Yousafzai, S.Y.; Foxall, G.R.; Pallister, J.G. Technology acceptance: A meta-analysis of the TAM: Part 1. *J. Model. Manag.* **2007**, *3*, 251–280. [CrossRef]
88. Kim, H.; Kankanhalli, A. Investigating User Resistance to Information Systems Implementation: A Status Quo Bias Perspective. *MIS Q.* **2009**, *3*, 567–582. [CrossRef]
89. Markus, M.L. Power, politics, and MIS implementation. *Commun. ACM* **1983**, *6*, 430–444. [CrossRef]
90. Likert, R.A. Technique for the Measurement of Attitudes. *Arch. Psychol.* **1932**, *140*, 5–55.
91. Davis, F.D. A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1985. Available online: https://dspace.mit.edu/bitstream/handle/1721.1/15192/14927137-MIT.pdf (accessed on 14 January 2020).

92. Doll, W.J.; Hendrickson, A.; Deng, X. Using Davis's perceived usefulness and ease-of-use instrument for decision making: A confirmatory and multi-group invariance analysis. *Decis. Sci.* **1998**, *4*, 839–869. [CrossRef]
93. Jackson, T.F. System User Acceptance Thru System User Participation. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*; American Medical Informatics Association: Bethesda, MD, USA, 1980; Volume 3, pp. 1715–1721.
94. Lin, W.T.; Shao, B.B.M. The relationship between user participation and system success: A simultaneous contingency approach. *Inf. Manag.* **2000**, *27*, 283–295. [CrossRef]
95. Luna, D.R.; Lede, D.A.R.; Otero, C.M.; Risk, M.R.; de Quirós, F.G.B. User-centered design improves the usability of drug-drug interaction alerts: Experimental comparison of interfaces. *J. Biomed. Inform.* **2017**, *66*, 204–213. [CrossRef]
96. Kujala, S. User involvement: A review of the benefit and challenges. *Behav. Inf. Technol.* **2003**, *1*, 1–16. [CrossRef]
97. König, C.; Hofmann, T.; Bruder, R. Application of the user-centred design process according ISO 9241-210 in air traffic control. *Work* **2012**, *41*, 167–174. [CrossRef] [PubMed]
98. DLR Institute of Flight Guidance. TriControl—Multimodal ATC Interaction. 2016. Available online: http://www.dlr.de/fl/Portaldata/14/Resources/dokumente/veroeffentlichungen/TriControl_web.pdf (accessed on 14 January 2020).
99. Ohneiser, O. *RadarVision—Manual for Controllers, Original German Title: RadarVision—Benutzerhandbuch für Lotsen*; Internal Report 112-2010/54; German Aerospace Center, Institute of Flight Guidance: Braunschweig, Germany, 2010.
100. Brooke, J. SUS: A "quick and dirty" usability scale. In *Usability Evaluation in Iindustry*; Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L., Eds.; Taylor & Francis: London, UK, 1996; pp. 189–194.
101. Brooke, J. SUS: A Retrospective. *J. Usability Stud.* **2013**, *2*, 29–40.
102. Sauro, J.; Leis, J.R. When Designing Usability Questionnaires, Does It Hurt to Be Positive? In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, 7–12 May 2011; pp. 2215–2224.
103. Bangor, A.; Kortum, P.T.; Miller, J.T. An Empirical Evaluation of the System Usability Scale. *Int. J. Hum.-Comput. Interact.* **2008**, *24*, 574–594. [CrossRef]
104. Bangor, A.; Kortum, P.T.; Miller, J.T. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *J. Usability Stud.* **2009**, *4*, 114–123.
105. Brinkman, W.-P.; Haakma, R.; Bouwhuis, D.G. Theoretical foundation and validity of a component-based usability questionnaire. *Behav. Inf. Technol.* **2009**, *28*, 121–137. [CrossRef]
106. Jakobi, J. Prague—A SMGCS Test Report. 2010. Available online: http://emma2.dlr.de/maindoc/2-D631_PRG-TR_V1.0.pdf (accessed on 14 January 2020).
107. Bishop, P.A.; Herron, R.L. Use and Misuse of the Likert Item Responses and Other Ordinal Measures. *Int. J. Exerc. Sci.* **2015**, *3*, 297–302.
108. Burnard, P.; Gill, P.; Stewart, K.; Treasure, E.; Chadwick, B. Analysing and presenting qualitative data. *Br. Dent. J.* **2008**, *8*, 429–432. [CrossRef] [PubMed]
109. Nørgaard, M.; Hornbæk, K. What do usability evaluators do in practice? An explorative study of think-aloud testing. In Proceedings of the 6th conference on Designing Interactive systems, University Park, PA, USA, 26–28 June 2006; ACM: New York, NY, USA, 2006; pp. 209–218.
110. Battleson, B.; Booth, A.; Weintrop, J. Usability Testing of an Academic Library Web Site: A Case Study. *J. Acad. Librariansh.* **2001**, *3*, 188–198. [CrossRef]