CEAS Aeronautical Journal
https://doi.org/10.1007/s13272-017-0275-y

ORIGINAL PAPER

A practical example for validation of ATM security prototypes

Michael Finke[1] · Tim H. Stelkens-Kobsch[1]

Received: 21 March 2017 / Revised: 20 November 2017 / Accepted: 27 November 2017

Abstract
The insights presented in this article are outcomes of a security research project that was initiated to collate and interpret the latest findings gathered in the domain of air traffic management security. The concept of a holistic approach to security management has been evaluated. Due to the large scope of the project, only an excerpt of the findings is provided in this article. This article focuses on a brief description of a security prototype validation methodology, developed within the project. To provide tangible application of the methodology, the adoption to a security prototype is developed, which is intended to enhance security of the air traffic control voice communication system.

Keywords ATM security · Validation · Security prototype · ATC radio communication

1 Introduction

One underestimated challenge to the existing air traffic management (ATM) is the existence of security threats imposed by intentional attacks on the infrastructure. Security measures to avoid exploitation of vulnerabilities, or to mitigate successful exploitation, have increased both in number and in their effectiveness over time [1]. Nevertheless, since the main impact on ATM on September 11, 2001, the awareness about new and serious threats has increased. This consequently results in the need for security solutions which propose (1) security controls to avoid the penetration of the ATM system, (2) measures to mitigate the influence of successful intrusion, and (3) countermeasures to keep the impact on the ATM system at acceptable levels.

Programs to toughen up ATM for the future such as SESAR, NextGen, or CARATS provide extensive guidance material for enhancing safety and security [2–4]. However, when it comes to implementation, the number of projects or initiatives is extremely small.

The Project GAMMA[1] marked the first implementation of SESAR (Single European Sky ATM Research) guidance material regarding security risk assessment and treatment. The work in the project followed the Security Risk Assessment (SecRAM) [2] and treatment postulated by SESAR. This was accompanied by the application of the minimum set of security controls (MSSC) [3]. As a summary, the following steps were taken in the development phase of GAMMA regarding all kinds of processes throughout the known ATM:

- Identify the primary and supporting assets.
- Identify vulnerabilities.
- Invent the attack scenarios.
- Name the security objectives.
- Analyze and treat the risks.
- Define security controls.

This article discusses a subset of the project outcomes. It briefly describes the postulated security prototype validation methodology and, more comprehensively, its practical application. This may be used as the blueprint and an example for validating other ATM security-oriented systems.

A dedicated prototype for secure air traffic control (ATC) voice communication shall serve as a practical example for applying the described validation methodology. This prototype will close the known security

✉ Michael Finke
michael.finke@dlr.de

[1] Deutsches Zentrum fur Luft- und Raumfahrt e.V., Braunschweig, Germany

[1] GAMMA, http://www.gamma-project.eu. The research leading to the results presented in this article has received funding from the European Union's Seventh Framework Programme under Grant Agreement no. 312382.

Published online: 24 January 2018

vulnerability of open, analogue, and unsecured air–ground voice communication which is still widely used in ATC.

Strategies to counteract this security threat have already been investigated, such as the promising approach of adding an almost unnoticeable digital watermark signature to every audio transmission, which can be decoded and used to identify the station from which the corresponding transmission was sent. This solution is primarily intended to submit aircraft identification tags (AIT) to ATC, but it can be used in the same way to confirm the identification of the responsible ATC unit to pilots. Several studies have shown the technical feasibility [5–7]. Nevertheless, this (intrusive) approach would require a change of technical radio equipment.

The prototype described in this article uses a new, non-intrusive approach to secure the air–ground voice communication. More information regarding the concept of this prototype and the conducted validation is given below.

2 Validation approach for security prototypes in ATM

Another important goal of the GAMMA project is the validation of the newly developed security controls (enhanced to security prototypes). The security controls represent the core element of the postulated holistic security concept [8].

The validations in this project were set up according to the protocols of the European Operational Concept Validation Methodology (E-OCVM) [9]. E-OCVM attempts to make the expected benefit measureable. After verification of the system, the next task is to find out if "the right system was built" (the driving question of validation). The overarching question is: are the newly introduced systems and functions (respectively, processes) worth implementing?

To shorten up the work done in the project, the resulting approach to successfully validate ATM security prototypes consists of the following steps:

- Define security key performance indicators (KPI) to quantify the efficiency of the security controls.
- Set up the validation plan including specific validation goals, metrics, and acceptance criteria.
- Conduct the validation exercises.
- Evaluate the results.

The above shall now be elaborated in the following chapters taking one prototype as a tangible example.

3 An example: the SACom prototype

The secure ATC communications (SACom) prototype is a modular system which was developed as one of seven prototypes within the project, which were designed to close known security gaps in present or near future ATM systems [10]. Air–ground voice communication was recently described again to be one of the most endangered communication means for spoofing and/or spamming [11]. In particular, the SACom aims at improving the security of air–ground voice communication between pilots and air traffic control. For decades, this communication has been performed with analogue radio transceivers, which neither use encryption techniques nor are in any other way secure [12]. Physically, the air–ground communication is freely accessible and can easily be eavesdropped. Apart from the availability of appropriate radio communication equipment, there is no technical hurdle which has to be overcome for an unauthorized person to take part and disturb the ATC–pilot communication. The security threat which is now in the focus is the insertion of "fake ATC clearances" by an unauthorized speaker with the goal to seriously spoil the safe, orderly and economic flow of air traffic. Recent examples for such security events are published in [13, 14].

The SACom prototype uses voice analysis algorithms to verify the speaker identity of every utterance like proposed in [15]. The prototype furthermore adds several other indicators of a different nature. This combination of several indicators is done to not just identify a possible unauthorized transmission directly, but also to detect likely consequences of a successful inserted "fake ATC clearance". This approach is expected to make the system more robust, for example, in case of (1) failing to determine a single indicator, (2) false results of a single indicator, or (3) a single indicator being bypassed by the attacker, e.g., by voice reproduction software. In addition to that, it makes the system more flexible: SACom indicators can also provide benefits in other security/safety incidents, for example, hijacking or emergencies.

The detailed concept is described in the following section.

3.1 Concept of the SACom prototype

The SACom prototype is designed as a pure detection system. It continuously monitors the audio stream of the air–ground voice communication as well as the position and behaviour of the air traffic. The prototype directly detects (1) unauthorized transmissions (e.g., "fake ATC clearances") by means of voice analysis, but it also detects possible consequences of such transmissions, (2) increased stress level of all participants exposed to the incident, (3)

aircraft deviating from the cleared flight path while reacting to unauthorized transmissions, or (4) valid ATC clearances with the potential of a collision risk.

The basic idea of the SACom prototype is not just to provide, but also to combine and correlate all these indicators. Apart from increased robustness and flexibility, this shall enable the detection of either attempted or successfully infiltrated "fake ATC clearances". It shall also help to distinguish them from events with a pure safety background (unintentional mistakes in air–ground voice communication such as callsign mix-up, navigation errors, etc.). It must be considered that circumstances without any safety or security threat can also trigger individual SACom indicators (e.g., a student pilot under high stress). In this context, it is assumed that such events often trigger only one of the SACom indicators (1)–(4), while security events likely trigger more than one of these indicators at once and/or repeatedly.

It was one of the design requirements for the SACom prototype that it shall not modify or manipulate the existing air–ground voice communication (e.g., by means of encryption, by adding audio signatures/watermarks, by blocking content, etc.). Furthermore, the SACom prototype has to be able to automatically report security-relevant information to a defined security management entity (e.g., a Security Operation Center) [8].

As SACom is a pure detection system, the initiation of countermeasures still rests with the person using this system [e.g., the air traffic controller (ATCo)]. Therefore, the SACom primarily increases the situational awareness and enables the user to react more quickly and more appropriately to such security-relevant events as time plays a key role in handling them.

The SACom:

- Enables the recipient of a message to directly identify "fake ATC clearances" (by verifying the speaker identity).
- Can provide a rough indication concerning the current workload or confusion (by providing stress detection functions).
- Acts as a safety net (by providing conflict detection functions).
- Enables all users to immediately initiate corrective measures concerning the affected traffic (e.g., correcting aircraft deviating from their clearance).
- Enables all users to initiate preventive measures concerning traffic which is not (yet) affected (e.g., by sending aircraft to a backup frequency, increasing the separation between aircraft or reducing the traffic load on the ATC sector).
- Significantly accelerates the passing of security-relevant information to entities responsible for security management (security operation centers, authorities, etc.) by automatic reports.

According to the desired functions, the SACom prototype has a modular architecture and consists of the following sub-systems:

- Speaker verification module containing the voice authorization function (SV).
- Stress detection module containing the stress detection function (SD).
- Conformance monitoring module containing conformance monitoring (CM) and conflict detection functions (CD).
- Security management interface containing the correlation function and automatic reporting functions.

Speaker verification and stress detection modules require the input of the audio stream from the air–ground voice communication as well as a database of known and authorized speakers. These modules deliver a score value to the security management interface. A low speaker verification score indicates a mismatch of the voice of an utterance with the known voice characteristics of authorized speakers. A high stress score indicates the presence of known voice patterns that are typical for stressed speech in this utterance based on known parameters. The conformance monitoring module requires the input of surveillance data as well as complete, correct and up-to-date data about given ATC clearances.

The conformance monitoring function of this module works on one hand in a trajectory-based way for monitoring aircraft along complete flight paths such as approach procedures or arrival routes. On the other hand, it is also done in a non-trajectory-based format for explicitly instructed levels, headings, speeds, and rates of climb/descent.

The conflict detection function of this module in the same way consists of two parts. Received ATC clearances are on one hand converted into a predicted trajectory. On the other hand, the current aircraft state vector is simply extended to predict the future positions of aircraft in case of no further aircraft manoeuvers. Both predictions are continuously cross-checked for all known flights to detect possible conflicts in the future (in an adjustable time period). These detected conflicts may be caused by the given (maybe false) ATC clearances or by the current aircraft movements if they are maintained. The conflict detection function, therefore, can well be compared to the very common short-term conflict alert function (STCA) of state-of-the-art ATC radar systems, but also covers the medium-term conflict detection (MTCD) which today is mainly used in the upper airspace [16, 17].

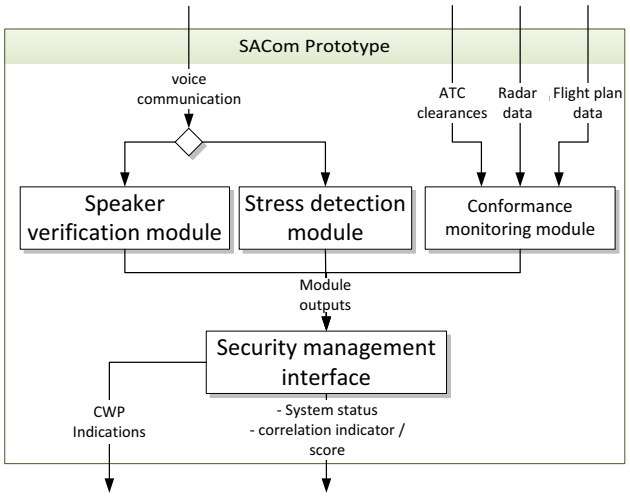

Fig. 1 Architecture of the SACom prototype

In case of detected deviations from the given ATC clearance or predicted collision risks between two aircraft, the conformance monitoring module will give appropriate alerts.

The security management interface requires the input of at least one of the above-mentioned modules and produces a correlated score value. According to the concept described above, this correlated score combines the different individual SACom indications and shall provide a specific value representing the overall security threat situation. The correlated score is obtained by adding up all individual indicators in a defined time period, applying weighting factors and again adding them up to an overall score.

A low-correlated score indicates normal operations, while a rapidly increasing score indicates suspicion of the presence of an unauthorized speaker inserting false instructions. This score is used by the prototype as a trigger for automatic alerting and reporting functions together with a defined alert threshold. The raw data and/or the correlated score can automatically be submitted to a security management entity or directly provided to user displays (air traffic controller tools or cockpit systems of a pilot).

Figure 1 sketches the architecture of the SACom prototype including different modules as well as inputs and outputs.

The speaker verification module and the stress detection module were developed and built by the Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia. The conformance monitoring module and the security management interface were developed and built by the Institute of Flight Guidance, German Aerospace Center (DLR), Braunschweig, Germany.

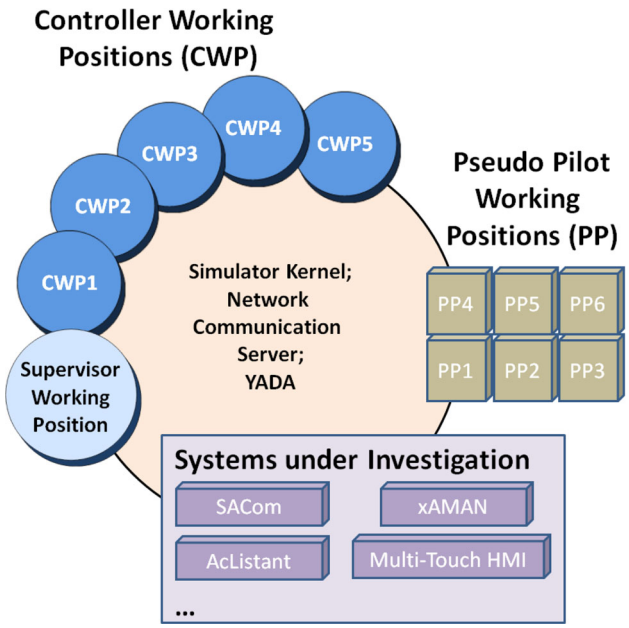

Fig. 2 Simulation facility ATMOS

3.2 Simulator implementation

The air traffic management and operations simulator (ATMOS) of DLR's Institute of Flight Guidance located in Braunschweig is an experimental facility for simulating air traffic in real time. In a simulated airspace, the ATMOS can be used to test, e.g., new procedures, ATM concepts, or supporting systems in terms of safety, feasibility, efficiency, and traffic capacity. This facility allows researchers and air traffic controllers to jointly evaluate new working methods for controlling and influencing air traffic. One area of research is determining the strain on controllers, including critical situations.

The simulator is primarily designed for performing assessments with interacting participants, which is why the traffic situations have to be simulated in real time. It is generally possible to use any airspace in the world, including one or more airports as desired. If necessary, the selected airport can be adjusted in line with different air traffic control sectors.

As shown in Fig. 2, the facility consists of five controller working positions (CWP), a supervisor working position and six pseudo-pilot working positions (PP). The air traffic generator used to establish simulated traffic is the NARSIM (NLR's Air Traffic Control Research Simulator). The system is completed with a flexible software solution (YADA) for voice over IP (VoIP) communication between controllers and pseudo-pilots.

This simulation facility was dedicated as the validation platform to conduct the security validation of the SACom prototype.

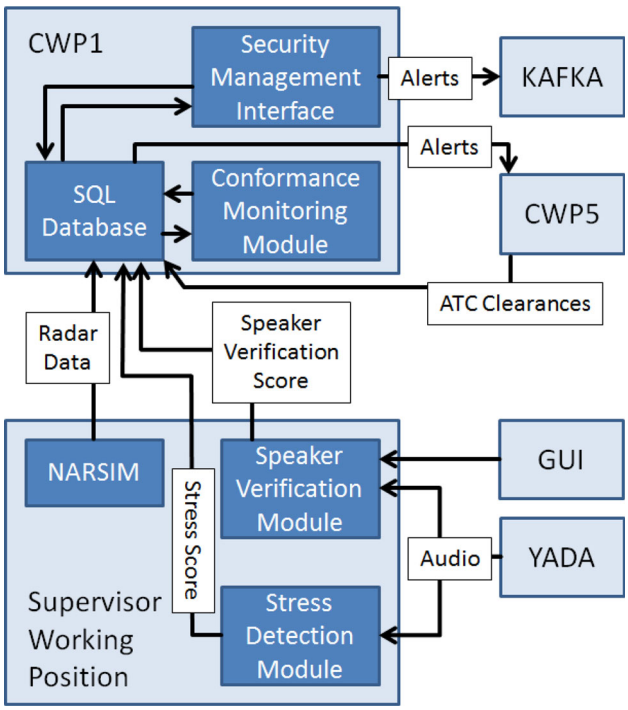

Fig. 3 SACom prototype implementation in ATMOS

The speaker verification module and the stress detection module were implemented as a software application on the supervisor working position. This software is granted a direct access to the voice communication audio stream, while the built-in database of authorized speakers could be changed or enhanced using a graphical user interface (GUI).

A separate controller working position (CWP1) housed the conformance monitoring module and the security management interface. In addition, an SQL database was used as data exchange interface between all prototype modules as well as the simulator components (NARSIM). This database was also managed from this station. The security management interface was equipped with a connection to an external security management entity. Automatic alerts and reports to this entity are passed through a separate module based on Apache KAFKA due to the local network architecture used in the simulator environment. As an example, this security management entity could be the security management platform, which is another security prototype developed by Leonardo, Italy.

The workstation CWP5 was intended to be used by the test person acting as radar approach controller during the validation trials. The radar display software was also able to access the SQL database on CWP1, which contained all alerts related to the SACom prototype.

Figure 3 illustrates the SACom implementation in the ATMOS which was just described above.

4 Validation methodology and validation objectives

The principle fitness for purpose of a technical mean is shown when it meets postulated requirements. Consequently, a prototype is fit for purpose—and, therefore, fulfils the research question—if it satisfies the requirements to address the threat. How well the prototype satisfies these requirements may be difficult to assess; therefore, the next task is to split these requirements into more measurable sub-requirements, which are reflected in security key performance indicators. Examples for this approach as well as possible general security key performance indicators (e.g., number of supporting assets affected, recorded time until mitigation of the attack, etc.) are available in [18].

Furthermore, the system configurations for baseline and conceptual solution have to be determined.

In case of the SACom prototype, the attack scenario consists of an intruder into the ATC voice communication who interacts with aircraft and issues deliberately safety critical instructions to pilots ("False Air Traffic Controller"—"False ATCo"). The overall goal of the exercise is to validate whether or not the SACom prototype is fit for purpose to address this threat.

The validation objectives which were addressed in the validation exercise are the following:

- To validate that the SACom prototype as a composition of different indicators supports the detection of a "False ATCo" attack or its consequences.
- To validate that the SACom prototype is usable, useful and leads to a better situational awareness of the user when reacting to a "False ATCo" attack or its consequences.
- To validate that the performance of the SACom prototype is acceptable for a system in this state of development.

In general, these validation objectives are considered as fulfilled when:

- determined System Usability Scale (SUS) scores are above 50;
- scores for user trust in the system (SATI) are above 2 (= "sometimes");
- detection rates are considerably higher than false alarm rates;
- times until detection are equal or better than human performance.

The objectives were the basis to elaborate several KPI which are specific to the prototype and its functions. These KPIs are the following:

- *Speaker verification: false alarm rate* Number of authorized transmissions which were incorrectly classified as unauthorized transmissions, divided by the total number of authorized transmissions.
- *Speaker verification: detection rate* Number of unauthorized transmissions which were correctly classified as unauthorized transmissions, divided by the total number of unauthorized transmissions.
- *Stress detection: false alarm rate* Number of utterances which are believed to be free of stress but nevertheless show a stress score above a defined alert threshold, divided by the total number of utterances which are believed to be free of stress.
- *Stress detection: detection rate* Number of utterances which are believed to be compromised with stress and which show a stress score above a defined alert threshold, divided by the total number of utterances which are believed to be compromised with stress.
- *Conformance monitoring: false alarm rate* Number of incorrect conformance monitoring alerts, divided by the total number of all conformance monitoring alerts.
- *Conformance monitoring: detection rate* Number of correctly detected deviations from a valid ATC clearance, divided by the total number of deviations from a valid ATC clearance. This KPI was determined for the performance of the SACom prototype but also for the performance of the ATCo.
- *Conformance monitoring: time until detection* Time difference between the timestamp when the deviation was detected and the timestamp when the deviation could have been detected at the earliest. This KPI was determined for the performance of the SACom prototype but also for the performance of the ATCo.
- *Conflict detection: false alarm rate* Number of incorrect conflict alerts, divided by the total number of all conflict alerts.
- *Conflict detection: detection rate* Number of correctly detected conflict situations, divided by the total number of all conflict situations. This KPI was determined for the performance of the SACom prototype but also for the performance of the ATCo.
- *Correlation: false alarm rate* Number of seconds of a simulation where the correlation score went beyond a defined alert threshold without the presence of a "False ATCo" attack, divided by the total number of seconds of a simulation without the presence of a "False ATCo" attack.
- *Correlation: detection rate* Number of seconds of a simulation where the correlated score went beyond a defined alert threshold during the presence of a "False ATCo" attack, divided by the total number of seconds of a simulation with a presence of a "False ATCo" attack.
- *Correlation: time until detection* Time difference between the timestamp of the first "False ATCo" action and the time stamp where the correlated score reached a defined alert threshold.
- *User acceptance* This KPI is reflected in the SUS score (system usability) as well as the SATI score (situational awareness), obtained from the questionnaires.

The application of the validation methodology led to the validation procedure described hereafter and was a straightforward approach to provide evidence that the SACom prototype is fit for purpose, to fulfill the postulated requirements, to achieve the validation goals and to meet pre-defined acceptance criteria corresponding to security key performance indicators.

5 Validation procedure

As the SACom system could be used in aircraft cockpits and at controller working positions, it should be validated for both use cases.

The procedure described in the following always refers to the validation of the SACom installed on the ATC side. The validation of the SACom used in aircraft cockpits is not considered further in this article.

It was decided to perform the validation of the SACom prototype as human-in-the-loop real-time simulation. Active air traffic controllers from the German Air Navigation Service Provider DFS and from the Australian Air Navigation Service Provider Air Services Australia were recruited as test persons to take part in the validation campaign.

5.1 Steps of a validation run

To reflect the validation methodology described above, the following validation procedure was chosen:

1. Briefing of the test person

The test person taking part in the simulations is comprehensively briefed about the projects' concept, the simulation trials, about the working environment and about the goal of the validation campaign.

2. Adding the test person to the database of authorized speakers (voice enrolment)

The test person is asked to read a number of prepared ATC clearances. This audio stream is recorded and used to create a new voice enrolment for the speaker verification module. Afterwards, the test person reads again the same prepared ATC clearances to verify if the reliability of the new enrolment is sufficient.

3. Simulator training

The simulator training is performed as a short-time simulation with low workload to give the test person the opportunity to make himself or herself familiar with the airspace structure, the local operating procedures, and the used CWP equipment. No SACom indications are displayed to the controller in this simulation. Nevertheless, this step is already used to test the stress detection function for evaluation of false alarms. Due to the low workload, it is expected that no stress is present in this simulation.

4. Short-time simulations

The next step of one validation run is a set of 20 short-time simulations, split up into four blocks consisting of five short-time scenarios each. These short-time simulations contain pre-defined safety or security events which cause aircraft to deviate from given clearances or even introduce a collision risk between two aircraft. The task of the test person is to react to these events while providing ATC service to all aircraft under control. The SACom prototype will work in the background, but the corresponding indications will not be displayed to the controller. More information about this validation step is given in Chapter 5.3.

5. SACom briefing

In this validation step, the SACom prototype is introduced to the test person. He or she is briefed about the functions, the purpose, and corresponding indications on the controller's Human–Machine Interface (HMI).

6. SACom training

This is a short-time simulation run with low workload. It is used to give the test person the opportunity to become acquainted with the SACom indications which were just briefed.

7. Long simulation run containing a complete security incident

This validation step consists of a 45 min simulation run, which contains a phase of normal operations (15 min), a phase of multiple intrusions of an unauthorized person into the air–ground voice communication (20 min) and again a phase of normal operations. The test person has the task to react to these events and to compensate the effects as much as possible by making use of the SACom functions. More information about this validation step is given in Sect. 5.4.

8. Questionnaires

In the next step, the test person is then asked to give answers to prepared online questionnaires. These are standardized questionnaires aiming at situational awareness (Situational Awareness for SHAPE (SASHA) Questionnaire of EUROCONTROL [19]), usability (System Usability Scale (SUS) [20]), and trust (SHAPE Automation Trust Index (SATI) [21]). In addition, tailor-made questionnaires are filled which focus on very specific aspects of interest. Such aspects may be, e.g., realism of the validation exercise, the GAMMA concept as whole or potential improvements.

9. Debriefing

Each validation run ends with a de-briefing session. Here, the controller can give any other feedback and the validation exercise is discussed.

5.2 CWP equipment and working environment

For all simulations, the same equipment was installed at the CWP:

- A radar display respecting EUROCONTROL's recommendations for air traffic control radar systems as far as feasible and reasonable within research [22].
- The open source voice over IP radio communication simulator YADA [23].
- A separate speech recognition system which delivers necessary data about given ATC clearances [24] with the possibility to make manual inputs.

The simulated airspace was the terminal control area of Düsseldorf airport, Germany. The test person acted as radar approach controller guiding the aircraft from the initial contact to the final approach segment. The simulated traffic consisted of flights approaching under Instrument Flight Rules (IFR). Runway 23R of Düsseldorf airport was used for all landings. No departures or flights under visual flight rules were simulated.

Minimum vectoring altitude was defined to be 3000 ft MSL in general for simplification, because all invited controllers did not hold local ratings for the Düsseldorf approach sector.

All flights had to be separated using the standard radar separation or standard vertical separation (3NM lateral separation and 1000 ft vertical separation).

5.3 Short-time simulations

The reason for choosing several short-time scenarios instead of one big scenario is the need for comparability and the need to produce a sufficient amount of data.

It often depends on exact traffic constellations and an exact timing whether specific events can be simulated equally for all exercise runs. Having said this, a long simulation run is far too dynamic and offers too much flexibility for the evolution of the traffic situation. From the

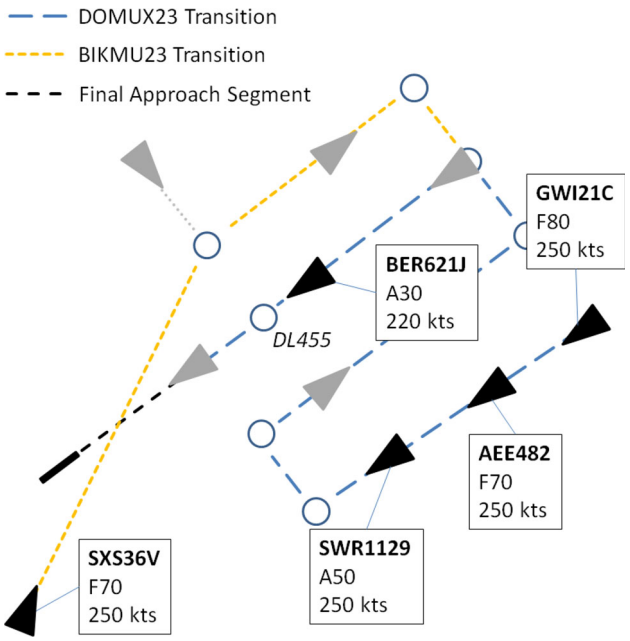

Fig. 4 Short-time scenario number 16

simulation design point of view, it is nearly impossible to inject exactly the same events with exactly the same consequences in different validation runs when using long simulation scenarios. Practically, only the first few minutes of a simulation can be considered as comparable by setting specific initial aircraft positions, state vectors, and traffic constellations. Nevertheless, due to the differences in performance, work habits, and experience from controller to controller, it is absolutely natural and cannot be completely avoided that the same traffic situation may be solved in a different way every time.

When using a set of short-time scenarios instead, it is possible to set up pre-defined traffic constellations and state vectors recurrently, confronting all test persons with exactly the same preconditions.

For the SACom validation, 20 short-time simulations were performed. Each simulation took between 3 and 6 min and contained a pre-defined traffic situation as well as pre-defined events with a safety or security background. All those events are categorized and were distributed over the short-time simulations according to their nature, which can be:

- Lateral, vertical, or speed deviations.
- Manoeuvers without any clearance.
- Wrong compliance with a new clearance, e.g., a left turn instead of a right turn.
- Non-compliance to a new clearance.
- Deviations caused by safety issues (e.g., technical failures).
- Deviations caused by security issues (e.g., simulated fake ATC clearances).

All short-time scenarios need to be completed by the test person without SACom support, which allows the direct comparison of the performance of the unsupported air traffic controller (= "baseline") with the performance of the SACom prototype (= "best case conceptual solution").

Figure 4 gives an example of short-time scenario number 16 as it was simulated during the validation trials.

In this short-time scenario (like in all other short-time scenarios), several aircrafts are placed along the published approach procedures of Düsseldorf Airport (BIKMU23 and DOMUX23 transitions) in a defined altitude, at a defined heading, with a defined speed and with defined initial ATC clearances. The task of the test person is to guide the aircraft until the final approach segment.

The pre-defined event in this example is a simulated fake ATC clearance given by an unauthorized person to BER621J. This fake clearance is issued at waypoint DL455 and instructs this flight to discontinue the approach and to climb straight ahead to flight level 70. The intention of this fake clearance is to cause a collision hazard with SXS36 V, which follows the BIKMU 23 transition in flight level 70, overflying the runway centerline.

To make sure that this pre-defined event can successfully be simulated, it is important to design every situation in a way which leaves only very few options open to guide the traffic. With regard to the example presented above, this is on one hand achieved by placing BER621J on the final track. It is very unlikely that a controller turns this aircraft away, because it is short prior reaching the final approach segment. On the other hand, to avoid that the controller changes the flight path of SXS36V, the DOMUX 23 transition is occupied by other aircraft (SWR1129, AEE482, and GWI21C), leaving no gap open where SXS36V could be fed in directly.

5.4 Security incident simulation

Within this simulation, a complete intrusion of an unauthorized person infiltrating fake ATC clearances into air–ground voice is simulated.

This simulation was designed to last 45 min, while a medium workload is imposed.

During the first 15 min, no special event will be simulated (normal operations). The main purpose of this phase is to observe the SACom prototype in a situation without any threat and to investigate if and why false alarms occur and how they affect the correlation function.

In the second phase of the simulation, an unauthorized speaker will insert fake ATC clearances to cause a collision risk between two aircraft or at least confusion and an

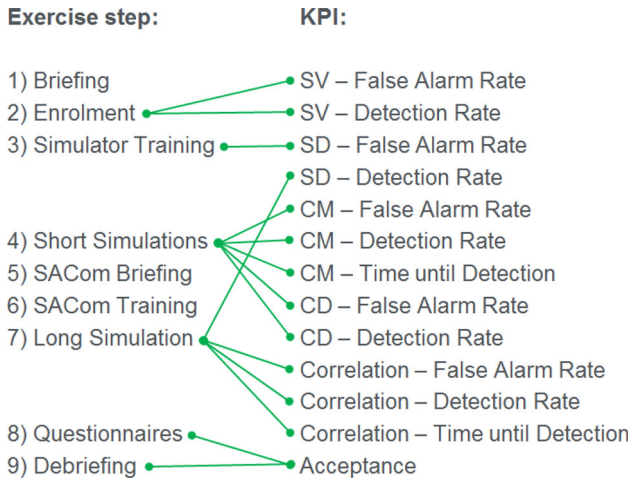

Fig. 5 Connection between exercise steps and KPIs

increased workload, which may lead to a loss of safety and capacity. The second phase was planned with a length of 20 min, whereas the unauthorized speaker tries to inject fake clearances at intervals of at least 90 s and after 300 s at the latest (rule of thumb). As potentially dangerous unlawful actions cannot be predetermined and planned before, the person acting as unauthorized speaker needs to have knowledge in air traffic control procedures and has to decide spontaneously according to the traffic situation just as a real attacker would do. To simulate block-out effects in voice communication and to increase the probability that the interference is not immediately detected, the pseudo-pilots are instructed not to read back all of the given fake clearances.

The main purpose of this second phase is to observe the correlation process of the SACom prototype while being exposed to a security threat, but also to learn more about the typical reactions and countermeasures of an air traffic controller experiencing such an attack.

The third phase of this simulation again contains normal operations without any interference, because the SACom prototype should also cease to produce warnings after the incident is over.

6 Results' overview

During late summer and fall of 2016, several validation exercises were performed, which are listed in Table 1.

6.1 Assessment of validation results

Figure 5 shows the relations between the KPIs (right) and the exercise steps (left), which create the data needed to determine the individual KPIs (SV = Speaker Verification, SD = Stress Detection, CM = Conformance Monitoring, and CD = Conflict Detection).

Some of the steps will be described in more detail below to give a better understanding of the result assessment.

6.1.1 Enrolment

During this exercise step, the test person is added to the database of authorized speakers. Participants were asked to read a defined number of phrases one after the other. All utterances are recorded and fed into the speaker verification module to extract the voice characteristics.

6.1.2 Simulator training

All participants conducted a training session within this facility to familiarize with the setup, the system, and the software. During the simulator training, it is expected that no stress is present due to the very low workload and the low demands. In this simulation, the controllers were instructed to "play" with all the CWP functions and to make themselves familiar with them.

Nevertheless, as the stress detection module was working in the background, a corresponding stress value was determined for every utterance. It is expected that the stress detection module does not measure any stress and, therefore, produces stress scores of or near to zero.

6.1.3 Short-time simulations

During this exercise step, the following data were recorded (beside other):

- Complete radar data for all aircraft.
- Given ATC clearances.
- All SACom alerts.

In addition, a psychologist observed the test person to exactly log his or her reaction to pre-defined events.

As expected, some of the pre-defined events could not successfully be simulated during the short-time scenarios due to preventive actions taken by the air traffic controller. In contrast, some unplanned aircraft deviations and even conflict situations occurred. The majority of these unplanned events were caused by pseudo-pilot errors, but some were also caused by hasty controller actions without taking the new situation into full account.

Table 2 shows in detail that a high percentage of pre-planned events could be simulated successfully throughout all short-time scenarios, resulting in a very good comparability between the test persons.

Table 1 Validation exercise overview

Exercise ID	Date	Test person
VL–C1	24th/25th Aug	ATCo 1 from DFS
GP–C2	21st/22nd Sept	ATCo from Air Services Australia
AL1–C3	17th/18th Oct	ATCo 2 from DFS
AL2–C4	18th/19th Oct	ATCo 3 from DFS
AL3–C5	19th/20th Oct	ATCo 4 from DFS
AL4–C6	20th/21st Oct	ATCo 5 from DFS

6.1.4 Long simulation

Within this scenario, a briefed person acting as a "False ATCo" deliberately gives instructions to pseudo-pilots. During this exercise step, the same data were recorded as in the short-time simulations. In addition, the person acting as "False ATCo" logged every inserted fake ATC clearance and observed the effects on the traffic flow. This person also assessed if the intervention caused or contributed to a safety–critical situation. Table 3 gives an overview of given fake clearances and their effects. About 50% of all fake clearances had no effect on the traffic flow, mainly because the fake clearance (interfering with other utterances) just led to a short block-out on the frequency. This simply compelled the controller to repeat his instruction. In some cases, the controller gave an alternate clearance before the pilots were able to comply with the fake clearance. Most fake clearances were recognized and corrected by the controllers, which just caused additional workload and delay. However, interestingly, in two cases, a safety–critical situation could be provoked, in one case even with causing a serious collision risk (closest proximity 1,02NM/200 ft on nearly opposite headings).

6.1.5 Questionnaires and debriefing

Prepared questionnaires were filled out electronically and standardized scales were used [System Usability Scale (SUS) and SHAPE Automation Trust Index (SATI)]. This enabled the use of automatic analysis tools, establishing that strong comparability is achieved between trials.

Table 3 Inserted fake clearances and their effects

Exercise ID	Inserted fake clearances	With an effect on the traffic flow (delay)	With a safety–critical consequence
VL–C1	5	3	0
GP–C2	6	4	0
AL1–C3	9	3	0
AL2–C4	10	5	1
AL3–C5	6	6	0
AL4–C6	13	6	1

The usability scale (SUS) has a range from 0 to 100, values above 50 being considered as acceptable, and values above 70 indicating good usability [20]. The trust scale (SATI) ranges from 0 = "never" to 6 = "always" [21].

The de-briefing session is a last opportunity to collect subjective impressions, opinions, and feedback which may be the basis for further developments, improvements, and for finding new fields of research.

6.2 Quantitative results, controller feedback, and discussion

The following sections will state the results and conclusions specific to the different SACom functions.

6.2.1 Speaker verification function

Table 4 shows the results obtained for the speaker verification function in a very brief way.

It can be seen that the pure performance of the speaker verification function under laboratory conditions (good audio quality, no background noise, etc.) was very satisfying with a low false alarm rate and a high detection rate.

Nevertheless, the system got usability ratings between 43 (= "poor") and 75 (= "good"), and the trust index got a mean rating of 2.2 (= "sometimes"). During the debriefing session, the controllers stated that the reason for this is the used HMI, which is an additional window on the screen showing almost always non-relevant information, because most of the transmissions were authorized. It is expected

Table 2 Successfully simulated and unplanned events during short-time simulations

Exercise ID	Successful events	Successful events (%)	Unplanned events
VL–C1	17/20	85	8
GP–C2	18/20	90	7
AL1–C3	18/20	90	4
AL2–C4	16/20	80	9
AL3–C5	17/20	85	4
AL4–C6	17/20	85	1

Table 4 Speaker-verification KPIs, usability, and trust

Exercise ID	Opt. alert thr.	False alarm rate (%)	Detection rate	SUS	SATI
VL–C1	15	3.3	100	NA	NA
GP–C2	14–30	0	100	50	2.33
AL1–C3	27–35	0	100	43	1.67
AL2–C4	21	0	96	58	1.5
AL3–C5	31–40	0	100	75	3.33
AL4–C6	33	0	91.7	50	2.17
Average	26.25	0.55	97.95	55	2.2

that, with an HMI redesign, the usability and trust rating can be improved significantly.

6.2.2 Stress detection function

Table 5 shows the results obtained for the stress detection function.

The alert threshold was set to a stress score of 30 during the exercises as this was estimated to be a suitable value because of the results of preliminary trials.

Apart from exercise run AL4–C6, the stress detection function unfortunately did not indicate an increased stress score during the attack phase compared to the phase with normal operations.

This can be interpreted in the following ways:

- The induction of stress during the simulation was not successful.
- The classification of utterances as "compromised with stress" or as "free of stress" is not accurate.
- The stress detection function is not sensitive enough.
- The reliability of the stress detection function depends on the voice characteristics and is not equal for all speakers.

Separate trials with the stress detection function, which were conducted in January 2017 in Braunschweig (results will be published in autumn 2017), showed that the stress detection in ATC voice communication under laboratory conditions may be successful for some speakers, but without the reliability which would be expected from a security system. This stems from the fact that stress detection from voice is still science in its infancy. As this trend was already visible during pre-validation tests, the stress score was not displayed to the controller during the validation exercises. Consequently, no SATI or SUS was filled out for the stress detection function.

6.2.3 Conformance monitoring function

Table 6 shows results obtained for the conformance monitoring function.

Table 5 Stress detection KPIs

Exercise ID	False alarm rate (%)	Detection rate (%)
VL–C1	23.0	27.6
GP–C2	1.3	0.7
AL1–C3	3.1	0.6
AL2–C4	21.8	14.6
AL3–C5	20.3	23.3
AL4–C6	6.3	20.0
Average	12.6	14.5

It was recognized that most of the false alarms occurred because of incorrect, incomplete, early or late information about the given ATC clearances, which initially led to a very high false alarm rate. For the result analysis, these input errors were eliminated subsequently, resulting in a second value for the false alarm rate representing the amount of false alarms which really come from the system. Consequently, two types of false alarm rates were defined for the result analysis:

- False alarm rate type I, which is calculated from the amount of all false alarms including those that are caused by incorrect inputs.
- False alarm rate type II, which is calculated from the amount of false alarms excluding those that are caused by incorrect inputs.

The results show that the performance of the conformance monitoring function was satisfying with an average detection rate of 91.2% and an average false alarm rate type II of 8.8%. In most exercise runs, the detection rate of the prototype was higher than the detection rate of the air traffic controller without support.

Another big advantage of the conformance monitoring function is that it can detect aircraft deviations significantly faster than the air traffic controller (25 s faster on the average).

The system got usability ratings between 35 (= "poor") and 75 (= "good"), and the trust was rated with 2.4 on average (= "sometimes"). During the debriefing session, the controllers stated that the reason for this is the high

Table 6 Conformance monitoring KPIs, usability, and trust

Exercise ID	False alarm rate type I (%)	False alarm rate type II (%)	Detection rate SACom (%)	Detection rate ATCo (%)	Time until detection SACom	Time until detection ATCo	SUS	SATI
VL–C1	63.8	10.3	88.0	92.0	16.5	41.6	NA	NA
GP–C2	71.0	7.2	93.3	76.7	11.8	39.4	67.5	4
AL1–C3	34.3	2.9	95.8	91.7	15.8	43.1	75	3.83
AL2–C4	56.7	9.0	96.7	80.0	14.5	38.7	35	1
AL3–C5	58.9	12.5	88.5	84.6	13.9	38.9	40	1.33
AL4–C6	52.8	11.1	85.0	85.0	14.1	34.7	60	1.83
Average	56.3	8.8	91.2	85.0	14.4	39.4	55.5	2.4

Table 7 Conflict detection KPIs, usability, and trust

Ex. ID	False alarm rate type I (%)	False alarm Rate type II (%)	Detect. rate SACom (%)	Detect. rate ATCo (%)	SUS	SATI
VL–C1	66.7	0.0	80.0	100.0	NA	NA
GP–C2	84.5	7.1	88.2	76.4	72.5	4.33
AL1–C3	87.7	16.7	72.7	90.9	NA	NA
AL2–C4	87.9	12.5	88.9	88.9	67.5	6
AL3–C5	81.4	20.0	66.7	83.3	85	5
AL4–C6	78.8	20.0	70.0	90.0	55	1.33
Av.	81.2	12.7	77.8	88.3	70	4.2

false alarm rate (type I) during the simulation, because these false alarms which result from input errors cannot be eliminated in real time.

6.2.4 Conflict detection function

Table 7 shows the results obtained for the conflict detection function in a very brief way.

As the conflict detection function uses the same data about given ATC clearances, it is affected by incorrect inputs in the same way as the conformance monitoring function. Therefore, the distinction between false alarm rate type I and false alarm rate type II (see conformance monitoring function) applies here too.

The performance of the SACom prototype in terms of false alarm rate (type II) and detection rate is still acceptable, but it did not reach a performance which is comparable to the one of the air traffic controller. Preventing collisions between two aircraft is one of the main tasks of air traffic control; therefore, the controllers are well-trained to detect and react immediately to conflict situations.

Nevertheless, this function got very good usability ratings between 55 (= "good") and 85 (= "excellent"). The average trust rating was 4.2 (= "often"). An explanation might be that controllers are used to all kinds of conflict detection tools and have already years of experience in using them. Following this, it is assumed that the attitude towards conflict detection functions in general is much better than towards new functions which are not yet widely used.

6.2.5 Correlation function

Table 8 shows the results obtained for the correlation function.

The correlation function had to be modified significantly after exercise runs VL–C1 and GP–C2; therefore, only the last four runs are comparable without restrictions.

The duration of the second phase of this simulation (20 min) was defined as the phase, where an attack by an unauthorized speaker giving false instructions was present.

Table 8 Correlation KPIs

Ex. ID	Optimum alert thresh.	False alarm rate (%)	Detection rate (%)	Time until detect. (s)
VL–C1	118	0.0	31.6	25
GP–C2	120	38.4	61.8	278
AL1–C3	28	50.5	80.7	29
AL2–C4	20	56.9	88.7	96
AL3–C5	25	68.3	78.6	164
AL4–C6	61	16.3	89.0	44
Average C3–C6	34	48.0	84.3	83

The performance of the correlation function was best for exercise run AL4–C6, but produced high false alarm rates during the other exercise runs. These false alarm rates were a direct consequence of false indications of the individual SACom modules, especially because of a well-developed aftereffect of the attack phase during exercise step 7.

The correlated score is the basis for automatic reporting to the security management entity and was not displayed to the controller in these validation runs. Therefore, no SATI or SUS questionnaire was filled out.

7 Conclusions

The described SACom prototype takes into account the security countermeasures defined during the development phase of the GAMMA project. This prototype has the potential to reduce risks affecting the future ATM system. The validation example explains the way to validate if generated security information is usable, beneficial, and reliable for users and operators. The effectiveness of the postulated security countermeasures was successfully measured with the help of introduced security KPIs.

The amount of objective measurements and subjective feedbacks of ATM experts were meaningful and comprehensive. Reliability can be seen as one of the most important features of a security system or a specific security function, which was measured using the KPIs false alarm rate, detection rate, and time until detection.

Regarding the discussed ATM security prototype, the adherence to the proposed ATM security management validation methodology appears to be straightforward and clearly focuses on the development of tailor-made validation exercises.

The used short-time simulations have shown a high success rate in introducing numerous and comparable situations throughout the whole validation campaign. These simulations even revealed a "weak point", because 4 of 6 air traffic controllers experienced a mid-air collision during short-time scenario number 16 (see Sect. 5.3). Obviously, air traffic controllers do rarely notice level deviations of aircraft when they are already on the final approach using an Instrument Landing System (ILS).

The security incident simulation, which replicates a complete attack by an unauthorized speaker infiltrating fake clearances, showed that the SACom prototype can provide valuable support in handling these events and—with reduced false alarms and improved HMI design—will be a very helpful addition, but it cannot guarantee that safety is always maintained.

The lack of a commonly accepted validation methodology for ATM security prototypes, tools, and systems shows that the community is in dire need of defining it. The security validation approach presented in this article has the potential to be the sought-after construction kit and to serve as a guideline for similar validation activities.

Acknowledgements Funding was provided by European Union's Seventh Framework Programme (Grant Agreement no. 312382).

References

1. http://www.eurocontrol.int/tags/atm-security
2. SESAR ATM Security Risk Assessment Methodology, D02, 00.01.04, 02/05/2013
3. Minimum Set of Security Controls, SESAR Project 16.02.05, D05-006, Edition 00.06.00 (2013)
4. Joint Planning and Development Office, Security Annex, "Concept of operations for the next generation air transportation system," Version 2.0, Washington, DC (2007)
5. Fantacci, R., Menci, S., Micciullo, L., Pierucci, L.: A secure radio communication system based on an efficient speech watermarking approach. Secur. Commun. Netw. **2**(4), 305–314 (2009)
6. Prinz, J., Sajatovic, M., Haindl, B.: S2EV—Safety and security enhanced ATC voice system. In: IEEE Aerospace Conference, Big Sky, MT, USA (2005)
7. Hering, H., Hagmüller, M., Kubin, G.: Safety and security increase for air traffic management through unnoticeable watermark aircraft identification tag transmitted with the VHF voice communication. In: IEEE/AIAA Digital Avionics System Conference (DASC) (2003)
8. GAMMA Consortium.: GAMMA CONOPS, the ultimate ATM security framework, newsletter, no 1, pp. 2–3. (2015)
9. EUROCONTROL.: European operational concept validation methodology, version 3.0, https://www.eurocontrol.int/publications/european-operational-concept-validation-methodology-eocvm (2010). Accessed June 2015
10. Stelkens-Kobsch, T.H., Hasselberg, A., Mühlhausen, T., Carstengerdes, N., Finke, M., Neeteson, C.: Towards a more secure ATC voice communications system. In: Digital Avionics Systems Conference (DASC), 2015 IEEE/AIAA 34th. Prague, Czech Republic (2015). https://doi.org/10.1109/dasc.2015.7311419
11. Strohmeier, M., Schäfer, M., Pinheiro, R., Lenders, V., Martinovic, I.: On perception and reality in wireless air traffic communication security. In: IEEE transactions on intelligent transportation systems, vol. 18(6), pp.1338–1357. (2017)
12. International Civil Aviation Organization.: Communication Systems, Annex 10 to the Convention on International Civil Aviation, vol. III, 2nd edn (2007)
13. LiveATC: Fake ATC in Action (LTBA-Istanbul).: http://www.liveatc.net/forums/atcaviation-audio-clips/25-may-2011-fake-atc-in-action-(ltba-istanbul) (2011). Accessed June 2015
14. The Age: Lone-Wolf Radio Hoaxer Hacks Melbourne Air Traffic Control.: http://www.theage.com.au/victoria/lonewolf-radio-hoaxer-hacks-melbourne-airtraffic-control-afp-20161107-gsk12o.html (2016). Accessed Dec 2016
15. Neffe, M., Van Pham, T., Hering, H., Kubin, G.: Speaker segmentation for air traffic control. In: Müller C. (eds) Speaker

Classification II. Lecture Notes in Computer Science, vol. 4441. Springer, Berlin (2007)
16. EUROCONTROL.: Specification for short term conflict alert, 1.0 (edn). http://www.eurocontrol.int/publications/short-term-conflict-alert-stca-specification (2007). Accessed July 2016
17. EUROCONTROL.: Specification for medium-term conflict detection, 1.0 (edn). http://www.eurocontrol.int/publications/medium-term-conflict-detection-mtcd-specification (2010). Accessed July 2016
18. Montefusco, P., Casar, R., Stelkens-Kobsch, T.H., Koelle, R.: Addressing security in the ATM environment. In: ARES 2016, 11th International Conference on Availability, Reliability and Security (2016)
19. SASHA—Situation Awareness for SHAPE.: https://ext.eurocontrol.int/ehp/?q=node/1609 (2012). Accessed Sep 2016
20. Brooke, J.: SUS—System Usability Scale. https://measuringu.com/sus/ (1986). Accessed Sep 2016
21. Dehn, D.M.: Assessing the impact of automation on the air traffic controller: the SHAPE questionnaires. ATC Q. **16**(2), 127–146 (2008)
22. EUROCONTROL.: A Human–Machine interface reference system for EnRoute air traffic control (1995)
23. https://github.com/lfv-mssm/yada. Accessed March 2017
24. Helmke, H., Rataj, J., Muehlhausen, T., Ohneiser, O., Ehr, H., Kleinert, M., Oualil, Y., Schuldner, M., Klakow, D.: Assistant-based speech recognition for ATM applications. In: 11th USA/Europe Air Traffic Management Research and Development Seminar (ATM2015), Lissabon/Portugal (2015)