

# AUTOMATIC SPEECH RECOGNITION IN THE COCKPIT: A COMPARATIVE STUDY OF ASR MODELS FOR PILOT COMMUNICATION

S. Ternus\*, K.K.R. Nareddy†, J. Niebling†, A. Papenfuss\*

\* German Aerospace Center, Institute of Flight Guidance, Lilienthalplatz 7, 38108 Braunschweig, Germany

† German Aerospace Center, Institute of Data Science, Mälzerstraße 5, 07745 Jena, Germany

## Abstract

Automatic Speech Recognition (ASR) has seen significant advances in aviation, particularly in Air Traffic Control (ATC), however intra-cockpit communication between pilots has remained largely unexplored despite its central role in teamwork and decision-making. This paper takes an application-oriented perspective and examines how openly available state-of-the-art ASR models perform when applied to intra-cockpit communication without any domain-specific adaptation. We evaluate OpenAI's Whisper (Large-v3 and turbo variant), Wav2Vec2-XLSR-53 as a base model with fine-tuned English, German and multilingual versions, and Meta's Massively Multilingual Speech (MMS) model. Using a dataset of 409 manually transcribed speech segments collected from simulator flights, this paper classifies cockpit communication into six categories and assess performance using Word Error Rate (WER) for each model and category. Results show that Whisper Large consistently achieves the lowest average error rates and demonstrates strong multilingual handling, though it is prone to outliers and occasional hallucinations. Wav2Vec-based models, while less accurate overall, avoid generative errors, with monolingual fine-tuned models working better in language-specific contexts and multilingual variants being able to adapt to code-switching in some cases. The findings highlight trade-offs between consistency, multilingual capability, and computational work, and point to the potential of domain-specific fine-tuning, as this enables improvements in specialized terminology handling. These insights provide a foundation for applying ASR to cockpit communication in both human factors research and future Human–AI Teaming (HAT) applications.

## Keywords

Automatic Speech Recognition; Cockpit Communication; Human–AI Teaming

## 1. INTRODUCTION

Automatic Speech Recognition (ASR) refers to the process of automatically converting spoken language into text. In recent years, advances in deep learning and end-to-end architectures have significantly improved recognition accuracy across a wide variety of domains. Within aviation, ASR has received growing attention, particularly in Air Traffic Control (ATC), where it has been applied to enhance operations and safety through functionalities such as call sign highlighting, radar label pre-filling, and readback error detection [1–5].

Progress in this area has been enabled by large-scale dataset development and domain-specific fine-tuning [2, 6, 7]. Major corpora such as the *ATCO2 corpus* [7] and the *Singapore S-ATC corpus* [8] have provided the foundation for benchmarking models, leading to Word Error Rates (WER) as low as 4–15% when systems are fine-tuned on ATC data [1–3].

While these advances demonstrate the possibilities for ASR in aviation, most research has concentrated on the structured and standardized communication between ATCOs and pilots. In contrast, intra-cockpit communication, the interaction between flight crew

members, remains largely unexplored despite its central role in teamwork, decision-making, and adherence to standard operating procedures.

### 1.1. Motivation

The ability to passively capture and transcribe intra-cockpit speech has important implications for both research and operational practice.

First, ASR provides a non-intrusive and effective means for human factors research, offering the opportunity to analyze teamwork and communication strategies [9]. Manual transcription of crew communication is typically very time-consuming, whereby ASR could substantially reduce the effort for such investigations.

Second, ASR could serve as a key enabling component for systems that continuously monitor crew communication and support situation awareness during operations. Automated systems could verify whether checklist items have been verbally confirmed, whether standard briefing elements are completed, or whether procedural steps are omitted, all without requiring active input from the crew. In this way, ASR could contribute to another safety layer, ensuring pro-

cedural adherence and detecting potential deviations in real time.

Third, accurate transcription of cockpit dialogue could serve as the foundation for future speech-based interfaces supporting Human-AI Teaming (HAT). In line with EASA's concept paper [10], cockpit systems are anticipated to encompass not only Level 1 applications that assist or augment pilots, but also Level 2 applications enabling HAT. At Level 2A ("human-AI cooperation"), the AI supports pilot decision-making by suggesting or implementing directed actions while full authority remains with the pilot. At Level 2B ("human-AI collaboration"), interaction becomes more dynamic, e.g. with the AI capable of sharing situation awareness, adjusting strategies in real time, and communicating to support collaborative decision-making [10]. This might require dialogue in spoken natural language [10, 11]. Therefore, ASR in intra-cockpit communication could lay the groundwork for bidirectional speech-based interaction between pilots and intelligent systems, providing an intuitive modality that aligns with cockpit workflows.

## 1.2. Research Objectives and Contributions

Against this background, the paper takes an application-oriented perspective, focusing on the systematic evaluation of openly available, state-of-the-art ASR models for intra-cockpit communication. The objective is to determine how well such models perform without use-case specific adaptation when applied to different scenarios in intra-cockpit communication. Specifically, the paper:

- Classifies cockpit communication scenarios, so the variation in speech structure, lexical content, and multilingual exchanges is captured.
- Compares multiple ASR models across these scenarios using WER as the evaluation metric.
- Analyzes model-specific strengths and weaknesses across different communication scenarios.

## 1.3. Structure of the Paper

Hereby, the paper is organized as follows: Section 2 introduces and classifies intra-cockpit communication and outlines the associated challenges for ASR. Section 3 describes the methodology, including the dataset, models, metrics, and normalization procedures. Section 4 presents the results for overall model performance and per communication scenario. Section 5 discusses the presented results and analyzes strengths and weaknesses of the models, followed by limitations and future work in Section 6. Finally, Section 7 concludes the paper.

## 2. COCKPIT COMMUNICATION

The following subsections first provide an overview and classification of the main types of cockpit communication and then discuss specific challenges cockpit communication characteristics pose for ASR.

### 2.1. Types of Communication in the Cockpit

Communication in aviation can be broadly distinguished into three domains: pilot-to-pilot (intra-cockpit), pilot-to-ATC, and pilot-to-cabin communication. Pilot-to-pilot or intra-cockpit communication refers to all verbal exchanges between the pilot flying and the pilot monitoring within the cockpit. Pilot-to-ATC communication encompasses the standardized phraseology exchanged between the flight crew and air traffic controllers. Pilot-to-cabin communication describes the interaction between the flight crew and cabin crew.

The present study focuses on intra-cockpit communication, for the reasons presented in Section 1.1. Based on discussions with a retired airline pilot, domain experts and the analysis of cockpit recordings obtained in simulator studies, intra-cockpit communication can be classified into six principal categories:

- 1) *Checklists*: Standardized procedures and phraseology used to systematically verify the correct configuration of aircraft systems and the completion of required actions across different phases of flight and pre-flight. Communication is highly structured and follows a fixed call-and-response pattern, with highly specific technical vocabulary almost exclusively in English.
- 2) *Briefings*: Structured pre-flight or phase-of-flight briefings covering essential elements such as departure and approach procedures, weather conditions, and alternate aerodromes. While the points to be discussed are fixed, the phrasing is more flexible, combining repeated technical terms in English with explanatory speech in the crew's native language (here German).
- 3) *Aircraft handling*: Verbal exchanges directly related to the control and monitoring of the aircraft, including mandatory callouts such as "V1", "Rotate", or "Positive climb". These are standardized and mostly in English, but are sometimes mixed with short, spontaneous German comments on handling and performance, producing a mixture of codified callouts and situational remarks.
- 4) *Navigation and Flight Path Management*: Communication concerned with planning, monitoring, and adjusting the flight path, including references to waypoints and altitudes. Utterances combine structured confirmations in English with evaluative, interactive dialogue in German.
- 5) *Decision-making and situational communication*: Exchanges that support the joint assessment of the situation, negotiation of options, and agreement on a course of action, particularly in dynamic or unexpected contexts. Here, the language here is less predictable and more discursive, with reasoning and evaluation in German combined with English technical terms.
- 6) *Abnormal and emergency procedures*: Communication during the structured execution of abnormal or emergency checklists, either via the Electronic

Centralized Aircraft Monitor (ECAM) or the Quick Reference Handbook (QRH). ECAM-related exchanges are characterized by a dense sequence of technical terminology and procedure-specific items, with a primary focus on the systematic identification of malfunctions and the articulation of corresponding mitigation steps. While the procedural items themselves are standardized in English, crews may read aloud or comment in German, resulting in a multilingual exchange (e.g., “Wir haben nur noch einen Spoiler pro Seite.”).

In addition, there is also social communication between pilots, which may include small talk or non-operational remarks. While such exchanges are part of the natural communication environment, they are excluded from the present analysis due to their limited operational relevance.

## 2.2. Challenges for ASR in Cockpit Communication

ASR in the cockpit environment faces several challenges that distinguish it from other application domains. A primary factor is background noise, which includes continuous engine sounds, airflow, alarms, and radio static. Another challenge is that non native English speaking flight crews often operate in bilingual or multilingual contexts, differences in accents, dialects etc. Another important aspect is aviation-specific phraseology and vocabulary. All of this becomes even more complex given the heterogeneity of communication types within the cockpit. As displayed in the previous section, communication ranges from highly structured and predictable phrases during checklists and callouts, to semi-structured briefings expressed in formalized but natural language, and to decision-making and situational exchanges. The latter are often guided by structured procedures such as FOR-DEC (Facts, Options, Risks & Benefits, Decision, Execution, Check), a standardized decision-making model in aviation, but still realized in more flexible and less predictable language.

## 3. METHODOLOGY

### 3.1. Dataset

The dataset was compiled from recordings obtained in four independent studies, comprising a total of 18 cockpit simulator flight recordings. In all studies, informed consent was obtained from the participants. The recordings were produced in two different simulator environments: a full flight simulator and a flat panel simulator. The original recordings were in video format and were converted to WAV audio files for further processing. The audio material is multilingual, containing a mix of German and English, reflecting the language use commonly observed among German pilots. All speakers in the dataset are male pilots.

Segmentation of the audio was carried out automatically through a Python script using the `pydub` library<sup>1</sup>. Each recording was processed to identify pauses in communication, with a maximum segment length of 30 seconds and a minimum silence duration of 500 ms used for detection. The 30-second limit was chosen because the Hugging Face ASR pipeline processes inputs of up to 30 seconds, requiring longer recordings to be split accordingly. To avoid cutting off words in the middle, which could degrade recognition accuracy, segmentation points were aligned with detected silences. Silence was defined relative to the average signal level, with a threshold of  $-35$  dBFS. This threshold was adjusted when necessary depending on the noise level of the recording. The script first detected all silent intervals meeting these criteria, calculated their midpoints, and then split the audio at the latest silence point within each 30-second analysis window. Each segment was saved as a separate WAV file.

Category	Segments	Duration (minutes)
1	27	11.5
2	58	25.9
3	59	24.2
4	29	12.0
5	160	72.9
6	76	35.6
<b>Total</b>	<b>409</b>	<b>182.1</b>

**TAB 1. Overview of dataset distribution across categories**

Manual ground truth transcripts were created for each segment. A set of categories representing the speech situation was defined in collaboration with a pilot and domain-experts as discussed in Section 2 and each segment was assigned one or more categories according to its content. Since the segmentation process was independent of conversational boundaries, a single segment could contain content from multiple categories or include a category switch within its duration. To ensure that each sample in the final dataset represented a single, unambiguous category, all segments containing multiple categories or category switches were removed. Furthermore, segments containing content outside of the set categories, like social communication, communication with ATC or discussions with the researchers during the simulation were excluded. After this filtering step, the final dataset consists of 409 speech segments with an overall duration of about 182 minutes. The distribution of those segments across the different categories of cockpit communication is shown in Table 1. Each remaining segment has an associated WAV audio file, a verified transcript, and a single unique category label.

<sup>1</sup><https://www.pydub.com/>

### 3.2. Models

In order to compare different ASR models for the target application, we focused on open-source models with broad community adoption, reproducibility, and competitive benchmark performance. Another selection criterion was the availability of multilingual support and integration within the Hugging Face ecosystem, which ensures that all models can be evaluated using the same pipeline and allows for easy implementation and direct comparability across models. Based on these considerations, we restrict our comparison to Whisper, representing large-scale weakly supervised training with multiple model sizes to examine scaling trade-offs, Wav2Vec-XLSR-53, representing self-supervised representation learning with different finetuned heads and Massively Multilingual Speech (MMS), which extends this line of work to a big language coverage.

#### 3.2.1. OpenAI Whisper

Whisper [12] is a family of encoder-decoder transformer models trained on 680,000 hours of multilingual and multitask supervised data. It supports over 90 languages and has shown robustness to domain and environmental variability, making it suitable for noisy communication channels. Whisper is fully open source and provides different model sizes (tiny to large). In this work, we primarily focus on Whisper Large-v3<sup>2</sup>, hereafter mentioned as Whisper Large and Whisper Turbo<sup>3</sup>, since they offer better recognition accuracy compared to the smaller models while still being practical to deploy on modern hardware [12, 13]. Whisper is widely used in monolingual and multilingual ASR tasks [14, 15].

#### 3.2.2. Wav2Vec 2.0

Wav2Vec 2.0 [16] introduced a self-supervised learning paradigm for ASR by pretraining on large quantities of unlabeled speech data and fine-tuning on labeled datasets. The architecture consists of a convolutional feature encoder that transforms raw audio into latent representations, followed by a transformer network that contextualizes these representations. Pre-training is carried out using contrastive learning, and the model can then be fine-tuned with relatively small amounts of labeled data. This approach has demonstrated state-of-the-art results on benchmarks such as LibriSpeech and has shown strong adaptability to both high- and low-resource languages.

In this study, we evaluate two major pretrained base models derived from Wav2Vec 2.0 and their respective fine-tuned variants for ASR:

#### Wav2Vec2-XLSR-53

Wav2Vec2-XLSR-53<sup>4</sup> [17] is a pretrained base model built on Wav2Vec 2.0. It was trained on raw speech

<sup>2</sup><https://huggingface.co/openai/whisper-large-v3>

<sup>3</sup><https://huggingface.co/openai/whisper-large-v3-turbo>

<sup>4</sup><https://huggingface.co/facebook/wav2vec2-large-xlsr-53>

waveforms from 53 languages and learns cross-lingual speech representations through contrastive learning and shared quantization of latent features. As a base model, it is not directly optimized for speech recognition but requires fine-tuning on labeled downstream data for ASR. From this, the following models were derived:

- **XLSR-Multilingual-56**<sup>5</sup> [18]: Fine-tuned on 56 languages using Mozilla Common Voice.
- **XLSR-English**<sup>6</sup> [19]: Fine-tuned on English (Common Voice 6.1).
- **XLSR-German**<sup>7</sup> [20]: Fine-tuned on German (Common Voice 6.1).

#### MMS-1B

The MMS project [21], introduced by Meta AI, extends the Wav2Vec 2.0 architecture to a large multilingual setting. It is pretrained with Wav2Vec2's self-supervised training objective on about 500,000 hours of speech data in over 1,400 languages. From this base, the following variant was used in this work:

- **MMS**<sup>8</sup> [21]: Fine-tuned on 1,162 languages for multi-lingual ASR.

### 3.3. Metrics

Transcription accuracy was evaluated using WER and was computed with the `jiwer` library<sup>9</sup>, which provides a standardized implementation for text-based error measurement. WER is defined as

$$(1) \quad \text{WER}(\%) = \frac{S + D + I}{N} * 100\%,$$

where  $S$  represents the number of substitutions,  $D$  the number of deletions,  $I$  the number of insertions, and  $N$  the total number of words in the reference text. A lower WER indicates a more accurate transcription. Moreover, we measured the average processing duration (in seconds) required to automatically transcribe an audio segment. This metric was obtained on a workstation with an AMD 48-Core EPYC processor and an NVIDIA A100 GPU with 48 GB VRAM.

### 3.4. Normalization

Prior to computing WER, transcripts were normalized to eliminate differences that were not relevant for the evaluation. We applied OpenAI's normalization functions<sup>10</sup> [12], which handle case folding and removal of special characters, convert numeric expressions into Arabic numerals, and unify spelling variants. In addition, we implemented custom normalization tailored to cockpit communication. First, ICAO alphabet words

<sup>5</sup><https://huggingface.co/voidful/wav2vec2-xlsr-multilingual-56>

<sup>6</sup><https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english>

<sup>7</sup><https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german>

<sup>8</sup><https://huggingface.co/facebook/mms-1b-all>

<sup>9</sup><https://jitsi.github.io/jiwer/>

<sup>10</sup><https://github.com/openai/whisper>

Category	MMS	XLSR-Multil.	XLSR-Ger.	XLSR-Eng.	Whisper Large	Whisper Turbo
1	83.5 ± 16.7	<b>87.8 ± 12.7</b>	86.6 ± 14.0	75.7 ± 16.9	<b>28.2 ± 24.9</b>	40.1 ± 47.5
2	57.7 ± 19.5	79.2 ± 10.3	60.7 ± 13.5	<b>93.3 ± 4.7</b>	<b>22.3 ± 19.2</b>	24.9 ± 13.9
3	94.5 ± 49.5	93.9 ± 10.5	88.4 ± 19.9	<b>98.0 ± 21.8</b>	<b>53.4 ± 40.7</b>	63.5 ± 49.0
4	86.3 ± 22.0	90.7 ± 11.4	81.8 ± 18.0	<b>95.7 ± 20.4</b>	<b>43.1 ± 25.6</b>	53.9 ± 26.3
5	68.2 ± 17.7	85.0 ± 10.4	70.4 ± 14.0	<b>97.0 ± 4.2</b>	<b>25.7 ± 16.1</b>	30.2 ± 22.4
6	74.7 ± 20.2	<b>85.0 ± 10.5</b>	79.1 ± 11.4	77.0 ± 14.4	<b>29.4 ± 22.6</b>	33.1 ± 21.0

**TAB 2. Normalized mean WER(%) (± SD) for each model across communication categories. Red values indicate the worst performance within a category, while blue values highlight the best performance.**

were mapped to their corresponding single letters (e.g., *DELTA* → *D*). Next, compound word variants were normalized using regular expression matching, so that forms such as “take-off,” “takeoff,” and “take off” were replaced by a unified representation. Finally, a predefined list of filler words (e.g., *um*, *uh*, *ähm*) was filtered out by removing them from the text.

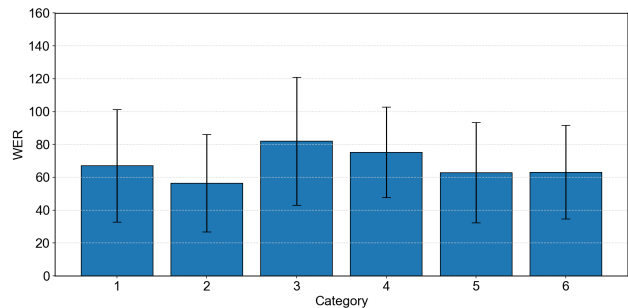
### 3.5. Procedure

The evaluation followed a within-subject design, with all models applied to every category. In a first step, transcripts were generated for each audio segment, that were transferred to 16kHz and the processing time required per segment was recorded. The transcripts were then normalized. Based on the normalized text, the WER was calculated for every segment. Finally, the results were aggregated by computing the mean and Standard Deviation (SD) of the WER values for each model within each category.

## 4. RESULTS

### 4.1. Accuracy

Table 2 summarizes the results and presents the normalized WER across the evaluated models. In the Whisper family, Whisper Large generally achieved the lowest error rates and consistently outperformed all other models, with lowest WERs of 22.3% in Briefings (2) and 25.7% in *Decision-making and situational communication* (5), as displayed in Figure 2. In the Wav2Vec family, MMS and XLSR-German provided comparatively lower error rates than XLSR-Multilingual and XLSR-English in those categories as well, but the overall error levels remain substantially higher than those of the Whisper models. XLSR-English, on the other hand, performed strongest for *Checklists* (1) (75.7%), within the Wav2Vec models. In *Aircraft Handling* (3) and *Flight Path Management* (4) all Wav2Vec models showed high error rates above 88%, with XLSR-German performing comparatively best, followed by the multilingual models and XLSR-English performing the worst, again all showing higher WERs than both Whisper model variants. In the category of *Abnormal and emergency procedures* (6), the multilingual Wav2Vec models differed more. MMS achieved the lowest error rate at 74.7%,



**FIG 1. Mean normalized WER (%) for all categories pooled over all models**

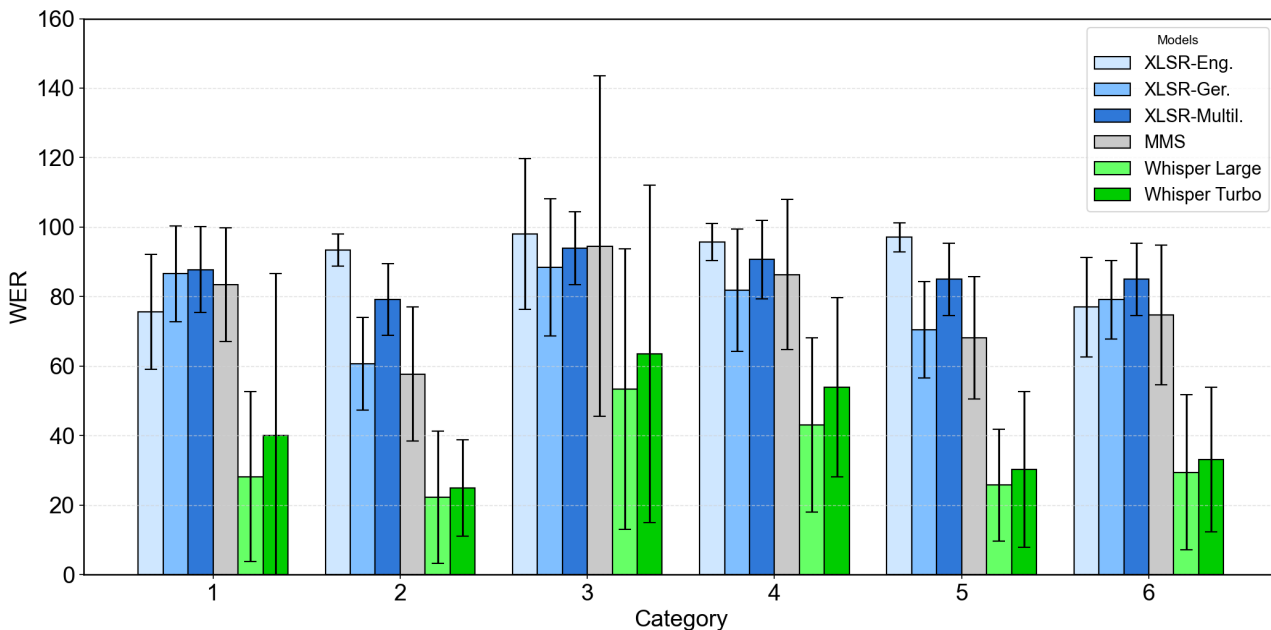
followed by XLSR-English, while XLSR-Multilingual showed the highest at 85.0%.

Category-wise, *Briefings* (2) presented as the most accurately transcribed communication category aggregated over all models, followed by *Decision-making and situational communication* (5), as can be seen in Figure 1. *Checklists* (1) and *Abnormal and emergency procedures* (6) also showed strong results for Whisper Large, but presented bigger differences in model performance for the Wav2Vec models. *Aircraft handling* (3) and *Navigation and Flight Path Management* (4) exhibited the highest error rates over all models.

### 4.2. Robustness

Across categories, Whisper models demonstrated overall lowest WERs. However, this performance at the average level was accompanied by considerably high variability, as can be seen in Figure 2. SDs for Whisper sometimes even exceeded 40%, showing that while many utterances were transcribed with high accuracy, others resulted in severe errors. In contrast, the XLSR models, particularly the multilingual variant, consistently showed lower variation (SD mostly around 10–15%), reflecting more stable performance across utterances. MMS consistently performed worse than the Whisper models in terms of accuracy, while at the same time exhibiting substantially higher SDs than the XLSR variants.

Whisper's fluctuations are further reflected in the following example. In several instances, Whisper models produced repetitions or hallucinations, where a briefly present cue in the audio was expanded



**FIG 2. Normalized mean WER (%) by communication category across all models (with SD)**

into an extended sequence that did not occur in the recording.

**Reference:** “Checklist. Before Takeoff, Flight Controls. Checked [...] 137, 137, 140, und Toga, Toga takeoff. [...]”

**Whisper Turbo:** “B4 tagger, Flakadrolls Checked [...] 137, 137, 140 und **Toga, Toga, Toga, Toga** [...]”  
 WER (whole segment): 241.1%  
 WER w.o. Hallucination: 45.5%

The repetitive expansion of the token “Toga” is absent from the reference and inflates segment-level WER. While infrequent, such episodes explain a non-trivial share of the observed outliers and the correspondingly elevated standard deviations in the Whisper results.

**4.3. Multilingual Capability**

Furthermore, as expected, the language-specific errors of the German- and English-finetuned Wav2Vec models are readily observable. For a German statement, XLSR-German remains close to the reference, whereas XLSR-English shows an English bias. For an English phrase, the opposite pattern emerges. The multilingual Wav2Vec models often perform nearly as well as the language-finetuned variants (e.g., here XLSR-German) and diverge less than models fine-tuned on a different language (e.g., here XLSR-English). The following excerpts illustrate these patterns <sup>11</sup>.

**Reference:** “Also einen Fakt möchte ich noch mit einbringen [...]”

**XLSR-German:** “also ein fakt möchte ich noch mit einbringen [...]”  
 WER: 12.5%

**XLSR-English:** “so in fact must not in bring [...]”  
 WER: 100%

**XLSR-Multilingual:** “also einfach möchte ich noch mit einbringen [...]”  
 WER: 25%

**MMS:** “also ein fakt michte ich noch mit einbringen [...]”  
 WER: 25%

In mixed-language segments, the multilingual variants sometimes manage the code-switch but often struggle with selecting the appropriate language, which frequently results in compounded tokens and non-words. Hereby, MMS generally handles the switches more reliably than the XLSR-Multilingual model, which places it above XLSR-Multilingual among the multilingual models but still below both Whisper variants, as displayed in the example below<sup>11</sup>.

**Reference:** “[...] slats und flaps sind slow und wir haben CAT 1 only [...]”

**MMS:** “[...] slats und flaps sin slow und biharm cat 1 only [...]”  
 WER: 27.3%

**XLSR-Multilingual:** “[...] slats and flapson slow und beham caton only [...]”  
 WER: 63.6%

<sup>11</sup>WER is computed only for the displayed part of the segment, to demonstrate the effect of deviation more clearly.

**XLSR-German:** “[...] slaps und flaps sind slo und bhm kat vononli [...]”  
WER: 63.6%

**XLSR-English:** “[...] slats on flaps in slow unihan cat 1 only [...]”  
WER: 45.5%

**Whisper Large:** “[...] slats und flaps sind slow und wir haben CAT 1 only [...]”  
WER: 0.0%

#### 4.4. Computational Work

To complement the accuracy evaluation, we also measured the average processing duration required to transcribe a segment. The results are summarized in Table 3. Among the Whisper models, Whisper Turbo was the faster option (0.65 s), whereas Whisper Large required considerably more time (1.86 s). The MMS-1B model also achieved relatively low latency (0.38 s). Within the XLSR family, the multilingual variant proved especially efficient (0.20 s), clearly outperforming the German (0.70 s) and English (0.78 s) versions.

Model	time
XLSR-German	0.70
XLSR-English	0.78
XLSR-Multilingual	0.20
MMS	0.38
Whisper Turbo	0.65
Whisper Large	1.86

**TAB 3. Average times (s) for transcribing a segment across ASR models.**

## 5. DISCUSSION

### 5.1. Accuracy

The results show that Whisper Large consistently achieved the lowest error rates across all categories compared to the Wav2Vec-based models, demonstrating comparatively strong performance across different communication types, speakers, and noise levels. When set against standard benchmarks however, the observed WER values appear high: Whisper Large reaches only 5.7% WER on Common Voice 15 for German and 9.3% for English [22], whereas the intra-cockpit results are in the range of 22–53%. At the same time, these values are comparable to zero-shot evaluations of Whisper Large on ATC corpora, where WERs of 17.98% and 29.05% have been reported after normalization [3]. MMS and XLSR models also perform poorly in comparison to general benchmarks, where WERs below 10% are common [16], often exceeding 70–90%. This demonstrates both the capabilities of the Whisper models

and the particular difficulty of cockpit speech, with its mixed-language phraseology, noise, and disfluencies. Furthermore, error rates are still strongly affected by the handling of specialized terminology. Domain-specific phrases that do not consist of words used outside of aviation were often transcribed inaccurately by all models. To address these limitations, both model families can be fine-tuned to domain-specific vocabularies. Whisper fine-tuning has been shown to improve transcription accuracy for technical terminology [23, 24]. For Wav2Vec-based models, fine-tuning has demonstrated reductions in WER across multiple domains, including low-resource and noisy environments [16, 25–27].

### 5.2. Robustness

While average accuracy provides one perspective, the degree of variation across utterances highlights another important aspects. Whisper models demonstrated the lowest average WER, suggesting strong robustness across different speakers and noise conditions at the aggregate level. However, their performance was characterized by high variability and occasional severe errors, reflected in large standard deviations and extreme outliers. Related to this phenomenon is another risk inherent to Whisper’s encoder–decoder Transformer architecture with generative capabilities: it is prone to “hallucinations,” where words or entire sentences are produced that are not present in the input audio [23, 28]. Such behavior was observed in the present study in the form of repetitions and expansions of short cues into long, incorrect sequences.

By contrast, the XLSR models, while not always reaching Whisper’s best-case accuracy, showed more stable performance with consistently lower variation across categories. Wav2Vec-based models are also structurally less prone to generative errors. MMS performance fell in between, with lower accuracy than whisper but higher variability than the XLSR model variants.

These findings suggest that model choice depends on the application context. In safety-critical domains, the more stable and non-generative nature of Wav2Vec-based models may be preferable, whereas in exploratory or less safety-critical contexts, such as human factors research where inherent multilingual capabilities and strong out-of-the-box accuracy are valuable, Whisper may be the more suitable option.

### 5.3. Multilingual Capability

A further factor to consider is multilingual capability. Whisper Large and Turbo are inherently multilingual [22] and prove strong capabilities when the spoken language is not predetermined, like in the multilingual cockpit environment studied here. In our examples, Whisper often managed multilingual switches more successfully than the other models, demonstrating its suitability for such mixed-language scenarios. In contrast, although multilingual

Wav2Vec variants such as XLSR-Multilingual and MMS were also trained on multiple languages, they demonstrated greater difficulty in distinguishing and consistently transcribing mixed-language inputs. The fine-tuned monolingual models performed better in their trained language: XLSR-English achieved the best results on English-dominated Checklists, while XLSR-German performed better on predominantly German heavy categories like Briefings, even though MMS sometimes outperformed it in those categories, since they still include English aviation terms. This suggests that a two-step pipeline with automatic language identification followed by transcription using a language-specific model could improve the accuracy in transcription for Wav2Vec models. Even though Whisper already fulfills this requirement, it sometimes introduces automatic translation in longer segments, which can drastically reduce accuracy, as noted in other studies [23].

#### 5.4. Computational Work

Lastly the practical application of Whisper, Wav2Vec2, MMS models raises questions regarding efficiency. Benchmarks indicate that Whisper models, despite offering strong zero-shot generalization, are relatively resource-intensive due to their transformer encoder–decoder architecture, with inference latency scaling with model size [12]. The results regarding the computational work, presented in Section 4.4, highlight the trade-off between model size and latency, with lighter architectures such as Whisper-turbo providing substantial efficiency gains relative to larger configurations such as Whisper Large. For efficient inference, various model adaptation methods have been used in conjunction with Whisper [13–15]. Overall, Whisper Large tends to provide better accuracy and multilingual coverage but at higher computational cost, whereas Whisper Turbo model is often more efficient for real-time transcription, while preserving a good transcription accuracy.

### 6. LIMITATIONS & FUTURE WORK

This study has several limitations that should be acknowledged. The first limitation concerns the dataset. Here, the speech data consisted exclusively of male speakers, which means the findings may not generalize to female voices or to a more diverse set of speakers. Moreover, the distribution of utterances across the different communication categories was not balanced, which may have biased performance comparisons and the data was collected exclusively in simulator environments, which may not fully capture the acoustic and operational characteristics of real-world cockpit communication. Furthermore, the recordings originated from different simulator setups, introducing variations in audio quality and background noise that may have influenced the observed error rates. To better capture the performance over a more diverse speaker set and noise levels and to enable

training of specialized ASR models, future work could focus on the creation and recording of a dedicated dataset. Such a dataset should employ a unified and standardized recording setup that closely reflects real cockpit acoustics while including diverse speakers of different ages, genders, and accents. Each audio file should hereby contain only one complete statement or phrase of operational relevance. To further enhance robustness, the dataset should systematically include variations in background noise levels and conditions. Furthermore, it could also deliberately include typical communication occurrences like overlaps, hesitations, disfluencies, and less distinct pronunciations, that reflect operational communication. Accordingly, fine-tuning existing models on aviation-specific data represents a key next step for improving the handling of technical terminology and reducing error rates. In addition to dataset limitations, the scope of this study was restricted to a limited set of ASR architectures. While Whisper, Wav2Vec2, and MMS capture important families of current approaches, future work should extend the evaluation to additional architectures such as HuBERT [29], WavLM [30], SEW/SEW-D [31, 32], and VoxTral [33] in order to obtain a more comprehensive picture of strengths and weaknesses across design paradigms. Finally, all utterances were weighted equally in the evaluation, meaning that longer and more complex phrases were treated the same as shorter ones. Future analyses could account for utterance length or operational criticality to provide more nuanced insights.

### 7. CONCLUSION

In conclusion, this paper provided a systematic evaluation of openly available ASR models for intra-cockpit pilot communication from an application-oriented perspective. By examining different Whisper and Wav2Vec 2.0-XLSR variants, as well as MMS on a dataset of simulator recordings, we demonstrated how models pretrained on general, non-domain specific audio data perform in this setting without additional adaptation.

The results showed that Whisper Large consistently achieved the lowest error rates, but it was also prone to variability and occasional hallucinations. Wav2Vec-XLSR models were less accurate overall but displayed more stable behavior without generative errors. Whisper further proved most capable in handling multilingual exchanges and code-switching, whereas language-finetuned Wav2Vec models performed comparatively better only when the spoken language matched their training, and multilingual variants struggled more with mixed-language input. In terms of computational efficiency, Whisper Turbo provided a practical balance between speed and accuracy, while Whisper Large required considerably more processing time.

Taken together, these findings highlight important trade-offs between accuracy, robustness, multilingual capability, and computational efficiency. They also

underline that current models, while already useful for human factors research and exploratory applications, still fall short of the reliability required for direct operational use in safety-critical environments. Future work should therefore explore domain-specific fine-tuning, targeted handling of aviation terminology, and hybrid pipelines that combine automatic language identification with specialized ASR models. These steps will be essential for moving from proof-of-concept transcription towards robust integration of ASR in cockpit workflows and, ultimately, for enabling speech-based interaction in future HAT concepts.

#### Contact address:

[sarah.ternus@dlr.de](mailto:sarah.ternus@dlr.de)

#### References

- [1] Zhuang Wang, Peiyuan Jiang, Zixuan Wang, Boyuan Han, Haijun Liang, Yi Ai, and Weijun Pan. Enhancing air traffic control communication systems with integrated automatic speech recognition: Models, applications and performance evaluation. *Sensors*, 24(14), 2024. ISSN: 1424-8220. DOI: [10.3390/s24144715](https://doi.org/10.3390/s24144715).
- [2] Moritz May, Matthias Kleinert, and Hartmut Helmke. Automatic transcription of air traffic controller to pilot communication - training speech recognition models with the open source toolkit coquistt. In *DLRK, Deutscher Luft- und Raumfahrtkongress 2024*, pages 1–8, Dezember 2024. DOI: [10.25967/630171](https://doi.org/10.25967/630171).
- [3] Jan van Doorn, Junzi Sun, J.M. Hoekstra, Patrick Jonk, and Vincent de Vries. Whisper-atc: Open models for air traffic control automatic speech recognition with accuracy. In Eric Neiderman, Marc Bourgois, Dave Lovell, and Hartmut Fricke, editors, *Proceedings International Conference on Research in Air Transportation*, 2024.
- [4] Hartmut Helmke, Matthias Kleinert, Arthur Linß, Lucas Klamert, Petr Motlicek, Julia Harfmann, Nuno Cebola, Hanno Wiese, Hörður Arilíusson, and Teodor Simiganoschi. The haawaii framework for automatic speech understanding of air traffic communication. In *13th SESAR Innovation Days 2023, SIDS 2023*, November 2023.
- [5] S. Shetty, H. Helmke, M. Kleinert, and O. Ohneiser. Early callsign highlighting using automatic speech recognition to reduce air traffic controller workload. In Katie Plant and Gesa Praetorius, editors, *Human Factors in Transportation. AHFE 2022 International Conference*, volume 60 of *AHFE Open Access*, USA, 2022. AHFE International. DOI: [10.54941/ahfe1002493](https://doi.org/10.54941/ahfe1002493).
- [6] Juan Zuluaga-Gomez, Petr Motlicek, Qingran Zhan, Karel Vesely, and Rudolf Braun. Automatic speech recognition benchmark for air-traffic communications, 2020. <https://arxiv.org/abs/2006.10304>.
- [7] Juan Zuluaga-Gomez, Karel Vesely, Igor Szöke, Alexander Blatt, Petr Motlicek, Martin Kocour, Mickael Rigault, Khalid Choukri, Amrutha Prasad, Seyyed Saeed Sarfjoo, Iuliia Nigmatulina, Claudia Cevenini, Pavel Kolčárek, Allan Tart, Jan Černocký, and Dietrich Klakow. Atco2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications, 2023. <https://arxiv.org/abs/2211.04054>.
- [8] Phuong Tuan Dat, Luong Trung Tuan, Jayakrishnan Melur Madhathil, and Tran Huy Dat. Automatic speech recognition and understanding over noisy air traffic control vhf channels in singapore. In *SESAR Innovation Days*, Rome, Italy, November 2024. SESAR Joint Undertaking. 12–15 November 2024.
- [9] Anne Papenfuss and Christoph Andreas Schmidt. Using automatic speech recognition to evaluate team processes in aviation - first experiences and open questions. In Don Harris and Wen-Chin Li, editors, *Engineering Psychology and Cognitive Ergonomics*, pages 501–513, Cham, 2023. Springer Nature Switzerland. ISBN: 978-3-031-35389-5.
- [10] European Union Aviation Safety Agency (EASA). Easa concept paper: Guidance for level 1 & 2 machine learning applications, issue 02. Technical report, EASA, Cologne, Germany, 2023.
- [11] Mallory S Graydon, Jon B Holbrook, Natasha A Neogi, Jeffrey M Maddalon, and G Frank McCormick. Challenges, research, and opportunities for human-machine teaming in aviation. 2025.
- [12] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. <https://arxiv.org/abs/2212.04356>.
- [13] Keisuke Kamahori, Jungo Kasai, Noriyuki Kojima, and Baris Kasikci. Liteasr: Efficient automatic speech recognition with low-rank approximation, 2025. <https://arxiv.org/abs/2502.20583>.
- [14] Thomas Palmeira Ferraz, Marcely Zanon Boito, Caroline Brun, and Vassilina Nikoulina. Multilingual distilwhisper: Efficient distillation of multi-task speech models via language-specific experts, 2024. <https://arxiv.org/abs/2311.01070>.
- [15] Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. Lora-whisper: Parameter-efficient and extensible multilingual asr, 2024. <https://arxiv.org/abs/2406.06619>.

- [16] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.
- [17] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition, 2020. <https://arxiv.org/abs/2006.13979>.
- [18] voidful. voidful/wav2vec2-xlsr-multilingual-56: Fine-tuned XLSR model on 56 languages. <https://huggingface.co/voidful/wav2vec2-xlsr-multilingual-56>, 2024. Accessed: 2025-09-17.
- [19] Jonas Grosman. Fine-tuned XLSR-53 large model for speech recognition in English. <https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english>, 2021. Accessed: 2025-09-17.
- [20] Jonas Grosman. Fine-tuned XLSR-53 large model for speech recognition in German. <https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german>, 2021. Accessed: 2025-09-17.
- [21] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages, 2023. <https://arxiv.org/abs/2305.13516>.
- [22] OpenAI. Whisper github repository. <https://github.com/openai/whisper>, 2022. Accessed: 2025-09-17.
- [23] Kartheek Nareddy, Sarah Ternus, and Julia Niebling. Analyzing and fine-tuning whisper models for multilingual pilot speech transcription in the cockpit. 06 2025. DOI: 10.48550/arXiv.2506.21990.
- [24] Jan van Doorn, Junzi Sun, Jacco Hoekstra, Patrick Jonk, and Vincent de Vries. Whisper-atc open models for air traffic control automatic speech recognition with accuracy. In *Proc. Int. Conf. Res. Air Transp.(ICRAT)*, 2024.
- [25] Morgane Riviere, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. Unsupervised pretraining transfers well across languages. In *ICASSP*, pages 7414–7418. IEEE, 2020.
- [26] Qiantong Xu, Tatiana Likhomanenko, Gabriel Synnaeve, and Ronan Collobert. Self-training and pre-training are complementary for speech recognition. In *ICASSP*, pages 3030–3034. IEEE, 2021.
- [27] Alejandro Vasquez, Wei-Ning Hsu, Tatiana Likhomanenko, Abdelrahman Mohamed, et al. Benchmarking self-supervised speech representation learning. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1212–1228, 2022.
- [28] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, page 1672–1681, New York, NY, USA, 2024. Association for Computing Machinery. ISBN: 9798400704505. DOI: 10.1145/3630106.3658996.
- [29] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.
- [30] Sanyuan Chen, Chengyi Wang, Yu Chen, Shujie Wu, Zhuo Chen, Ziqiang Liu, Yan Chen, Zhu Yao, Jinyu Li, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.
- [31] Yu Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, and Ming Zhou. Sew: Self-supervised learning with extracted weights for speech recognition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 123–132, 2021.
- [32] Sanyuan Chen, Chengyi Wang, Yu Chen, Shujie Wu, et al. Sew-d: Speeding up conformer with dilated convolutions for efficient speech recognition. In *Interspeech*, pages 3625–3629, 2021.
- [33] Soumi Maiti, Yifan Peng, Shukjae Choi, Jee weon Jung, Xuankai Chang, and Shinji Watanabe. Voxlm: unified decoder-only models for consolidating speech recognition/synthesis and speech/text continuation tasks, 2024. <https://arxiv.org/abs/2309.07937>.