Towards Robust Multilingual Cockpit ASR with Whisper Fine-Tuning and Waveform Augmentation

Kartheek Kumar Reddy Nareddy[1], Eric Guenl[1], Sarah Ternus[2], Boris Djartov[2], Anne Papenfuß[2], and Julia Niebling[1]

[1] Institute of Data Science, German Aerospace Center, Mälzerstraße 5, 07745 Jena, Germany
kartheek.nareddy@dlr.de
[2] Institute of Flight Guidance, German Aerospace Center, Lilienthalplatz 7, 38108 Braunschweig, Germany sarah.ternus@dlr.de

Abstract. We present a domain adaptation study of Whisper models for automatic speech recognition (ASR) of cockpit speech. The task is characterized by several significant challenges, including domain-specific aviation vocabulary, code-switching between German and English, short command-like utterances, and acoustic conditions that differ substantially from those represented in the large-scale pretraining data of general ASR models. A corpus of approximately 180 minutes of cockpit recordings was collected and manually annotated by native speakers with an aviation background. The data were split into training and test sets and evaluated using the Whisper model family. The effect of transcript normalization on word error rate (WER) was examined, and a set of normalization schemes was proposed to reduce mismatches between reference and predicted transcripts. The Whisper Large-v3 model was fine-tuned using parameter-efficient Low-Rank Adaptation (LoRA), reducing WER from 57.61% to 24.90% through fine-tuning and the proposed transcript normalization scheme. To improve robustness under limited data, waveform-level augmentation methods were further applied, including jittering, time warping, window slicing, window warping, scaling, and magnitude warping. Fine-tuning on the augmented training set provided an additional WER reduction of approximately 3–4% on the held-out test set. These results show that careful normalization, parameter-efficient fine-tuning, and simple audio augmentation can substantially improve Whisper-based ASR in low-resource, domain-specific cockpit speech settings.

1 Introduction

Automatic Speech Recognition (ASR), the task of transforming audio signals into text, plays a central role in natural language processing [6]. Speech signals vary substantially across speakers, accents, speaking rate, recording conditions, background noise, and conversational context, which makes ASR a difficult problem in practice [6]. ASR has been used in applications such as automatic call handling [11] and personalized AI assistants [31]. Conventional ASR systems are usually built as a pipeline of acoustic feature extraction, acoustic modeling, language modeling, and decoding, often formulated through Bayes' decision rule [4,40,33,15]. Deep learning has changed both acoustic and language modeling, and has gradually led to end-to-end ASR models that learn much of this pipeline jointly [19,5,7,1,3,41].

Publicly available datasets such as LibriSpeech [35], Common Voice [2], and SpeechStew [8] have been important for training and evaluating modern ASR systems. Still, the growth in model size has often outpaced the amount of carefully labeled speech data available for many domains, leaving models prone to overfitting and weak generalization under distribution shift [17]. This limitation has motivated the use of large-scale unlabeled or weakly labeled speech corpora, including BigSSL [50], GigaSpeech [10], and People's Speech [16]. Among recent ASR models, OpenAI's Whisper [41] is notable for its large-scale weakly supervised training on 680,000 hours of multilingual and multitask speech data, which gives it strong generalization across many languages and recording conditions [12].

While Whisper and other ASR models, including Wav2Vec [3], SpeechStew [8], and DeepSpeech [1], perform well on general speech data, their accuracy can degrade in domain-specific settings [49]. Fine-tuning adapts a pre-trained model to a target dataset and has become a practical way to improve ASR performance under such conditions [46,9]. It has been effective in specialized domains and in low-resource languages. For instance, fine-tuning Whisper for Nepali speech produced substantial reductions in Word Error Rate (WER), with improvements of up to 36.2% for the small model [45].

Data augmentation offers a complementary way to improve robustness when labeled speech is limited. Rather than collecting and transcribing new recordings, augmentation creates additional label-preserving training examples from existing audio [37]. Prior work has explored augmentation for low-resource speech recognition [25,42], vocal tract length perturbation [24], noise injection by mixing clean and noisy signals [18], speed perturbation on raw audio [27], and simulated room acoustics for far-field ASR [26]. More recent work has further extended augmentation to neural ASR settings through feature-space transformations such as time and frequency masking in SpecAugment [37], as well as mixup-based augmentation for low-resource end-to-end ASR [32]. Related ideas have also been studied for keyword spotting [43] and multi-stream ASR training [30].

In aviation, ASR has been widely studied for Air Traffic Control (ATC) communication. Domain-specific models such as Whisper-ATC have achieved WER values as low as 1.17% on ATCOSIM simulated data and up to 60% relative improvement through regional fine-tuning [12]. Intra-cockpit communication between pilots, however, remains less explored, especially in the context of fine-tuning large pre-trained ASR models. Recent cockpit-focused studies have begun to benchmark and fine-tune modern ASR systems for multilingual pilot communication [34,48]. Earlier studies considered Hidden Markov Model based transcription systems for cockpit conversations [47,36]. Accurate transcription of cockpit speech can support human factors research, help analyze teamwork and decision-making, and provide a basis for future speech-driven cockpit assistance systems. The cockpit setting remains difficult, as it contains overlapping speech, multilingual exchanges, high noise levels, short command-like utterances, and use-case specific vocabulary.

Early resources such as the ATCOSIM corpus provided clean ATC speech from simulation environments and highlighted the need for domain-specific corpora in aviation speech processing [20]. More recent work has introduced larger ATC datasets and benchmarks, including ATCO2 and related ASR benchmarks, which address challenges

such as accented speech, noisy VHF channels, limited in-domain supervision, callsign recognition, and domain-specific phraseology [52,51]. The Airbus ATC speech recognition challenge further showed that pilot and controller speech differ in difficulty, with pilot speech being notably harder to transcribe than controller speech [38]. Other studies have explored the use of contextual and operational knowledge for ATC speech processing, including speaker-role identification, command extraction, and assistant-based speech recognition systems for controller support [39,44].

Air-traffic communication follows a highly constrained radiotelephony style, where standard phraseology is used to reduce ambiguity and improve mutual understanding between pilots and controllers [22]. Typical conventions include the use of aircraft call signs, concise controller instructions, pilot readbacks, standardized words for clearances and instructions, and careful pronunciation of letters and numbers. Readback is particularly important for safety-critical information such as runway assignments, altitude clearances, headings, and route instructions, since it allows the controller and pilot to verify that the instruction was correctly received [14]. Prior safety studies also identify call-sign confusion, simultaneous transmissions, radio interference, and deviations from standard phraseology as recurring issues in air-ground communication [13]. These conventions make aviation speech different from ordinary conversational speech: utterances are often short, formulaic, number-heavy, and strongly dependent on operational context.

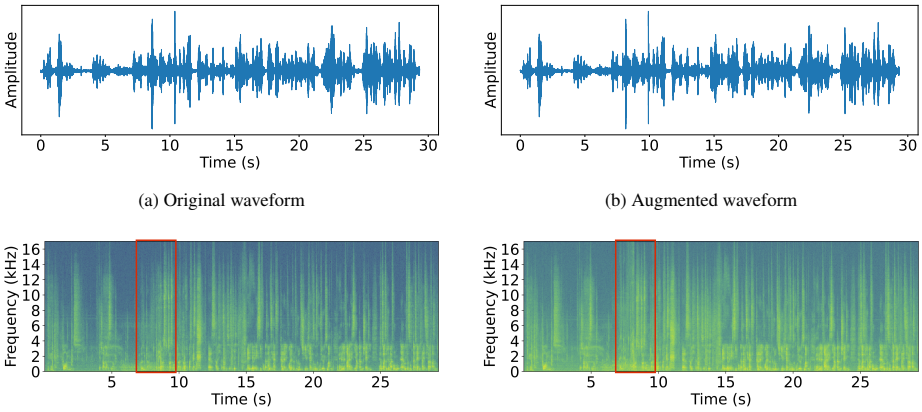
Although our work focuses on intra-cockpit speech rather than controller–pilot radio communication, several of these conventions remain relevant. Cockpit briefings and checklist interactions contain callouts, confirmations, aviation abbreviations, ICAO alphabet terms, procedural commands, and readback-like responses between crew members. Moreover, multilingual cockpit communication may mix English aviation phraseology with German conversational or explanatory speech. This combination creates a difficult ASR setting, where the model must preserve the spoken language, avoid unintended translation, and correctly recognize domain-specific terms that may be rare in general speech corpora.

To address these challenges, we explore the fine-tuning of Whisper models for multilingual pilot communication in the cockpit. We use the Hugging Face fine-tuning pipeline and evaluate different Whisper model variants under cockpit-specific scenarios. Our study examines the role of transcript normalization, proposes aviation-aware normalization schemes for reliable WER computation, and investigates waveform-level data augmentation, including jittering, time warping, window slicing, window warping, magnitude warping, and scaling, to improve robustness under limited labeled data. The goal is to provide a practical account of adapting Whisper to intra-cockpit ASR and to establish a baseline for future speech recognition systems in this domain.

2 Methodology

2.1 Dataset

The dataset used in this study was collected from cockpit simulator recordings and pilot interview recordings. The cockpit recordings cover several operational situations



(a) Original waveform

(b) Augmented waveform

(c) Original spectrogram

(d) Augmented spectrogram

Fig. 1: Comparison of the original and window-warp augmented audio samples in the waveform and spectrogram domains. The effective augmentation was applied between 7 and 10 seconds of the audio, and the resulting spectral changes are highlighted by the red rectangle.

and were categorized into four scenario groups: takeoff scenarios, including the associated briefings and checklists; structured abnormal and emergency procedure execution via the Electronic Centralized Aircraft Monitor (ECAM) or the Quick Reference Handbook (QRH); decision-making procedures based on FOR-DEC, a standardized aviation decision-making model comprising Facts, Options, Risks and Benefits, Decision, Execution, and Check; and landing scenarios, including the associated briefings and checklists. The speech is multilingual, with German and English often appearing in the same recording and mixing both languages, which reflects the way German pilots communicate in operational settings.

The cockpit data contains approximately 180 minutes of conversations between pilots. The original cockpit videos were first converted into audio files. The audio signals were segmented into clips of up to 30 seconds. To avoid truncating spoken words, segmentation was performed via a script that identified points of silence to establish split boundaries. Finally, the segments were resampled to 16 kHz to align with the input requirements of the Whisper models. Manual transcripts were prepared for all recordings and used as ground-truth references for evaluation and fine-tuning. The resulting dataset was split into a training and test datasets with 80:20 split ratio.

2.2 Evaluation Metrics

We evaluate transcription accuracy using Word Error Rate (WER) and Character Error Rate (CER). WER is the standard metric for ASR evaluation and measures the edit distance between the predicted transcript and the reference transcript at the word level:

$$WER = \frac{S + D + I}{N}, \tag{1}$$

Table 1: WER(↓) and CER(↓) (%) of the pre-trained Whisper model family under different transcript normalization schemes on the held-out test set. Most normalization schemes substantially reduce transcription errors compared with evaluation without normalization.

	Tiny		Base		Small		Medium		Turbo		Large	
Normalizer	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER
No-norm	120.09	84.33	94.98	63.35	62.87	40.37	55.86	34.65	53.75	31.24	57.61	34.40
Basic	98.56	71.78	86.96	59.41	52.67	36.75	42.23	30.66	40.07	27.23	46.13	31.62
Number	119.92	84.01	95.00	63.29	62.91	40.42	55.91	34.68	53.75	31.29	57.59	34.47
English	97.85	71.16	86.98	59.32	52.55	36.83	42.05	30.63	39.95	27.27	45.87	31.69
Proposed I	98.73	73.01	86.96	60.43	52.48	36.95	41.98	30.49	40.35	27.23	46.05	31.56
Proposed II	98.89	72.39	87.13	60.32	52.43	36.97	41.91	30.49	40.25	27.46	45.29	31.76
Proposed III	98.41	72.28	86.83	60.23	52.26	36.91	41.67	30.37	40.11	27.20	45.65	31.60

where S denotes the number of substitutions, D the number of deletions, I the number of insertions, and N the total number of words in the reference transcript. A lower WER indicates better transcription performance. CER computation follows similar mechanism as WER, with a distinction of character substitutions, deletions, and insertions were counted in place of words.

2.3 Transcript Normalization

Post-transcription normalization is an important part of the evaluation pipeline, since small formatting differences may otherwise be counted as recognition errors. This is particularly relevant in cockpit conversations, where the same spoken content may appear in different written forms. For example, aviation expressions, numbers, compound words, and ICAO alphabet terms can lead to mismatches between the predicted and reference transcripts even when the intended transcription is largely correct.

We compare several normalization schemes. The basic normalizer lowercases text and removes special characters. The number normalizer converts spoken or written numeric forms into a common representation. The English normalizer combines text, spelling, and number normalization. In addition to these standard normalizers, we introduce a custom normalization scheme, referred to as Proposed I normalization. This scheme includes processing for punctuation, spelling, and number-related variants, and adds cockpit-specific transformations. In particular, it converts ICAO alphabet terms into their corresponding standard letters, for example converting “DELTA” into “D”. It also removes selected filler words and normalizes compound expressions so that forms such as “take-off”, “takeoff”, and “take off” are treated consistently. The novel Proposed II normalization applies Proposed I first, followed by the English normalizer. Proposed III applies the English normalizer first, followed by Proposed I.

2.4 Waveform-level Data Augmentation

Data augmentation is used to increase the effective size of the training set and improve robustness under limited labeled data. This is a common strategy in ASR, especially

when collecting and manually transcribing additional speech is expensive. Data augmentation methods for speech expose the model to small variations in timing, amplitude, frequency, and acoustic structure, while preserving the transcript label. In this work, we apply augmentation at the waveform level before Whisper feature extraction. The methods used in this paper differ from SpecAugment-style methods, which operate directly on log-mel spectrograms by masking or warping time and frequency regions [37]. An example of the window warp augmentation can be observed in Figure 1.

We apply six waveform-level augmentations before Whisper feature extraction, which are adapted techniques from the survey conducted by Iwana and Uchida [23], which evaluates data augmentation approaches for time-series data. A detailed description and illustrative examples of each of the six waveform-level augmentations can be found at the appendix.

- **Jittering:** Gaussian noise with $\sigma = 0.002$ is added to the waveform to mimic mild background noise and recording artifacts.
- **Scaling:** waveform amplitude is multiplied by a random factor with $\sigma = 0.1$, representing variations in speaker loudness, microphone gain, and recording distance.
- **Time warping:** the temporal axis is smoothly deformed using a cubic spline with $\sigma = 0.15$ and 4 knots, introducing local variations in speaking rate and pitch.
- **Magnitude warping:** the signal amplitude is smoothly varied over time using a cubic spline with $\sigma = 0.15$ and 4 knots, modelling gradual loudness changes within an utterance.
- **Window slicing:** a 90% temporal segment of the waveform is selected and resampled to the original length, introducing moderate temporal variation while preserving the transcript label.
- **Window warping:** a local window covering 8% of the waveform is stretched or compressed using scales from $\{0.7, 1.2, 1.5\}$, while the final signal length is kept fixed.

The augmented samples are used solely for training, while the test set remains unchanged, ensuring that the reported WER reflects generalization to the original cockpit speech distribution.

2.5 LoRA Fine-tuning

The fine-tuning stage adapts pre-trained Whisper models to the cockpit speech domain using the prepared audio-transcript pairs. We use the Hugging Face Transformers framework for model loading, feature extraction, tokenization, training, and evaluation. For each audio segment, log-mel spectrogram features are computed using the Whisper feature extractor, while the corresponding transcript is converted into token labels using the Whisper tokenizer. A data collator is used to pad the input features and token labels appropriately during batch training.

In the place of full model fine-tuning, we use Low-Rank Adaptation (LoRA) [21]. LoRA updates only a small set of additional trainable parameters (a.k.a. adapters) while keeping the pre-trained model parameters fixed. Post-training inference involves joint evaluation with both adapters and pre-trained model. This is useful in our setting because the available labeled cockpit speech data is limited, and full fine-tuning of large

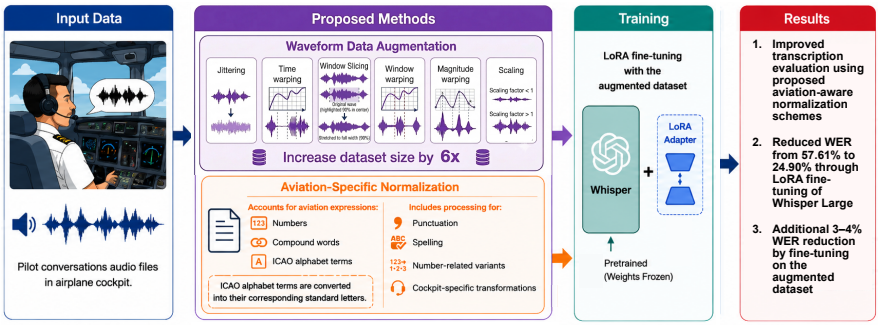


Fig. 2: Overview of the proposed aviation-domain ASR framework, combining waveform-level data augmentation, aviation-specific transcript normalization, and LoRA fine-tuning of Whisper. The proposed approach reduces WER from 57.61% to 24.90%, with augmented training providing a further 3–4% reduction.

Whisper models can be computationally expensive and more prone to overfitting. LoRA therefore provides a parameter-efficient method to adapt Whisper to cockpit-specific vocabulary, multilingual utterances, and acoustic conditions. Fine-tuning was performed on an NVIDIA Tesla V100 GPU. We tested learning rates from $\{1e^{-5}, 1e^{-4}, 1e^{-3}\}$ and availed early stopping criterion while finetuning to avoid overfitting and training divergence. Across the tested models, LoRA introduces only about 0.8% to 1.6% additional trainable parameters, while still allowing the model to adapt to the target domain.

3 Experimental Results

In this section, we report the transcription performance of Whisper models on cockpit and pilot speech recordings. Figure 2 summarizes the proposed aviation-domain ASR framework, which combines waveform-level data augmentation, aviation-specific transcript normalization, and LoRA fine-tuning of Whisper. The framework is evaluated against pre-trained Whisper models and existing normalization schemes[3].

3.1 Baseline Results

The audio files from the five scenarios were transcribed using the pre-trained multilingual Whisper models. The predictions were compared against the manually prepared reference transcripts using WER and CER and shown in the first row of Table 1. The results show that the smaller Whisper models are not sufficiently robust for this domain without adaptation. Whisper Tiny model obtain WER values exceeding 100% in some cases. This is not unusual in difficult ASR settings, since WER is not upper bounded by 100% when insertions or deletions dominate. The Small and Medium models provide a clear improvement over Tiny and Base, with average WER values in the range of

[3] This figure was generated with the assistance of generative AI tools and was subsequently reviewed, edited, and finalized by the authors.

Table 2: WER(↓) (%) of Whisper Large after LoRA fine-tuning with different waveform-level augmentation methods under each transcript normalization scheme. Best-performing method for each normalizer highlighted in bold.

	Without augmentation		LoRA fine-tuning with augmentation						
Normalizer	Pre-trained	Fine-tuned	All aug	Window warp	Jittering	Magnitude warp	Window slicing	Time warp	Scaling
No-norm	57.61	37.50	33.82	**33.44**	33.61	33.90	34.35	34.55	34.00
Basic	46.13	25.39	21.69	**21.45**	21.63	21.76	22.02	22.25	21.93
English	45.87	25.37	21.55	**21.29**	21.46	21.57	21.97	22.10	21.81
Proposed I	46.05	25.08	21.40	**21.14**	21.36	21.43	21.97	21.91	21.58
Proposed II	45.29	24.90	21.17	21.06	**20.96**	21.09	21.64	21.59	21.32
Proposed III	45.65	24.93	21.28	**21.03**	21.06	21.05	21.70	21.76	21.35

Table 3: WER(↓) (%) of Whisper Turbo after LoRA fine-tuning with different waveform-level augmentation methods under each transcript normalization scheme. Best-performing method for each normalizer highlighted in bold.

	Without augmentation		LoRA fine-tuning with augmentation						
Normalizer	Pre-trained	Fine-tuned	All aug	Window warp	Jittering	Magnitude warp	Window slicing	Time warp	Scaling
No-norm	53.75	39.72	36.05	34.34	34.22	34.60	**34.10**	37.45	48.62
Basic	40.07	26.86	23.56	**21.71**	22.19	22.21	21.89	24.89	35.45
English	39.95	26.78	23.52	**21.75**	22.09	22.18	21.95	24.83	35.30
Proposed I	40.35	26.87	23.52	**21.70**	22.12	22.18	21.85	24.54	38.40
Proposed II	40.25	26.46	23.22	**21.58**	22.07	22.04	21.81	24.29	38.35
Proposed III	40.11	26.60	23.39	**21.59**	22.01	21.98	21.74	24.38	38.18

55-63% without any normalization. Whisper Turbo achieves a mean WER of 53.75%, while Whisper Large-v3, referred to as Large in the following, obtains 57.61% WER. Although the larger models perform better, the baseline error rates remain high. This suggests that general-purpose Whisper models, even when trained on large-scale multilingual data, do not directly transfer well to intra-cockpit communication. The results motivate the need for domain adaptation using model finetuning and transcript normalization.

3.2 Effect of Transcript Normalization

Table 1 compares the normalization schemes considered in this work. The raw Whisper output, denoted as no-normalization, gives the highest WER. The basic normalization scheme alone improves the WER by 10-20%, showing the necessecity of normalization. The number normalization alone has very little effect on WER since it handles only numericals, which are not often used in conversations. The English normalizer gives the best performance among the standard normalization schemes. This indicates that spelling, number, and text-level normalization reduce a meaningful part of the measured error. The proposed normalization schemes Proposed II and Proposed III obtain the lowest WER for most Whisper models.

3.3 Effect of LoRA Fine-tuning and Normalization

The ASR results obtained after fine-tuning Whisper Large and Whisper Turbo are shown in Tables 2 and 3. Only a small fraction of the model parameters is updated

during LoRA fine-tuning. For Whisper Turbo, fine-tuning introduces $6,553,600$ trainable parameters, corresponding to approximately 0.8% of the total $815,431,680$ parameters. Despite this limited number of trainable parameters, the WER decreases from 53.75% to 39.72% without transcript normalization. For Whisper Large, fine-tuning reduces the WER from 57.61% to 37.50%. Normalization and fine-tuning are individually useful, but their combination gives the largest improvement. For the pre-trained Whisper Large model, WER decreases from 57.61% without normalization to 24.90% with Proposed II normalization. A similar trend is observed for Whisper Turbo. The WER decreases from 53.75% for the pre-trained model without normalization to 26.46% after LoRA fine-tuning with Proposed II normalization. The best results are obtained when the model is first adapted to the cockpit speech with finetuning and the resulting transcripts are then evaluated using a normalization scheme that reflects the conventions of cockpit communication.

3.4 Effect of Waveform-level Data Augmentation

The augmented training dataset is created using transformations such as jittering, scaling, magnitude warping, time warping, window slicing, and window warping. The test set is kept unchanged, so the reported improvements reflect generalization to the original cockpit speech distribution. Across the augmentation experiments, LoRA fine-tuning with augmented data gives reduction in WER in most cases compared with the corresponding fine-tuned baselines. In Table 2, with Whisper Large models, the best configuration reduces WER from 24.90% to 20.96% with Proposed II normalization, corresponding to a WER reduction of 3.94% using jittering data augmentation method. In the case of Whisper Turbo Model, Table 3, the WER decreases from 26.46% to 21.58% with Proposed II normalization, showcasing 4.88% WER reduction with window warp augmentation method.

Among the different data augmentation methods employed, Window warping gives the best or near-best performance for most normalization schemes, while jittering gives the lowest WER for Proposed II normalization for Whisper Large model. In contrast, window slicing, time warping, and scaling are slightly less effective, but still provide clear improvements over the non-augmented fine-tuned baseline. For Whisper Turbo, window warping provides the best WER for most normalizers, while window slicing gives the best result with raw transcripts. Jittering and magnitude warping also yield competitive performance, whereas time warping gives smaller gains. Scaling is notably less stable for Whisper Turbo and increases the WER substantially compared with the non-augmented fine-tuned Whisper Turbo model.

3.5 Error Analysis: Multilingual Speech and Phonetic Confusions

In addition to aggregate WER values, we analyze representative transcription errors to better understand the limitations of Whisper language models for cockpit ASR. Two recurring error types are observed: unintended translation in multilingual speech and substitutions caused by close phonetic similarity. Table 4 shows examples where the model translates German or mixed German-English utterances into English rather than preserving the original spoken form. For example, *Gut* is transcribed as *Good*, and

Table 4: Examples of unintended translation in ASR output. The reference contains German or mixed German-English speech, while the prediction is partially or fully translated into English.

Reference	**Prediction**
Ist confirmed	**That was confirmed**
Gut	**Good**
Blaues system ist natürlich verloren	**The blue system is of course lost**
daraufhin ein spoiler-pair	**then a spoiler pair**

Blaues system ist natürlich verloren is rendered as *The blue system is of course lost*. While such outputs may preserve part of the semantic meaning, they are still incorrect for verbatim ASR evaluation. Table 5 presents cases where the predicted text is phonetically close to the reference but still incorrect. Examples include *read status* being recognized as *wave status*, *slats low* as *sled low*, and *CAT3 single* as *cut three single*. These errors suggest that the model often captures the acoustic pattern of the utterance but fails to map it to the correct aviation term.

The effect of fine-tuning on transcription accuracy is further illustrated through the qualitative example presented below. The pre-trained Whisper Turbo failed to transcribe the utterance from the second pilot, who said *"ja habe ich die auch gemacht landing performance moment ich muss nochmal kurz suchen wo ich es aufgeschrieben habe da unten"*, resulting in a WER of 41%. In contrast, the fine-tuned Whisper Turbo was able to transcribe the complete utterance, resulting in near-perfect transcription accuracy.

Reference: *"ah landing performance fur zurich hast du wahrscheinlich auch schon durch das sollte kein problem sein mit dem gewicht ja habe ich die auch gemacht landing performance moment ich muss nochmal kurz suchen wo ich es aufgeschrieben habe da unten so wir haben also mit flaps 3 71 tonnen und airport elevation 1500 haben wir eine landing distance"*

o **Pre-trained Whisper Turbo Prediction:** *"landing performance fur zurich hast du wahrscheinlich auch schon durch das sollte kein problem sein mit dem gewicht so wir haben also mit flaps tri 71 tonnen und airport elevation 1500 haben wir einen landing distance"*

Finetuned Whisper Turbo Prediction: *"ah landing performance fur zurich hast du wahrscheinlich auch schon durch das sollte kein problem sein mit dem gewicht ja habe ich die auch gemacht landing performance moment ich muss nochmal kurz suchen wo ich es aufgeschrieben habe da unten so wir haben also mit flaps 3 71 tonnen und airport elevation 1500 haben wir eine landing distance"*

Table 5: Examples of transcription errors caused by close phonetic similarity between the reference and prediction.

Reference	Prediction
clear flight control	**okay, flight control**
clear flight control	**flight control**
read status	**wave status**
slats low	**sled low**
CAT3 single	**cut three single**
Inop systems	**In-hub systems**

3.6 LLM-based ASR Error Post-correction

We also explored whether the errors discussed in Section 3.5 could be reduced using LLM-based ASR post-correction, motivated by recent work on generative correction of ASR outputs [29,28]. We tested prompt-based correction of Whisper hypotheses, prompts with aviation-specific context and vocabulary, and translation-based post-processing where both hypotheses and references were translated into a common language before computing WER. These approaches did not consistently improve performance over the best normalization-based setting. In several cases, the LLM produced more fluent text, but also paraphrased the utterance, translated multilingual speech, expanded abbreviations, or replaced cockpit-specific terms with more common expressions.

4 Discussion

Pre-trained Whisper models require domain adaptation for cockpit speech. Although Whisper models are trained on large-scale multilingual data, the baseline results show high WER on cockpit communication, indicating that cockpit ASR is not only a general multilingual transcription problem, but also a domain adaptation problem involving procedural speech, aviation vocabulary, multilingual switching, and challenging acoustic conditions.

Normalization is not merely a cosmetic post-processing step. The difference between raw and normalized WER is large across several configurations. This shows that transcript formatting decisions can strongly affect measured ASR performance. For cockpit speech, normalization plays cruicial role in computation of WER, since the same spoken content may be written in several acceptable forms. ICAO alphabet terms, compound words, filler words, and numeric expressions need to be handled consistently to avoid counting formatting differences as recognition errors.

LoRA fine-tuning provides an efficient adaptation route. The fine-tuning results show that training an additional 0.8–1.5% percent of the model parameters using a small high quality labeled cockpit conversations can still provide meaningful WER reductions, showcasing the advantages in small domain-specific ASR datasets, where full fine-tuning may be computationally expensive and more prone to overfitting.

Fine-tuning and normalization have complementary effects. The strongest improvements are obtained when LoRA fine-tuning is combined with an appropriate normalization scheme. Fine-tuning improves the raw speech recognition capabilities of the Whisper model, while normalization reduces avoidable mismatches in evaluation. Finetuned model produces more domain-relevant transcripts and the text-level inconsistencies are corrected via systematic normalization.

Waveform augmentation improves robustness under limited data. Training on waveform augmented dataset gives an additional WER reduction of about 3–4%. Testing on held-out original and un-augmented test dataset shows that, the trascription accuracy gains indicate improved robustness to the natural variability of cockpit speech. The results are consistent with the broader ASR studies stating that augmentation is particularly useful when labeled speech is limited.

4.1 Future Work

Though LoRA fine-tuning, transcript normalization, and waveform-level data augmentation to expand training dataset substantially improve ASR performance, the transcription of cockpit conversations needs further refinements to enable deployment in AI-assistants for cockpit, improved analysis of communication between pilots, etc. An important direction to consider is to expand the training data, since the current dataset is limited in duration and speaker diversity. Additional recordings with more speakers, different accents, more acoustic conditions, and a wider range of cockpit procedures would likely improve domain adaptation. More systematic augmentation policies, in spectral-domain, can be considered for an additional degree of expanding the training dataset. The evaluation protocol can be improved with emphasis on cockpit conversations. Context-aware WER variants, semantic similarity measures, or task-specific scoring methods that better reflect whether the intended operational meaning has been preserved could make more impact on evaluation.

5 Conclusion

In this paper, the transcription of cockpit conversations using Whisper models was explored. To reduce unintended mismatches between reference and predicted transcripts, domain-specific post-processing protocols were introduced to address ICAO alphabet standardization, compound-word segmentation, and the removal of non-lexical fillers. Fine-tuning the Whisper models using Low-Rank Adaptation, combined with the proposed normalization schemes, reduced the WER from 53.75% for the pre-trained Whisper Turbo model without normalization to 24.46% for the fine-tuned Whisper Turbo model using the Proposed II normalization scheme on the test dataset. Fine-tuning on the waveform-augmented dataset provided an additional improvement, reducing the normalized WER from 24.46% to 21.58% for Whisper Turbo. Future work will focus on expanding the dataset and developing more robust evaluation frameworks.

Appendix

Waveform Data Augmentation

This section presents the selected augmentation techniques for time-series waveform data. The chosen techniques are drawn from the survey conducted by Iwana and Uchida [23], which evaluates data augmentation approaches for time-series data. The focus here is on transformation-based methods that operate independently of complex assumptions, such as the similarity scores. Illustrative examples of each technique are provided at the end of the section.

Jittering Jittering is an augmentation technique that introduces Gaussian noise into the time-series data. The degree of noise added is controlled by an adjustable hyperparameter, specifically the standard deviation of the Gaussian distribution. By modifying this parameter, the intensity of perturbations applied to the original time-series can be adjusted.

Scaling The scaling operation alters the magnitude of a time-series. A random scaling factor is drawn from a Gaussian distribution with a mean of one. To apply this transformation, the mean of the time-series is first subtracted from all values, ensuring that scaling is applied relative to a zero-centered baseline. The data is then multiplied by the scaling factor, and the original mean is subsequently added back. If the scaling factor is less than one, values above the mean decrease while values below the mean increase, effectively compressing the range. Conversely, when the scaling factor is greater than one, the variation in magnitude is amplified. The primary hyperparameter for this transformation is the standard deviation of the Gaussian distribution used to sample the scaling factor.

Magnitude Warping Magnitude warping is a transformation that modifies the amplitude of a time-series using a cubic spline. This spline is generated by randomly selecting a set number of knots and connecting them smoothly. The original time-series is then multiplied by the spline, causing variations in magnitude. When the spline values are smaller than one, the augmented time-series exhibits a reduced magnitude compared to the original. Conversely, when the spline values exceed one, the magnitude is increased. The key hyperparameters for this transformation include the number of knots and the standard deviation of the Gaussian distribution from which the knot values are drawn.

Window Slice Window slicing is a transformation that alters the time axis of a time-series by selecting a consecutive segment and rescaling it to the original length. A window of a specified percentage of the original time-series is selected, with its starting point chosen at random. This extracted segment is then interpolated back to the full length of the original time-series, effectively modifying the temporal structure while preserving the overall trend. The key hyperparameter for this augmentation is the window size percentage, which is typically set to 90%.

Window Warping A consecutive window of a specified percentage is extracted from the time-series and then either stretched by a factor of two along the time axis or compressed by a factor of 0.5. The remaining parts of the time-series are adjusted using interpolation to ensure that the augmented sequence retains the same overall length as the original. The primary hyperparameter for this transformation is the window size, which we set to 8%.

Time Warping Time warping is a transformation that modifies the temporal structure of a time-series without altering the magnitude of its values. Instead of adjusting amplitude, this technique changes the time steps at which values appear. Similar to magnitude warping, a cubic spline is generated using a predefined number of knots, which are randomly sampled from a Gaussian distribution. In regions where the spline values are below one, the time axis is stretched, causing data points to be spaced further apart. Conversely, in areas where the spline values exceed one, the time axis is compressed, resulting in a denser arrangement of data points. The primary hyperparameters for this transformation are the number of knots and the standard deviation of the Gaussian distribution from which the knot values are drawn.

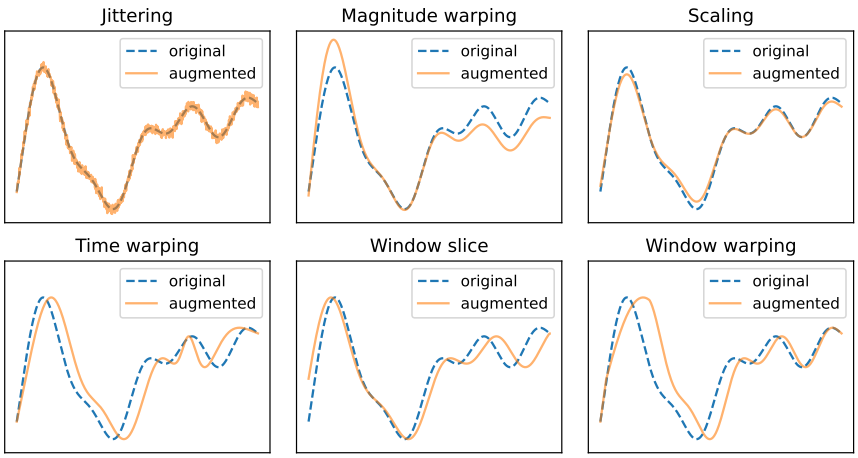


Fig. 3: Time series wave form augmentations on example data

References

1. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., et al.: Deep speech 2 : End-to-end speech recognition in english and mandarin. In: Proceedings of ICML. vol. 48, pp. 173–182 (2016)
2. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G.: Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670 (2019)

3. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems. vol. 33, pp. 12449–12460 (2020)
4. Bayes, T., Price, N.: An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London **53**, 370–418 (1763). https://doi.org/10.1098/rstl.1763.0053
5. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of machine learning research **3**, 1137–1155 (2003)
6. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C.: Automatic speech recognition and speech variability: A review. Speech communication **49**(10-11), 763–786 (2007)
7. Chan, W., Jaitly, N., Le, Q.V., Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: Proceedings of ICASSP. pp. 4960–4964 (2016). https://doi.org/10.1109/ICASSP.2016.7472621
8. Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q., Norouzi, M.: Speechstew: Simply mix all available speech recognition data to train one large neural network. arXiv preprint arXiv:2104.02133 (2021)
9. Chang, J., Nam, H., et al.: Exploring the feasibility of fine-tuning large-scale speech recognition models for domain-specific applications: A case study on whisper model and kspon-speech dataset. Phonetics and Speech Sciences **15**(3), 83–88 (2023)
10. Chen, G., Chai, S., Wang, G., Du, J., Zhang, W., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J.: Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. arXiv preprint arXiv:2106.06909 (2021)
11. Das, S.S., Chan, N., Wages, D., Hansen, J.H.: Application of automatic speech recognition in call classification. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 4, pp. IV–3896. IEEE (2002)
12. van Doorn, J., Sun, J., Hoekstra, J., Jonk, P., de Vries, V.: Whisper-ATC: Open models for air traffic control automatic speech recognition with accuracy. In: International Conference on Research in Air Transportation (2024)
13. EUROCONTROL: European action plan for air ground communications safety. Tech. rep., EUROCONTROL (May 2006), https://www.eurocontrol.int/publication/european-action-plan-air-ground-communications-safety
14. Federal Aviation Administration: Aeronautical information manual: Section 4. atc clearances and aircraft separation (2026)
15. Gales, M., Young, S.: The application of hidden markov models in speech recognition. Foundations and Trends in Signal Processing **1**(3), 195–304 (2008). https://doi.org/10.1561/2000000004
16. Galvez, D., Diamos, G., Ciro, J., Cerón, J.F., Achorn, K., Gopi, A., Kanter, D., Lam, M., Mazumder, M., Reddi, V.J.: The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. arXiv preprint arXiv:2111.09344 (2021)
17. Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
18. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., Ng, A.: Deep Speech: Scaling up end-to-end speech recognition. arXiv (2014)
19. Hinton, G., Deng, L., Yu, D., Dahl, G., r. Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine **29**(6), 82–97 (2012). https://doi.org/10.1109/MSP.2012.2205597

20. Hofbauer, K., Petrik, S., Hering, H.: The ATCOSIM corpus of non-prompted clean air traffic control speech. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08) (2008)
21. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022), https://arxiv.org/abs/2106.09685
22. International Civil Aviation Organization: Procedures for Air Navigation Services: Air Traffic Management. Montreal, Canada, 16th edn. (2016)
23. Iwana, B.K., Uchida, S.: An empirical survey of data augmentation for time series classification with neural networks. Plos one **16**(7) (2021)
24. Jaitly, N., Hinton, G.: Vocal Tract Length Perturbation (VTLP) improves speech recognition. In: ICML Workshop on Deep Learning for Audio, Speech and Language Processing (2013)
25. Kanda, N., Takeda, R., Obuchi, Y.: Elastic spectral distortion for low resource speech recognition with deep neural networks. In: ASRU (2013)
26. Kim, C., Misra, A., Chin, K., Hughes, T., Narayanan, A., Sainath, T., Bacchiani, M.: Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home. In: INTERSPEECH (2017)
27. Ko, T., Peddinti, V., Povey, D., Khudanpur, S.: Audio Augmentation for Speech Recognition. In: INTERSPEECH (2015)
28. Li, S., Chen, C., Kwok, C.Y., Chu, C., Chng, E.S., Kawai, H.: Investigating ASR Error Correction with Large Language Model and Multilingual 1-best Hypotheses. In: Interspeech 2024. pp. 1315–1319 (2024). https://doi.org/10.21437/Interspeech.2024-368
29. Ma, R., Qian, M., Manakul, P., Gales, M., Knill, K.: Can generative large language models perform asr error correction? arXiv preprint arXiv:2307.04172 (2023)
30. Mallidi, S.H., Hermansky, H.: Novel neural network based fusion for Multistream ASR. In: ICASSP (2016)
31. Matarneh, R., Maksymova, S., Lyashenko, V., Belova, N.: Speech recognition systems: A comparative review (2017)
32. Meng, L., Xu, J., Tan, X., Wang, J., Qin, T., Xu, B.: MixSpeech: Data Augmentation for Low-Resource Automatic Speech Recognition. In: Proceedings of ICASSP. pp. 7008–7012 (2021)
33. Mohri, M., Pereira, F., Riley, M.: Weighted finite-state transducers in speech recognition. Computer Speech & Language **16**(1), 69–88 (2002). https://doi.org/10.1006/csla.2001.0184
34. Nareddy, K.K.R., Ternus, S., Niebling, J.: Analyzing and fine-tuning whisper models for multilingual pilot speech transcription in the cockpit. arXiv preprint arXiv:2506.21990 (2025), https://arxiv.org/abs/2506.21990
35. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In: In Proceedings of ICASSP. pp. 5206–5210. IEEE (2015)
36. Papenfuss, A., Schmidt, C.A.: Using automatic speech recognition to evaluate team processes in aviation - first experiences and open questions. In: Engineering Psychology and Cognitive Ergonomics. p. 501–513 (2023). https://doi.org/10.1007/978-3-031-35389-5_34, https://doi.org/10.1007/978-3-031-35389-5_34
37. Park, D.S., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E.D., Le, Q.V.: Specaugment: A simple data augmentation method for automatic speech recognition. In: Proceedings of Interspeech. pp. 2613–2617 (2019). https://doi.org/10.21437/Interspeech.2019-2680
38. Pellegrini, T., Farinas, J., Delpech, E., Lancelot, F.: The Airbus air traffic control speech recognition 2018 challenge: Towards ATC automatic transcription and call sign detection. arXiv preprint arXiv:1810.12614 (2018)
39. Prasad, A., Zuluaga-Gomez, J., Motlicek, P., Sarfjoo, S.S., Nigmatulina, I., Ohneiser, O., Helmke, H.: Grammar based speaker role identification for air traffic control speech recognition. arXiv preprint arXiv:2108.12175 (2021)

40. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)
41. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International conference on machine learning. pp. 28492–28518. PMLR (2023)
42. Ragni, A., Knill, K.M., Rath, S.P., Gales, M.J.F.: Data augmentation for low resource languages. In: INTERSPEECH (2014)
43. Raju, A., Panchapagesan, S., Liu, X., Mandal, A., Strom, N.: Data Augmentation for Robust Keyword Spotting under Playback Interference. arXiv (2018)
44. Rataj, J., Helmke, H., Ohneiser, O.: AcListant with continuous learning: Speech recognition in air traffic control. In: Air Traffic Management and Systems IV, pp. 93–112. Springer (2021). https://doi.org/10.1007/978-981-33-4669-7_6
45. Rijal, S., Adhikari, S., Dahal, M., Awale, M., Ojha, V.: Whisper finetuning on nepali language (2023), preprint
46. Roushan, R., Mishra, H., Yadav, L., Koppula, S., Tiwari, N., Nataraj, K.: Optimizing speech recognition for medical transcription: fine-tuning whisper and developing a web application. In: 2024 IEEE Conference on Engineering Informatics (ICEI). pp. 1–6. IEEE (2024)
47. Schmidt, C., Stadtschnitzer, M., Koehler, J.: The Fraunhofer IAIS audio mining system: Current state and future directions. In: Speech Communication; 12. ITG Symposium. pp. 1–5 (2016)
48. Ternus, S., Nareddy, K.K.R., Niebling, J., Papenfuss, A.: Automatic speech recognition in the cockpit: A comparative study of asr models for pilot communication. In: Deutscher Luft- und Raumfahrtkongress 2025. Deutsche Gesellschaft für Luft- und Raumfahrt – Lilienthal-Oberth e.V., Bonn (2025). https://doi.org/10.25967/650258, https://doi.org/10.25967/650258
49. Williams, J.J., Borge, M., Levin, J., Guo, H., Rosé, C.P.: A comparative analysis of automatic speech recognition errors in small group classroom discourse. Proceedings of the ACM on Human-Computer Interaction **7**(CSCW2), 1–27 (2023). https://doi.org/10.1145/3565472.3595606
50. Zhang, Y., Park, D.S., Han, W., Qin, J., Gulati, A., Shor, J., Jansen, A., Xu, Y., Huang, Y., Wang, S.: Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. IEEE Journal of Selected Topics in Signal Processing **16**(6), 1519–1532 (2022)
51. Zuluaga-Gomez, J., Veselý, K., Szöke, I., Blatt, et al.: ATCO2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications. arXiv preprint arXiv:2211.04054 (2022)
52. Zuluaga-Gomez, J., Motlicek, P., Zhan, Q., Veselý, K., Braun, R.: Automatic speech recognition benchmark for air-traffic communications. arXiv preprint arXiv:2006.10304 (2020)