Analyzing and Fine-Tuning Whisper Models for Multilingual Pilot Speech Transcription in the Cockpit

Kartheek Kumar Reddy Nareddy
Institute of Data Science
German Aerospace Center
kartheek.nareddy@dlr.de

Sarah Ternus
Institute of Flight Guidance
German Aerospace Center
sarah.ternus@dlr.de

Julia Niebling
Institute of Data Science
German Aerospace Center
Julia.Niebling@dlr.de

Abstract

The developments in transformer encoder-decoder architectures have led to significant breakthroughs in machine translation, Automatic Speech Recognition (ASR), and instruction-based chat machines, among other applications. The pre-trained models were trained on vast amounts of generic data over a few epochs (fewer than five in most cases), resulting in their strong generalization capabilities. Nevertheless, the performance of these models does suffer when applied to niche domains like transcribing pilot speech in the cockpit, which involves a lot of specific vocabulary and multilingual conversations. This paper investigates and improves the transcription accuracy of cockpit conversations with Whisper models. We have collected around 85 minutes of cockpit simulator recordings and 130 minutes of interview recordings with pilots and manually labeled them. The speakers are middle aged men speaking both German and English. To improve the accuracy of transcriptions, we propose multiple normalization schemes to refine the transcripts and improve Word Error Rate (WER). We then employ fine-tuning to enhance ASR performance, utilizing performance-efficient fine-tuning with Low-Rank Adaptation (LoRA). Hereby, WER decreased from 68.49 % (pretrained whisper Large model without normalization baseline) to 26.26% (finetuned whisper Large model with the proposed normalization scheme).

1. Introduction

Automatic Speech Recognition (ASR), transforming audio signals into text, plays a key role in natural language processing [6]. The diversity in speech signals with variations like gender, accent, pace, external noise, etc. makes ASR a challenging problem [6]. ASR has found applications in automatic call handling [10], and personalized AI assistants [15]. Conventional ASR systems rely on a pipeline of components, including acoustic feature extraction, acoustic and language modeling, and decoding via Bayes' decision rule [4]. With the advent of deep learning, both acoustic and language modeling have been revolutionized [5], ultimately leading to end-to-end models [9].

Publicly available datasets like LibriSpeech [16], Common Voice [2], and SpeechStew [7] contributed towards training and testing newly upcoming ASR models. However, the increasing size of neural networks has outpaced the size of these labeled datasets, often resulting in overfitting and poor generalization [12]. This challenge has motivated the creation of large-scale unlabeled or weakly labeled datasets, such as BigSSL (1 million hours) [22], GigaSpeech [8], and People's Speech [11]. Among contemporary ASR models, OpenAI's Whisper stands out for its large-scale, weakly supervised training across 680,000 hours of multilingual data, incorporating both supervised and unsupervised techniques to achieve broad generalizability across diverse domains and languages [20].

While Whisper and other transformer-based models (e.g., Wav2Vec [3], SpeechStew[7], DeepSpeech[1]) perform impressively on general speech data, their accuracy can degrade in domain-specific contexts [21]. Fine-tuning, the process of adapting a pre-trained model to a specific task or dataset, has emerged as a powerful method to enhance ASR performance under such conditions [14]. Fine-tuning has proven effective across diverse application areas, including healthcare and low-resource languages. For example, adapting Whisper for Nepali speech led to substantial reductions in Word Error Rate (WER), with improvements up to 36.2% on the small model [18].

In aviation, ASR has also been widely explored for Air Traffic Control (ATC) communication. Domain-specific ASR models like Whisper-ATC have achieved as low results as 1.17% WER on ATCOSIM simulated data and up to 60% improvement through regional fine-tuning [20]. However, ASR for intra-cockpit communication between pilots and the fine-tuning of ASR models for that use-case remains relatively unexplored. Accurate transcription in this context can support human factors research, assess teamwork dy-

namics, and lay the groundwork for speech-driven cockpit automation systems. Studies in the past considered hidden markov models based transcription technologies aiming to transcribe cockpit conversations [17, 19]. Yet, the cockpit environment poses unique challenges, including overlapping speech, multilingual exchanges, high noise levels and a lot of use-case specific vocabulary.

To tackle these challenges, we explore the fine-tuning of Whisper models for multilingual pilot communication in the cockpit. Thereby, we adopt the Hugging Face fine-tuning pipeline. Our contributions aim to give an overview over fine-tuning Whisper for this domain, analyze model performance across different Whisper models and scenarios, proposing new normalization schemes, and establish groundwork for future ASR applications in the cockpit.

2. Methodology

2.1. Dataset

The dataset consists of 85 minutes of cockpit simulator recordings and 130 minutes of pilot interviews. The recordings cover various cockpit communication scenarios, including checklists, briefings, and emergency procedures, and reflect typical cockpit vocabulary. The audio is multilingual, with a mix of German and English as commonly spoken by German pilots. All audio was converted to MP3, segmented into 30-second clips, and resampled to 16 kHz. Manual transcripts were created as ground-truth references. The speakers are middle-aged male pilots.

2.2. Metrics

To evaluate transcription accuracy, WER was used as the primary metric. WER was computed using the jiwer library[1], which provides a standardized implementation for text-based error measurement. WER is a common metric in speech recognition and is defined as, WER = $\frac{S+D+I}{N}$, 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.

2.3. Transcript Normalization

In this paper, various normalization schemes are being compared. First, three normalization steps from Whisper were applied: Basic normalization, which includes case lowering and the removal of special characters; Number normalization, which converts numeric expressions into Arabic numerals; and the English normalizer, which combines text, number, and spelling normalization.

In addition to these, we introduce Proposed I, a custom normalization function incorporating similar processing for numbers, spelling, and punctuation with additional functions for transforming the ICAO-alphabet into standard letters (e.g., "DELTA" into "D") , removing filler words, and normalizing compound words (e.g., ensuring "take-off," "takeoff," and "take off" are treated as equivalent). Lastly, we evaluated two combined approaches: Proposed II, which applies Proposed I first, followed by English normalizer, and Proposed III, which applies English normalizer first, followed by Proposed I. Throughout the combined normalization approaches, the spelling and number normalizers are solely taken from the Whisper English normalizer.

[1] https://jitsi.github.io/jiwer/

2.4. Finetuning

The fine-tuning step of this study was conducted using the audio files and transcriptions from Section 2.1. For fine-tuning, the dataset was divided into a training set consisting of 158 audio files and a test set containing 40 audio files. Furthermore, the HuggingFace Transformers Python package was utilized to handle the fine-tuning procedure. Labels were extracted using the Whisper tokenizer. The log-mel spectrogram was computed using the feature extractor and processed as features. The data collator was used to ensure that the length of the features matched that of the input tokens. We used the LoRA [13] fine-tuning method, with learnable parameters amounting to approximately 1% of the model's total parameters. Fine-tuning of the Whisper models was performed using an NVIDIA Tesla V100. Hereby, multiple learning rates from {1e-5, 1e-4, 1e-3} were tested.

3. Results & Discussion

3.1. Baseline Results

We transcribed the audio files across the five scenarios using the family of multilingual whisper models. Then we computed the WER between predictions and the reference transcriptions. The results are shown in Figure 1. Whisper Tiny and Base models have WERs exceeding 100% in a few cases, indicating notable transcription errors. The Small and Medium Whisper models have considerable performance improvement over Tiny and Base, with WER in the 75-85% range on average. Whisper Turbo has 73.92% mean WER and Large-v3 (henceforth called as Large) has 66.44% WER, indicating the necessity for fine-tuning and text normalization.

3.2. Effect of Normalization

Table 1 shows a comparison of normalization schemes considered in this paper. The raw text predicted by the Whisper model is noted as no-norm(alization) text and has the highest WER. Basic text normalizer does have decent performance improvement over no-norm, while Number normalizer is falling behind the basic text normalizer. English normalizer on the other hand has best performance among the baseline normalizers considered. Three normalization

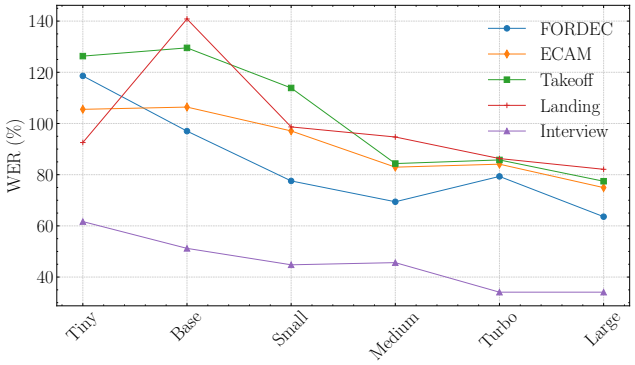

Figure 1. Baseline Word Error Rate (WER) comparison across family of Whisper models for various pilot speech scenarios.

Table 1. WER (in %) comparison of proposed normalization methods against baselines across family of Whisper models.

Normalizer	Tiny	Base	Small	Medium	Turbo	Large
No-norm	94.41	96.00	85.64	81.64	70.20	68.49
Basic	91.73	84.70	69.35	62.96	49.49	52.23
Number	89.13	88.70	77.41	70.84	62.18	59.76
English	88.37	84.58	69.16	62.43	48.88	52.08
Proposed I	85.68	83.10	69.05	63.19	49.68	52.74
Proposed II	88.41	84.25	69.87	62.60	48.69	52.00
Proposed III	88.21	82.96	68.76	62.54	48.70	52.41

schemes are presented in this paper, namely, Proposed I, II, and III. Among these, Proposed II and Proposed III have the lowest WERs for most of the Whisper models. This shows that normalizing ICAO-alphabets and removing filler words in combination with English normalizer results in the best performance.

3.3. Effect of Fine-Tuning

Fine-tuning the Whisper language models further resulted in improved transcription accuracy. The results of LoRA fine-tuning, with approximately 1% of learnable parameters, on Whisper Large and Turbo models are given in Tables 10 and 11. The fine-tuning results of the remaining Whisper models and relevant Python scripts can be found in the supplementary material. The Whisper Turbo model, when fine-tuned with $6,553,600$ parameters, representing 0.8% of its $815,431,680$ total parameters, achieved a WER reduction from 70.20% to 61.82% without any normalization. Similarly, the Whisper Large model showed a drop in WER from 68.49% to 55.65% with fine-tuning alone, depending on the learning rate. We experimented with different learning rates, and the optimal values varied between models. Whisper Turbo gave the best results at a learning rate of 1e-5, while Whisper Large models performed best with a 1e-3 learning rate.

3.4. Combining Normalization and Fine-Tuning

As shown in Section 3.2, normalization alone provides a notable improvement in transcription accuracy. However,

Table 2. LoRA fine-tuning on Whisper Large model with various learning rates. The numbers indicate WER in %.

Normalizer	pre-trained	lr=1e-5	lr=1e-4	lr=1e-3
No-norm	68.49	58.83	64.36	55.65
Basic	52.23	27.96	37.09	27.37
Number	59.76	46.10	48.30	50.08
English	52.08	27.80	36.71	26.36
Proposed I	52.74	32.72	37.35	38.37
Proposed II	52.00	27.65	36.41	26.26
Proposed III	52.41	28.24	36.60	27.00

Table 3. LoRA fine-tuning on Whisper Turbo model with various learning rates. The numbers indicate WER in %.

Normalizer	pre-trained	lr=1e-5	lr=1e-4	lr=1e-3
No-norm	70.20	61.82	64.67	65.06
Basic	49.49	28.18	29.04	31.02
Number	62.18	43.08	46.64	47.61
English	48.88	28.01	28.81	30.40
Proposed I	49.68	29.17	29.98	31.70
Proposed II	48.69	28.24	28.88	30.32
Proposed III	48.71	28.40	28.67	30.55

an interesting finding is that its effect becomes even more pronounced when applied after fine-tuning. For example, the Whisper Large model had a WER of 68.49% without normalization, which decreased to 52.00% when the proposed II normalizer was applied to the pre-trained model. After fine-tuning (with a learning rate of 1e-3), the model's WER without normalization was 55.65%, and further decreased to 26.26% when combined with the same normalization method. This demonstrates that while normalization improves performance on its own, its impact is more pronounced after the model has been fine-tuned. A similar trend was observed for the Whisper Turbo model, where the WER dropped from 70.20% (pre-trained, no normalization) to 61.82% after fine-tuning with a 1e-5 learning rate. When English normalization was applied, the WER further reduced to 28.01%. These results suggest that combining normalization with fine-tuning can yield greater improvements than using either approach independently.

4. Future Work

Though fine-tuning enhanced the ASR performance, the reported 26% WER in Section 3.3 is not suitable for reliable deployment. One promising direction is to utilize prompting to provide context and aid in recognition of domain-specific vocabulary. Further, we aim to generate more data for improving the fine-tuning performance. The WER computation does play a crucial role in determining the suitability of the language model for transcription. A context-based WER computation that overlooks minor grammatical varia-

tions typical of spoken language could provide a more accurate reflection of ASR model performance. Therefore, further improvements in normalization, as well as methods to assess whether the transcribed content conveys the intended meaning, should be considered.

5. Conclusion

In this paper the transcription of cockpit conversations using Whisper language models was explored. The audio files contain conversations between pilots in both German and English languages. The Whisper models transcribed the conversations with a high WER, which necessitates normalization and fine-tuning. Thereby, whisper normalization was utilized and own normalization schemes to normalize ICAO-alphabet, compound words and remove filler words were introduced, which resulted in better performance. Finetuning with Low-rank adaptation combined with normalization resulted in reduction of WER from 70.20% (pretrained Whisper Turbo model without normalization baseline) to 28.01% (fine-tuned Whisper Turbo model with the Proposed II normalization scheme) on the test dataset. The results emphasize the importance of domain adaptation for ASR models, particularly with technical vocabulary, multilingual speech, etc. Future work could include the exploration of prompting strategies, the creation of more training data, and more effective error computation approaches to further enhance the performance.

References

[1] D. Amodei et al. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 173–182. PMLR, 2016. 1

[2] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019. 1

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, pages 12449–12460, 2020. 1

[4] T. Bayes and N. Price. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. 1

[5] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003. 1

[6] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, and C. Ris. Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11):763–786, 2007. 1

[7] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi. Speechstew: Simply mix all available speech recognition data to train one large neural network. *arXiv preprint arXiv:2104.02133*, 2021. 1

[8] G. Chen, S. Chai, G. Wang, J. Du, W. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, and J. Zhang. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021. 1

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. 2011. 1

[10] S. S. Das, N. Chan, D. Wages, and J. H. Hansen. Application of automatic speech recognition in call classification. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages IV–3896. IEEE, 2002. 1

[11] D. Galvez, G. Diamos, J. Ciro, J. F. Cerón, K. Achorn, A. Gopi, D. Kanter, M. Lam, M. Mazumder, and V. J. Reddi. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *arXiv preprint arXiv:2111.09344*, 2021. 1

[12] R. Geirhos, J. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, Lu Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[14] et al. Liao, Y. Prompt-conditioning fine-tuning for domain-specific speech recognition. *arXiv preprint arXiv:2307.10274*, 2023. 1

[15] R. Matarneh, S. Maksymova, V. Lyashenko, and N. Belova. Speech recognition systems: A comparative review. 2017. 1

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing*, pages 5206–5210. IEEE, 2015. 1

[17] A. Papenfuss and C. A. Schmidt. Using automatic speech recognition to evaluate team processes in aviation - first experiences and open questions. In *Engineering Psychology and Cognitive Ergonomics*, page 501–513, 2023. 2

[18] S. Rijal, S. Adhikari, M. Dahal, M. Awale, and V. Ojha. Whisper finetuning on nepali language. 2023. Preprint. 1

[19] C. Schmidt, M. Stadtschnitzer, and J. Koehler. The Fraunhofer IAIS audio mining system: Current state and future directions. In *Speech Communication; 12. ITG Symposium*, pages 1–5, 2016. 2

[20] J. van Doorn, J. Sun, J. Hoekstra, P. Jonk, and V. de Vries. Whisper-ATC: Open models for air traffic control automatic speech recognition with accuracy. In *International Conference on Research in Air Transportation*, 2024. 1

[21] J. J. Williams, M. Borge, J. Levin, H. Guo, and C. P. Rosé. 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. 1

[22] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, and S. Wang. 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. 1

Analyzing and Fine-Tuning Whisper Models for Multilingual Pilot Speech Transcription in the Cockpit

Supplementary Material

6. Additional Results

6.1. Dataset Adaptation: Scenario Comparison

In the supplementary results, we additionally compare transcription performance across four distinct operational scenarios: Takeoff briefings and checklists (10 scenarios, 20 minutes), ECAM actions (11 scenarios, 30 minutes), FORDEC decision-making procedures (3 scenarios, 15 minutes), and landing briefings and checklists (5 scenarios, 20 minutes). Additionally, a controlled interview scenario incorporating aviation-specific vocabulary was included for comparison (12 scenarios, 130 minutes).

6.2. Effect of Normalization

A comparison of different normalization schemes is presented in Tables 4 to 8. The evaluation of various normalizers across a family of Whisper models on five distinct scenarios: ECAM, FORDEC, Interview, Landing, and Takeoff shows a consistent trends in performance improvements. Across all scenarios, the No-norm baseline exhibits the highest word error rate (WER), indicating that raw model outputs contain significant transcription errors. Among the baseline normalizers, the Basic and English approaches consistently outperform the Number normalizer, with notable reductions in WER. The proposed normalization techniques further refine these results, with Proposed II and Proposed III showing the most robust performance across different Whisper models. Larger models (Turbo and Large) tend to benefit more from normalization than smaller models (Tiny and Base), suggesting that model capacity influences the effectiveness of text normalization.

In the ECAM and FORDEC scenarios, the Proposed II and Proposed III normalizers achieve the lowest WER for Medium, Turbo, and Large models. Specifically, in ECAM, Proposed II achieves a WER of 49.48 for Large, while in FORDEC, Proposed III achieves a WER of 43.09 for Large. The Interview scenario follows a similar trend, with Proposed II yielding the best results across most model sizes, achieving a WER of 23.75 % for Large. The English normalizer performs comparably well, often ranking close to Proposed II. For Landing scenario, Proposed II achieves the lowest WER of 64.86 for Large, whereas for Takeoff, Proposed III yields the lowest WER of 44.89. Overall, the results emphasize the importance of selecting appropriate normalization strategies to enhance ASR accuracy, particularly in specialized domains where raw model predictions tend to exhibit high error rates.

Table 4. ECAM: Comparison of proposed normalizers with baselines.

Normalizer	Tiny	Base	Small	Medium	Turbo	Large
No-norm	105.49	106.49	97.04	82.81	84.02	74.32
Basic	96.44	94.62	79.79	67.91	65.82	50.27
Number	98.18	98.38	86.13	73.74	77.31	59.25
English	94.57	94.37	79.86	67.47	65.39	50.07
Proposed I	94.78	94.85	79.60	64.73	65.90	50.58
Proposed II	94.75	94.44	79.64	67.02	65.14	49.48
Proposed III	94.65	94.58	79.33	64.18	65.24	50.15

Table 5. FORDEC: Comparison of proposed normalizers with baselines.

Normalizer	Tiny	Base	Small	Medium	Turbo	Large
No-norm	118.98	96.48	77.56	69.42	64.20	63.59
Basic	104.85	84.42	61.19	54.51	46.46	43.38
Number	107.37	88.66	68.74	61.66	55.67	52.68
English	104.28	81.39	60.90	54.34	46.06	42.69
Proposed I	103.33	83.22	61.26	54.48	46.32	43.62
Proposed II	104.73	81.66	61.28	54.49	45.96	43.34
Proposed III	103.32	80.05	61.09	54.38	45.96	43.09

Table 6. Interview: Comparison of proposed normalizers with baselines.

Normalizer	Tiny	Base	Small	Medium	Turbo	Large
No-norm	68.26	51.21	45.08	45.64	34.10	34.10
Basic	59.26	41.49	34.95	37.49	25.18	23.82
Number	66.79	48.93	42.73	43.98	32.01	31.08
English	58.96	41.53	34.92	37.46	25.12	23.82
Proposed I	59.72	41.50	35.35	37.58	25.38	24.13
Proposed II	59.35	41.41	34.78	37.44	25.05	23.75
Proposed III	59.37	41.56	35.29	37.58	25.33	24.07

Table 7. Landing: Comparison of proposed normalizers with baselines.

Normalizer	Tiny	Base	Small	Medium	Turbo	Large
No-norm	95.86	140.83	98.66	84.34	86.30	82.10
Basic	90.78	118.61	86.56	67.72	80.12	65.59
Number	90.65	123.06	90.42	73.47	81.49	69.37
English	87.43	119.16	86.18	67.64	75.40	65.48
Proposed I	87.73	118.93	86.36	66.19	76.14	65.10
Proposed II	87.67	119.25	85.74	67.61	75.08	64.86
Proposed III	87.72	119.40	85.92	66.83	75.22	64.58

6.3. Effect of Finetuning

Table 9 provides details about LoRA fine-tuning for different sizes of Whisper models. It presents the total number of parameters in each model, the number of additional LoRA parameters introduced during fine-tuning, and the percent-

Table 8. Takeoff: Comparison of proposed normalizers with baselines.

Normalizer	Tiny	Base	Small	Medium	Turbo	Large
No-norm	119.52	121.93	113.88	94.72	85.41	77.44
Basic	123.36	110.78	93.39	60.46	60.67	46.15
Number	107.50	109.88	98.12	71.71	71.14	55.41
English	112.02	103.74	93.63	60.11	60.11	45.69
Proposed I	105.99	104.62	90.54	60.50	60.79	46.14
Proposed II	112.04	103.75	93.62	59.49	59.77	45.68
Proposed III	108.51	103.69	90.64	59.79	59.64	44.89

Table 9. LoRA Finetuning details

Model	Total parameters	LoRA parameters	Percentage (%)
Tiny	38,350,464	589,824	1.5380
Base	73,773,568	1,179,648	1.5990
Small	245,273,856	3,538,944	1.4429
Medium	773,295,104	9,437,184	1.2204
Turbo	815,431,680	6,553,600	0.8037
Large	1,559,219,200	15,728,640	1.009

age of LoRA parameters relative to the total model size. LoRA requires only a small fraction (0.8% to 1.6%) of the total model parameters, reducing the number of trainable parameters while still allowing effective adaptation.

LoRA fine-tuning on Whisper Large to Whisper Tiny models with various learning rates is given in Tables 10 to 15. The fine-tuning results across Whisper models of varying sizes (Tiny, Base, Small, and Medium) demonstrate that LoRA fine-tuning leads to significant reductions in WER across all configurations, with the extent of improvement depending on model size, normalization technique, and learning rate. The pre-trained models exhibit relatively high WER, particularly in the absence of normalization, with the No-norm baseline consistently yielding the worst performance. Fine-tuning improves recognition accuracy substantially, with Proposed II and English normalizers achieving the lowest WER across most scenarios.

For Whisper Medium and Small models, the optimal learning rate appears to be 1e-3, where Proposed II and English yield the lowest WER (32.67% and 32.97% for Medium; 39.18% and 39.11% for Small). However, for Whisper Base and Tiny models, higher learning rates (1e-3) occasionally lead to performance degradation before normalization. Notably, the No-norm baseline for Whisper Tiny at 1e-3 results in a WER of 96.31%, exceeding that of the pre-trained model, while the normalized WER being lower for finetuned model over pre-trained. Among the normalization techniques, Proposed II and English consistently outperform other approaches, demonstrating their effectiveness in improving ASR accuracy post-fine-tuning.

Table 10. LoRA fine-tuning on Whisper Large model with various learning rates. The numbers indicate WER in %.

Normalizer	pre-trained	lr=1e-5	lr=1e-4	lr=1e-3
No-norm	68.49	58.83	64.36	55.65
Basic	52.23	27.96	37.09	27.37
Number	59.76	46.10	48.30	50.08
English	52.08	27.80	36.71	26.36
Proposed I	52.74	32.72	37.35	38.37
Proposed II	52.00	27.65	36.41	26.26
Proposed III	52.41	28.24	36.60	27.00

Table 11. LoRA fine-tuning on Whisper Turbo model with various learning rates. The numbers indicate WER in %.

Normalizer	pre-trained	lr=1e-5	lr=1e-4	lr=1e-3
No-norm	70.20	61.82	64.67	65.06
Basic	49.49	28.18	29.04	31.02
Number	62.18	43.08	46.64	47.61
English	48.88	28.01	28.81	30.40
Proposed I	49.68	29.17	29.98	31.70
Proposed II	48.69	28.24	28.88	30.32
Proposed III	48.71	28.40	28.67	30.55

Table 12. LoRA fine-tuning on Whisper Medium model with various learning rates. The numbers indicate WER in %.

Normalizer	pre-trained	lr=1e-5	lr=1e-4	lr=1e-3
No-norm	81.64	60.10	66.84	63.85
Basic	62.96	35.69	36.12	33.22
Number	70.84	50.35	50.46	49.27
English	62.43	34.48	35.96	32.97
Proposed I	63.19	36.87	36.76	35.63
Proposed II	62.20	34.60	35.18	32.67
Proposed III	62.54	34.24	36.28	33.22

Table 13. LoRA fine-tuning on Whisper Small model with various learning rates. The numbers indicate WER in %.

Normalizer	pre-trained	lr=1e-5	lr=1e-4	lr=1e-3
No-norm	85.64	75.67	70.33	63.72
Basic	69.35	43.26	48.63	39.88
Number	77.41	57.39	67.56	61.53
English	69.16	42.74	47.81	39.11
Proposed I	69.05	43.11	61.84	56.30
Proposed II	68.87	42.49	47.73	39.18
Proposed III	68.76	42.39	49.09	40.19

7. Challenges with multi-lingual speech

Table 16 shows instances where the Whisper model transcriptions struggles with unexpected translation, often misinterpreting words or phrases based on phonetic similarities rather than contextual meaning. For example, ”Gut” is

Table 14. LoRA fine-tuning on Whisper Base model with various learning rates. The numbers indicate WER in %.

Normalizer	pre-trained	lr=1e-5	lr=1e-4	lr=1e-3
No-norm	96.00	88.56	81.45	73.95
Basic	84.70	60.06	62.64	56.57
Number	88.70	72.29	69.55	72.40
English	84.58	59.92	60.40	56.08
Proposed I	83.10	60.49	58.55	69.97
Proposed II	84.25	60.11	59.95	56.00
Proposed III	82.96	60.55	60.24	57.23

Table 15. LoRA fine-tuning on Whisper Tiny model with various learning rates. The numbers indicate WER in %.

Normalizer	pre-trained	lr=1e-5	lr=1e-4	lr=1e-3
No-norm	94.41	92.06	86.34	96.31
Basic	91.73	90.80	66.24	75.79
Number	89.13	86.93	76.28	84.88
English	88.37	84.68	66.33	74.49
Proposed I	85.68	82.94	67.17	74.91
Proposed II	88.41	84.78	66.10	74.69
Proposed III	88.21	84.99	67.04	74.17

Table 16. Transcription with unexpected translation

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

Table 17. Transcription errors: Words with close phonetics.

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

incorrectly transcribed as ”Good,” reflecting a bias toward English interpretations. Similarly, longer phrases exhibit structural differences that lead to errors in word order and meaning retention.

Table 17 presents cases where the model’s predictions are very similar to the reference text but still contain subtle inaccuracies. These transcription errors often involve homophones or phonetically similar words, such as ”slats low” misrecognized as ”sled slow” and ”CAT3 single” transcribed as ”cut three single.”