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	The blue system is of course lost
daraufhin ein spoiler-pair	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.”