Evaluating LoRA Rank Selection for Fine-Tuning Whisper on Air Traffic Control Communications

Cody Li[1*], **Jianhua Liu**[2]

Received February 18, 2026
Accepted October 2, 2026
Electronic access November 15, 2026

Clear and accurate communication between air traffic controllers and pilots is critical for aviation operations. Automatic speech recognition (ASR) can be used to improve both safety and operational efficiency by providing real-time transcription of air traffic control (ATC) communications and enabling the detection of exchange errors. However, state-of-the-art pre-trained ASR models, such as OpenAI's Whisper, exhibit limited performance on ATC speech due to domain-specific phraseology, rapid speech rates, diverse accents, and high levels of background noise. Fine-tuning Whisper using low-rank adaptation (LoRA) can improve transcription accuracy on ATC communications while mitigating catastrophic forgetting of its general speech recognition capabilities. This study examines how varying the LoRA rank parameter impacts the trade-off between the ability to maintain the model's overall performance in general domains and its ability to adapt to ATC-specific terminology. A publicly accessible ATC dataset was used to fine-tune and evaluate the Whisper small English model at several different LoRA rank values. The study indicates that a LoRA rank of 64 is a suitable compromise, as it significantly reduces the word error rate (WER) for ATC communication by over 85%, from 137.2% to 18.8%, while increasing the LibriSpeech WER from 3.1% to 8.3%.

Keywords: Automatic Speech Recognition, Air Traffic Control Communication, Whisper Fine-Tuning, LoRA Rank Optimization.

Introduction

Accurate voice communication between air traffic controllers and pilots is highly important to aviation safety and operational efficiency. In order to maintain safe and orderly flight operations, air traffic controllers are required to deliver precise and unambiguous verbal instructions to pilots during all flight phases, including taxi, takeoff, cruise, approach, and landing. Although the International Civil Aviation Organization (ICAO) and the Federal Aviation Administration (FAA) have published standardized procedures and language for air traffic control (ATC) communications[1,2], both air traffic controllers and pilots may sometimes deviate from these conventions and use different expressions, especially in high-stress situations.

Controlled studies have identified several factors that could negatively impact the accuracy of controller-pilot exchanges. These include, for example, transmissions with four or more information items, high pilot workload, and the absence of prosodic cues such as pauses. The effect of missing prosodic cues is particularly strong among non-native English-speaking pilots[3]. Other studies have shown that accent, speech rate, and information density also contribute to miscommunication. Controllers and pilots identify these characteristics as major sources of communication errors[4–6]. In fact, human factors are major contributors to all types of aviation accidents[7]. A recent study reports that human factors contribute to about 75% of aviation accidents and incidents around the world, with nearly one-third being caused by a lack of effective communication between pilots and controllers[8].

These findings demonstrate the need for robust technologies that can be used to improve the reliability of ATC communications. Accurate automatic speech recognition (ASR) models can serve as decision-support tools for both controllers and pilots by providing real-time transcription of voice communications. An ATC-trained ASR model can transcribe controller-pilot speech and automatically extract operational information such as call signs and runway numbers, as well as support conformance monitoring and controller training[9]. ASR-based support tools have also been reported to cut the time controllers spend entering and confirming clearances at their workstations by approximately 20 percentage points[10]. This capability can help mitigate communication errors, improve operational awareness, and offer an additional layer of redundancy to support aviation safety.

ASR research dates back to the 1950s, when the work demonstrated the ability to recognize spoken digits[11]. Within the ATC domain, a variety of approaches have been studied.

[1] *Carroll High School, Southlake, Texas*
[2] *Embry-Riddle Aeronautical University, Daytona Beach, Florida*
[*] *Email: codyli9219@gmail.com*

© The National High School Journal of Science 2026

Early systems were developed based on hidden Markov models (HMMs) to perform ATC speech recognition[12]. Later approaches used a hybrid HMM framework with a combination of factorized time-delay neural networks (TDNN-F) and convolutional neural networks (CNNs). Using training data consisting of over 170 hours of ATC speech collected from six separate ATC corpora, the models were able to achieve an average WER of 7.75% across four test sets[13]. A survey of the field identifies code-switching, variable speech rates, and radio channel distortion as major challenges for spoken instruction understanding in ATC[14]. Recent research has examined multiple techniques to overcome these issues. For example, one study proposed a single convolutional-recurrent model which was trained using a connectionist temporal classification objective for multilingual ATC recognition[15]. Another study demonstrated that acoustic characteristics derived directly from raw ATC waveform data yielded better performance than handcrafted acoustic features when subjected to radio channel distortion conditions[16]. Finally, it has been demonstrated that adapting language models using ATC instruction text can lead to improved results[17]. A detailed review of ATC corpora, models, and evaluation methods is also available[18].

Among the ASR models evaluated for ATC communications, Whisper has emerged as a good candidate for this type of application. Whisper is a family of ASR models developed by OpenAI and released in late 2022. It was trained on 680,000 hours of weakly supervised multilingual and multitask audio. In zero-shot evaluation, Whisper approached human-level robustness to accents, background noise, and technical language without any dataset-specific fine-tuning[19]. Research has been conducted on the potential use of Whisper as an ASR system for ATC speech recognition. The most common method of evaluating the effectiveness of Whisper for this purpose is full fine-tuning of the pre-trained model on ATC communication corpora[20,21]. As shown in studies, fine-tuning significantly improves the accuracy of transcripts produced by Whisper. One study reported WERs of 13.5% on the ATCO2 dataset and 1.17% on the ATCOSIM dataset, and improved performance by up to 60% in real-world applications[20]. Another study demonstrated that the WER for Southeast Asian-accented ATC speech reached 9.82% after accent-based fine-tuning[21].

However, fully fine-tuning the entire Whisper model is computationally intensive and typically requires a long period of time for training. Also, fine-tuning all parameters in the model on a comparatively small ATC-specific dataset may lead to an increased risk of overfitting, thus limiting the generalization capabilities of the model when exposed to varied accents, regions, and operational environments. Full fine-tuning may also result in catastrophic forgetting, the well-documented phenomenon in which previously acquired capabilities are significantly degraded when the model is trained sequentially on a new task[22]. Therefore, to avoid this issue, researchers have developed parameter-efficient fine-tuning (PEFT) methods as a viable option for training models[23]. One widely used PEFT technique is called low-rank adaptation (LoRA). When LoRA is used, the pre-trained model weights are frozen, and a set of trainable low-rank matrices is introduced into the transformer layers. LoRA was shown to reduce the number of trainable parameters for GPT-3 175B by up to four orders of magnitude compared with full fine-tuning[24]. Quantized LoRA (QLoRA) is an extension of LoRA that trains adapter layers over a 4-bit quantized base model. This significantly reduces the amount of memory needed for fine-tuning without a loss in performance when compared to 16-bit fine-tuning[25]. PEFT can help prevent catastrophic forgetting in speech processing tasks; however, PEFT does not always eliminate catastrophic forgetting. Bayesian regularization of adapter layer parameters has been proposed to maintain the information learned during pre-training[26].

When adapting a pre-trained large model using LoRA, the value of the rank parameter is particularly important. The rank parameter determines the number of additional trainable parameters that will be included in the modified model after the adaptation process. Additionally, the rank parameter has an impact on the trade-off between domain-specific fine-tuning and preserving the broad generalization capability of the pre-trained model. A direct comparison of LoRA and full fine-tuning on programming and mathematics tasks found that LoRA substantially underperforms full fine-tuning at commonly used low ranks[27]. As the rank of LoRA increases, the performance difference between LoRA and full fine-tuning becomes smaller, and for $r = 256$, LoRA's performance in some cases approaches that of full fine-tuning. Another study tested several LoRA ranks on reasoning and factual-recall question-answering tasks. It was observed that no one rank performed optimally across all tasks; however, a range of ranks ($r = 32 - 64$) performed better in multiple task environments[28].

In the area of speech recognition, LoRA was used to adapt Whisper for multilingual use and to expand its support for additional languages. To reduce language interference and allow for the addition of new languages, language-specific LoRA modules were implemented in Whisper. The results indicated that this method provided relative improvements of 18.5% for multilingual ASR and 23.0% for language expansion when compared to the baseline models, while requiring only approximately 5% of the total training parameters as compared to fine-tuning the entire model[29]. A separate study indicated that a combination of low-rank factorization and LoRA adapters can reduce training resources by up to 90%, with similar ASR performance[30]. LoRA has been utilized within the aviation sector to transcribe controller-pilot communications us-

© The National High School Journal of Science 2026

ing Distil-Whisper[31] and to provide transcription services for multilingual cockpit speech using Whisper[32]. Both of these studies focused on enhancing the accuracy of in-domain transcription. However, neither study systematically investigated the effect of LoRA rank on the trade-off between in-domain adaptation and retention of the model's general-domain capabilities.

Despite its importance, the optimal selection of LoRA rank for ATC-specific Whisper applications has received little research attention. Previous studies have either adopted the full fine-tuning approach, i.e., no LoRA was used[20,21], or applied LoRA but without evaluating its rank selection against the general speech recognition performance[31,32]. To the best of the authors' knowledge, no prior research has empirically examined how LoRA rank influences the balance between adapting to an application-specific domain (ATC) and maintaining the original speech recognition capabilities of Whisper for general English. Additionally, the transferability of the benefits associated with intermediate LoRA ranks from large text-based language models to speech recognition models (which have acoustic instead of lexical encoders) remains uninvestigated. The existence of these gaps in current literature serves as the primary rationale behind this study.

The main objective of this research is to examine the effect of LoRA rank value on the capability of the Whisper small English model to adapt to ATC communications. Note that the Whisper small English model was chosen because it offers a good balance between accuracy and the computational cost of fine-tuning. We hypothesized that lower ranks would allow the model to retain its pre-trained general speech recognition abilities better than higher ranks; however, lower ranks would be less capable of adapting to ATC-specific characteristics. In contrast, higher ranks would allow for more effective ATC domain adaptation, but they might also lead to an increased risk of overfitting or losing previously learned capabilities. These competing effects suggest that an optimal LoRA rank may exist that helps to improve ATC speech recognition while preserving the model's general speech recognition capability.

The remainder of this paper is organized as follows. Section II describes the methodology of this study, including the training and test datasets and the LoRA fine-tuning approach. Section III presents the experimental results. Section IV discusses the findings of this research. Section V provides a summary and suggests areas of further investigation.

Methods

Datasets

The ability to obtain high-quality ATC speech datasets is still a challenge in the field of aviation speech recognition research. While there are some large-scale datasets available (e.g., ATCO2), many ATC datasets require a license or purchase, which can limit access for researchers. For this study, we use a free-to-access dataset that combines publicly available portions of the UWB-ATCC corpus and the one-hour ATCO2 test subset.

The UWB-ATCC recordings were collected over Czech airspace by the Air Navigation Services of the Czech Republic, and then manually transcribed by the University of West Bohemia. This dataset consists primarily of English speech with a Czech accent. Speakers are identified only as controllers or pilots, and no information on age, gender, or identity is provided with the corpus[33]. The ATCO2 one-hour test subset is a freely available portion of the manually transcribed four-hour ATCO2 test set and contains real ATC communications recorded over very high frequency (VHF) radio channels. The ATCO2 collection contains data recorded at several different airports, including Sion, Bern, and Zurich in Switzerland. The speakers represented within the data are both native and non-native English speakers. Therefore, the data exhibits a large degree of variation in accent and operational acoustic environment. The one-hour subset of this data has been made available to aid in the design, testing, and evaluation of English ATC ASR systems[34]. Both parent corpora represent actual operational radio communications, as opposed to simulated or read speech. As such, they display many common characteristics found in typical ATC communications, including limited bandwidth, static, clipping, and variability in signal strength.

All audio in the consolidated dataset is single-channel and sampled at 16 kHz, matching the input sampling rate required by Whisper. The combined dataset of UWB-ATCC and ATCO2 samples is then randomly divided into training and test sets. The resulting training and test sets contain 11,868 training samples (approximately 11.2 hours) and 2,927 test samples (approximately 2.7 hours), and are hosted on Hugging Face[35].

To assess the retention of general speech recognition capability after domain-specific fine-tuning, the LibriSpeech test-clean subset is used as a general speech recognition benchmark[36]. LibriSpeech is composed of read audiobook speech. The characteristics of LibriSpeech are significantly different from those found in ATC communication in terms of vocabulary, channel characteristics, and speaker style. LibriSpeech therefore is used to indicate whether the ability of Whisper to recognize English speech remains intact following ATC-specific adaptation. The only portion of LibriSpeech utilized in this study is the test-clean split, which contains 2,620 samples (approximately 5.4 hours) and is used to assess the effectiveness of Whisper on an independent evaluation set. No LibriSpeech data was used in the training process. The test-clean split has become a common benchmark for evaluating ASR performance[19,36]. A disadvantage of utilizing the test-clean

split is that it contains clean read speech. Therefore, it only provides an assessment of general-domain retention in a relatively favorable acoustic environment, and the WER obtained from this split may not be representative of general-domain retention more broadly. Performance on more challenging general-domain speech, such as conversational or noisy audio, is not evaluated in this study and is left for future work.

The analysis of the ATC dataset indicates that there is a significantly larger number of short-duration clips than in the LibriSpeech dataset. Figure 1(a) shows that some of the ATC samples have a duration of only 1 second, while the median duration of the ATC samples is approximately 3.14 seconds. This reflects the fact that short, abrupt utterances are more common in ATC communications than in conventional speech.

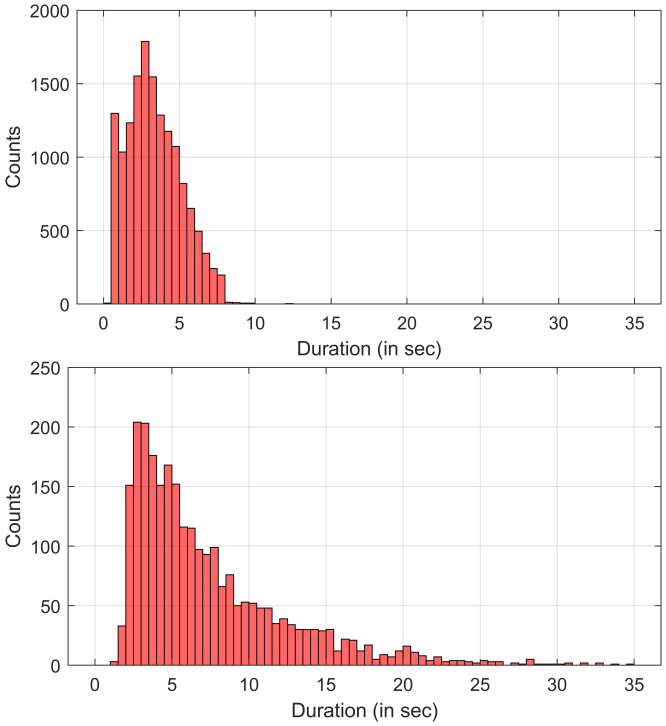


Figure. 1 Histogram of sample durations in (a) the ATC dataset and (b) the LibriSpeech test-clean dataset. The ATC dataset shown in (a) contains 14,795 samples (13.9 hours), including 11,868 training samples (11.2 hours) and 2,927 test samples (2.7 hours). The median sample duration is 3.14 seconds, with only 44 samples longer than 8 seconds. In contrast, the LibriSpeech test-clean dataset shown in (b) contains substantially longer samples, with a median duration of 5.8 seconds.

Upon examination of the original ATC data, it is evident that there are several quality issues, including (1) typographical errors, which include obvious misspellings that exist within the reference text (e.g., "actualy," "singarpoore," "qality"); (2) mistranscription, where specific phrases were incorrectly transcribed (e.g., "cleared to land" was transcribed as "cleaned to land"); and (3) segment truncation, where words located at the beginning or end of a segment were truncated, and the corresponding transcriptions were altered to match the truncated segments. All of these issues are typical of label noise in large-scale speech corpora.

Although manual cleanup could improve the quality of the dataset, a complete cleanup will require experts with domain knowledge to differentiate between transcription errors (such as misidentification) and natural variations found within ATC communications (e.g., different phraseologies, call-sign pronunciations, accents), as well as the effects of degraded radio audio. To avoid introducing reproducibility issues, the dataset is used in its original form, as this study focuses on the effects of LoRA rank selection rather than the absolute WER values. The presence of transcription noise can have an effect on WER values and therefore must be accounted for when interpreting the findings from this research. Systematic dataset cleanup and evaluation of its impact are left for future work.

Word Error Rate

Word error rate (WER), which is a common metric used to evaluate an ASR system's performance, measures how much of the ASR-produced output differs from the true or "correct" reference transcript of what was spoken[18,31,32]. WER is formally defined as

$$\mathrm{WER}\,(\%) = \left[\frac{S+D+I}{N}\right] \times 100 \tag{1}$$

where S represents the number of substitution errors (i.e., words misidentified as another word), D represents the number of deletion errors (i.e., words present in the reference transcript but missed from the ASR output), I represents the number of insertion errors (i.e., additional words generated by the ASR system that were not included in the reference transcript), and N represents the total number of words in the reference transcript. The WER value will be smaller if ASR performs better; therefore, a WER of 0% indicates that ASR has perfectly recognized the transcript. In this study, we calculated WER values with the use of the jiwer library*.

Algorithm

The ATC training dataset contains 11,868 samples, which were divided into five approximately equal-sized subsets. The five-fold cross-validation method was applied to every LoRA rank being tested. In this method, four of the five subsets were

*https://github.com/jitsi/jiwer

© The National High School Journal of Science 2026

utilized to train the model while one subset was utilized to validate the model. The entire procedure was repeated five times; therefore, each subset served as a validation set once.

The LoRA fine-tuning algorithm for the Whisper model was implemented using Unsloth[†] with the Hugging Face Transformers framework. LoRA adapters are applied to the query, key, value, and output projection matrices (`q_proj`, `k_proj`, `v_proj`, `out_proj`) in both the encoder and the decoder, including the decoder cross-attention blocks. All bias parameters were frozen (bias="none") and LoRA dropout was set to 0.05, identical to what Dettmers et al. employed[25]. For each LoRA rank r, the scaling factor α was set to $2r$. Thus, the effects of rank on fine-tuned performance were assessed with a consistent scaling ratio[27]. The learning rate was set to 2×10^{-4}. A cosine learning rate scheduler was used, and a warm-up period was set to 500 steps. Each graphics processing unit (GPU) had a batch size of 8. Gradient accumulation was performed over two steps, which resulted in an effective batch size of 16. The weight decay was set to 0.01 for regularization purposes.

Training was performed using the Hugging Face Transformers `Seq2SeqTrainer`. The maximum number of training epochs was limited to 20. Training stopped early if the validation WER did not improve for three consecutive epochs. The checkpoint with the lowest validation WER was kept as the final model. The training and validation loss curves were monitored for each rank and each fold in order to verify that there were no runs which demonstrated either divergence or large increases in loss. Experiments were run on the Google Colab Pro+ platform, which utilizes a G4 GPU with 96 GB VRAM. The typical fine-tuning time for each LoRA rank on one cross-validation fold is approximately 3 hours. Key LoRA fine-tuning hyperparameters are summarized in Table 1.

When performing ASR model training and computing WER, reference transcripts need to be normalized. In this work, the `EnglishTextNormalizer` from OpenAI's original Whisper package (`whisper.normalizers`) was used, which performed a standard set of text normalization steps, including lowercasing, punctuation removal, word-to-number conversion, and British-to-American spelling conversion[‡].

Table 1 Fine-tuning hyperparameter settings used in all experiments.

Parameter	Value
Base model	whisper-small.en
LoRA target modules	q_proj, k_proj, v_proj, out_proj
LoRA rank (r)	4, 8, 16, 32, 64, 128, 256
LoRA alpha (α)	$\alpha = 2r$
LoRA dropout	0.05
LoRA bias	none
Learning rate	2×10^{-4}
Learning rate scheduler	cosine
Warm-up steps	500
Weight decay	0.01
Batch size	8
Max. training epochs	20
Training early stopping criterion	No improvement in validation WER for 3 consecutive epochs
Training optimizer	AdamW (torch)
Cross-validation folds	5
Training platform	Google Colab Pro+

Results

First, the performance of the default Whisper models, from Tiny to Large variants, on both the ATC and LibriSpeech test datasets, is illustrated in Figure 2.

These results provide evidence that Whisper models perform very well on general English speech but perform poorly on ATC communications. Specifically, the Whisper small English model (which is used as the baseline model in this research) had a WER of 137.2% on ATC communications data, whereas it had a WER of only 3.1% on LibriSpeech data.

Figure 3 shows the performance comparison of the fine-tuned Whisper small English model evaluated at different LoRA rank values. The results from all five folds are shown. It is clear from this figure that WER on the ATC test dataset is substantially reduced after fine-tuning. Increasing the rank generally leads to a lower WER on the ATC dataset, while WER on the LibriSpeech dataset tends to increase.

To determine the statistical significance of differences in WER between successive LoRA ranks, paired t-tests were performed on WER difference scores for each pair of consecutive ranks. A smaller p-value signifies greater statistical evidence that the WER difference represents a true difference. Cohen's d is a common way to measure an effect size and was computed using the formula $d = \mu_\Delta/\sigma_\Delta$. Here, μ_Δ and σ_Δ are the mean and standard deviation of the paired WER difference scores for all five folds[37]. A larger absolute value of d indicates a larger mean difference relative to variability differences across folds.

As shown in Table 2, ATC WER generally decreases as rank increases through $r = 64$. Statistically significant reductions in WER are observed for the transitions from 4 to 8 and from 32 to 64. Beyond $r = 64$, no statistically significant improvements are observed. LibriSpeech degradation, by contrast, is statistically significant for all transitions from 16 to 32 through 128 to 256; for example, when r changes from 32

[†] `https://github.com/unslothai`

[‡] All source code is available at `https://github.com/C-L-South/whisper-atc`. All fine-tuned models are available at `https://huggingface.co/cody-li`.

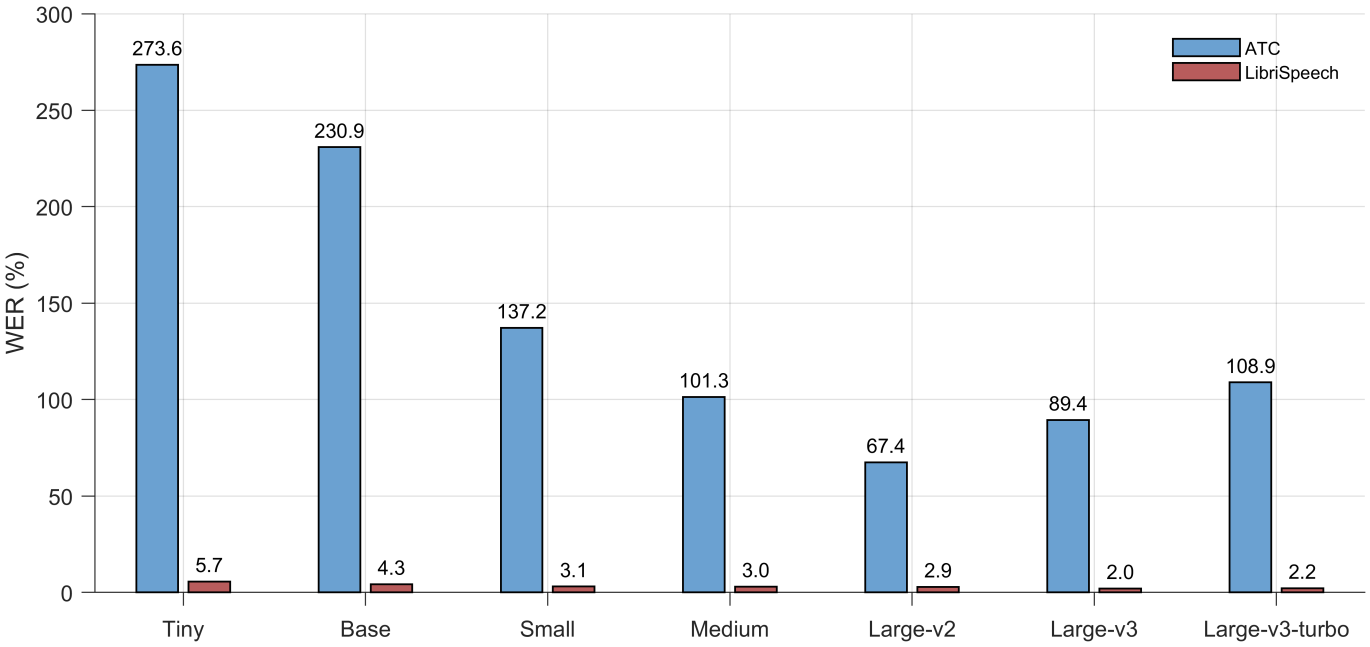


Figure. 2 Performance comparisons of the default Whisper models across the Tiny to Large variants on ATC and LibriSpeech test datasets.

to 64, $p = 0.004$ and $d = 2.62$, and when r changes from 64 to 128, $p = 0.0006$ and $d = 4.34$. The results of this analysis suggest that as the rank value increases above 64, additional improvements in ATC WER are small and not statistically significant, while LibriSpeech WER continues to increase. This finding is consistent with the interpretation that higher-rank adapters allow the model to change more during fine-tuning but can cause it to forget more of what it previously learned from pre-training, i.e., catastrophic forgetting [27].

Additionally, as shown in Table 3, an equal-weighted combined WER ($w = 0.5$) is minimized at $r = 64$, and this choice remains optimal even when weighting favors in-domain accuracy ($w = 0.7$), while also exhibiting the lowest cross-fold variance of any rank tested. Rank $r = 64$ is therefore considered the optimal rank for the fine-tuned model.

Discussion

While keeping the pre-trained Whisper weights frozen, LoRA improves domain-specific speech recognition performance by training only the adapter parameters. As shown in Table 4, the trainable LoRA parameters are 0.37% of the base model's parameter count at $r = 4$ and 23.42% at $r = 256$. When $r = 64$, the new parameters account for only 5.86% of the base model's parameter count. Furthermore, LoRA checkpoints store only adapter parameters, which substantially reduces checkpoint size as compared with a full fine-tuning scheme [24]. This method also allows a single frozen base model to support multiple lightweight, task-specific LoRA adapters.

The results of this study clearly show the rank-dependent trade-off between ATC adaptation and the retention of general English recognition capability. It should be pointed out that preserving general speech recognition capability is important because real-world ATC communications are not limited to standardized phraseology. In non-routine or unexpected situations, pilots and controllers may also use plain English. Therefore, an ATC-adapted ASR model should improve recognition of domain-specific communications without substantially degrading its ability to recognize general English speech.

Increasing LoRA rank was correlated with lower ATC WER until $r = 64$; however, at $r > 64$, there was no statistically significant increase in performance. Conversely, LibriSpeech WER increased as LoRA rank increased. The results suggest that increasing adapter capacity (LoRA rank) above $r = 64$ does not provide much benefit for improving performance on ATC speech recognition tasks, while simultaneously degrading the ability of the pre-trained model to recognize general-domain speech. In addition, as shown in Table 3, the combined-WER analysis further supports this conclusion, as $r = 64$ produces the best trade-off in terms of WER for the two weighting methods. Therefore, LoRA rank is an important hyperparameter to consider during model development for ATC. Evaluation of candidate LoRA ranks should be performed on both in-domain and out-of-domain speech recognition data to determine if the improvement in ATC recognition comes at the expense of general English recognition.

© The National High School Journal of Science 2026

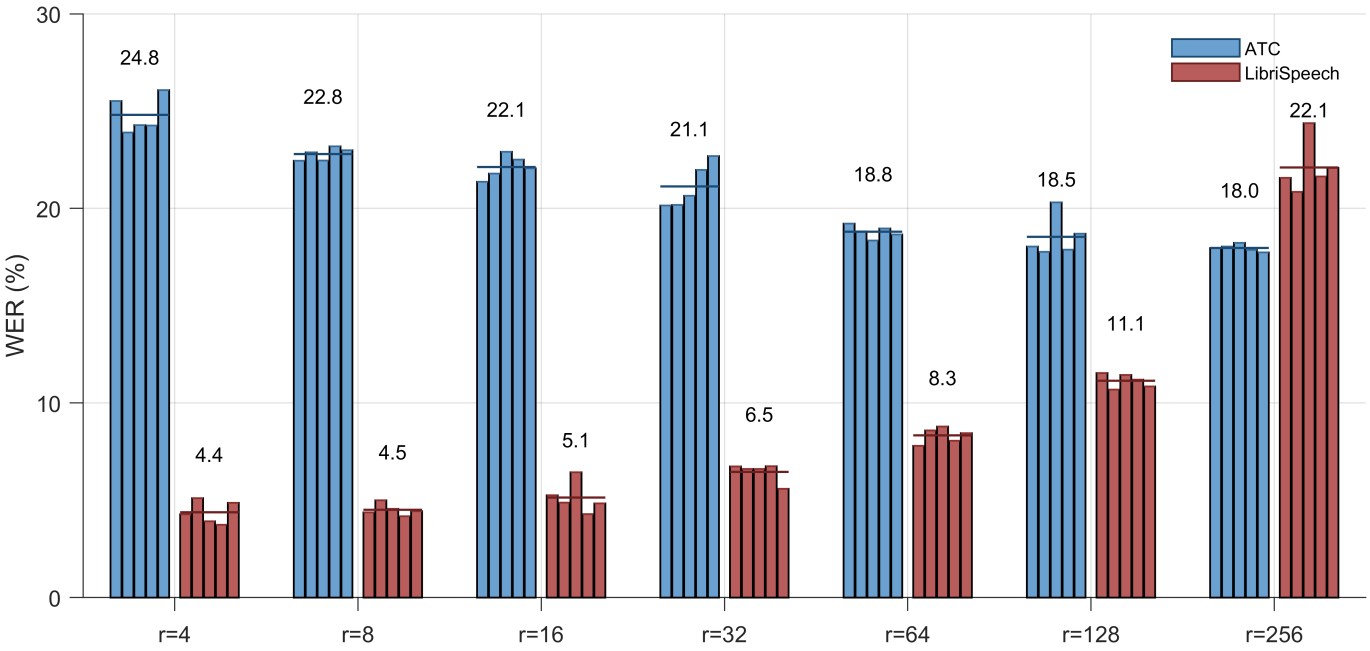


Figure. 3 WERs of the fine-tuned Whisper small English model on ATC and LibriSpeech test datasets with different LoRA fine-tuning rank parameters. For each rank, five models were evaluated using five-fold cross-validation.

Table 2 Paired t-test results (p-value) and effect size (Cohen's d) between consecutive LoRA ranks for ATC and LibriSpeech WER.

Rank change	**ATC** (p-value)	**ATC** (d)	**LibriSpeech** (p-value)	**LibriSpeech** (d)
$4 \rightarrow 8$	0.012	-1.97	0.539	0.30
$8 \rightarrow 16$	0.082	-1.03	0.150	0.80
$16 \rightarrow 32$	0.114	-0.90	0.029	1.49
$32 \rightarrow 64$	0.014	-1.86	0.004	2.62
$64 \rightarrow 128$	0.682	-0.20	0.0006	4.34
$128 \rightarrow 256$	0.254	-0.60	< 0.0001	9.13

Table 3 Combined WER (mean $\pm$ standard deviation) vs. LoRA rank.

Rank (r)	**WER**$_{\text{ATC}}$ **(%)**	**WER**$_{\text{Libri}}$ **(%)**	**WER**$_{\text{Combined}}$ **(%)**, $w = 0.5$	**WER**$_{\text{Combined}}$ **(%)**, $w = 0.7$
4	24.81 ± 0.94	4.38 ± 0.59	14.60 ± 0.61	18.68 ± 0.72
8	22.79 ± 0.33	4.51 ± 0.31	13.65 ± 0.20	17.31 ± 0.23
16	22.13 ± 0.60	5.14 ± 0.80	13.63 ± 0.58	17.03 ± 0.56
32	21.13 ± 1.15	6.46 ± 0.49	13.79 ± 0.43	16.73 ± 0.70
64	18.80 ± 0.33	8.34 ± 0.40	13.57 ± 0.08	15.66 ± 0.13
128	18.54 ± 1.05	11.14 ± 0.37	14.84 ± 0.62	16.32 ± 0.79
256	17.97 ± 0.18	22.10 ± 1.35	20.04 ± 0.73	19.21 ± 0.49

Note: $\text{WER}_{\text{Combined}} = w \times \text{WER}_{\text{ATC}} + (1 - w) \times \text{WER}_{\text{Libri}}$, where w is the weighting factor.

The importance of domain-specific adaptation can also be illustrated by comparing LoRA fine-tuned results with those obtained from default pre-trained Whisper models. The $r = 64$ Whisper small English model achieves a substantially lower ATC WER than the much larger original pre-trained Whisper large-v2 model. This result suggests that, for ATC speech

© The National High School Journal of Science 2026

Table 4 Trainable LoRA parameters as a percentage of total base model parameters across rank values for the Whisper small English model.

Rank (r)	4	8	16	32	64	128	256
Percentage (%)	0.37	0.73	1.46	2.93	5.86	11.71	23.42

Note: The Whisper small English base model contains 241,734,144 parameters.

recognition, domain-specific adaptation can provide greater performance benefits than simply increasing the size of the base model.

The ATC dataset used in this study was collected in European airspace. The characteristics of ATC communication can vary between regions in terms of accents, phraseology, radio conditions, and speaking speed. Therefore, the optimal rank identified in this study may not be the optimal choice for data from other regions. In future work, we will test larger ATC datasets from more regions to see if the same rank–performance trade-off observed in this study still holds.

Conclusion

This study evaluated the effect of the LoRA rank parameter value on fine-tuning the Whisper small English model for ATC communications. Seven rank values were evaluated, and the results showed that $r = 64$ provided the best overall balance, as it significantly improved ATC speech recognition while limiting the degradation in the Whisper model's general English speech recognition capability. Increasing the rank beyond 64 resulted in little additional improvement in ATC recognition while substantially degrading the general-domain performance. These results demonstrate that choosing a suitable LoRA rank is critical when fine-tuning the Whisper model. Future work includes the evaluation of medium and large Whisper models and the investigation of additional fine-tuning parameters. We also plan to test larger, higher-quality, and more diverse ATC datasets obtained from different regions. This will help to determine whether the rank value trade-off observed in this study holds for different accents and varying operating conditions.

References

1 ICAO, Manual of Radiotelephony, Doc 9432, 4th ed. Montreal, Canada: International Civil Aviation Organization, 2007., https://store.icao.int/en/manual-of-radiotelephony-doc-9432.

2 FAA, Air Traffic Control, Order JO 7110.65BB. Washington, DC, USA: Federal Aviation Administration, 2025., https://www.faa.gov/air_traffic/publications/atpubs/atc_html/.

3 B. R. C. Molesworth and D. Estival, Miscommunication in general aviation: The influence of external factors on communication errors. Safety Science. Vol. 73, pg. 73–79, Mar. 2015., https://doi.org/10.1016/j.ssci.2014.11.004.

4 Q. Wu, B. R. C. Molesworth, and D. Estival, An investigation into the factors that affect miscommunication between pilots and air traffic controllers in commercial aviation. The International Journal of Aerospace Psychology. Vol. 29, no. 1–2, pg. 53–63, 2019., https://doi.org/10.1080/24721840.2019.1604138.

5 Y. H. P. S. A. Y. Dissanayaka, B. R. C. Molesworth, and D. Estival, Miscommunication in commercial aviation: The role of accent, speech rate, information density, and politeness markers. The International Journal of Aerospace Psychology. Vol. 33, no. 1, pg. 79–97, 2023., https://doi.org/10.1080/24721840.2022.2154672.

6 H.-H. Yang, Y.-H. Chang, and Y.-H. Chou, Subjective measures of communication errors between pilots and air traffic controllers. Journal of Air Transport Management. Vol. 112, pg. 102461, 2023., https://doi.org/10.1016/j.jairtraman.2023.102461.

7 S. Shappell, C. Detwiler, K. Holcomb, C. Hackworth, A. Boquet, and D. A. Wiegmann, Human error and commercial aviation accidents: An analysis using the Human Factors Analysis and Classification System. Human Factors. Vol. 49, no. 2, pg. 227–242, Apr. 2007., https://doi.org/10.1518/001872007X312469.

8 R. A. Sihombing, N. Praptiningsih, and Y. Abimanyu, Analysis of miscommunication between pilot and air traffic control (ATC) and its contribution to flight incidents. International Journal of Humanities Education and Social Sciences. Vol. 5, no. 4, pg. 1737–1743, Feb. 2026., https://doi.org/10.55227/ijhess.v5i4.2194.

9 S. Badrinath and H. Balakrishnan, Automatic speech recognition for air traffic control communications. Transportation Research Record. Vol. 2676, no. 1, pg. 798–810, 2022., https://doi.org/10.1177/03611981211036359.

10 H. Helmke, O. Ohneiser, T. Mühlhausen, and M. Wies, Reducing controller workload with automatic speech recognition. 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 2016, pg. 1–10., https://doi.org/10.1109/DASC.2016.7778024.

11 K. H. Davis, R. Biddulph, and S. Balashek, Automatic recognition of spoken digits. The Journal of the Acoustical Society of America. Vol. 24, no. 6, pg. 637–642, Nov. 1952., https://doi.org/10.1121/1.1906946.

12 R. de Córdoba, J. Ferreiros, R. San-Segundo, J. Macías-Guarasa, J. M. Montero, F. Fernández, L. F. D'Haro, and J. M. Pardo, Air traffic control speech recognition system cross-task & speaker adaptation. IEEE Aerospace and Electronic Systems Magazine. Vol. 21, no. 9, pg. 12–17, Sep. 2006., https://doi.org/10.1109/MAES.2006.1705165.

13 J. Zuluaga-Gomez, P. Motlicek, Q. Zhan, K. Veselý, and R. Braun, Automatic speech recognition benchmark for air-traffic communications. Proceedings of Interspeech 2020, Shanghai, China, Oct. 2020, pg. 2297–2301., https://doi.org/10.21437/Interspeech.2020-2173.

14 Y. Lin, Spoken instruction understanding in air traffic control: Challenge, technique, and application. Aerospace. Vol. 8, no. 3, pg. 65, Mar. 2021., https://doi.org/10.3390/aerospace8030065.

15 Y. Lin, D. Guo, J. Zhang, Z. Chen, and B. Yang, A unified framework for multilingual speech recognition in air traffic control systems. IEEE Transactions on Neural Networks and Learning Systems. Vol. 32, no. 8, pg. 3608–3620, Aug. 2021., https://doi.org/10.1109/TNNLS.2020.3015830.

16 P. Fan, X. Hua, Y. Lin, B. Yang, J. Zhang, W. Ge, and D. Guo, Speech recognition for air traffic control via feature learning and end-to-end training. IEICE Transactions on Information and Systems. Vol. E106.D, no. 4, pg. 538–544, Apr. 2023., https://doi.org/10.1587/transinf.2022EDP7151.

17 J. Fan and W. Pan, Customization of the ASR system for ATC speech with improved fusion. Aerospace. Vol. 11, no. 3, pg. 219, Mar. 2024., https://doi.org/10.3390/aerospace11030219.

© The National High School Journal of Science 2026

18 Z. Wang, P. Jiang, Z. Wang, B. Han, H. Liang, Y. Ai, and W. Pan, Enhancing air traffic control communication systems with integrated automatic speech recognition: Models, applications and performance evaluation. Sensors. Vol. 24, no. 14, pg. 4715, Jul. 2024., https://doi.org/10.3390/s24144715.

19 A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning (ICML 2023), Honolulu, HI, USA, Jul. 2023. Vol. 202, pg. 28492–28518., https://proceedings.mlr.press/v202/radford23a.html.

20 J. van Doorn, J. Sun, J. M. Hoekstra, P. Jonk, and V. de Vries, Whisper-ATC: Open models for air traffic control automatic speech recognition with accuracy. Proceedings of the 11th International Conference on Research in Air Transportation (ICRAT 2024), Singapore, Jul. 2024, Paper ICRAT 2024-83.

21 M. Y. Z. Wee, J. J. H. Wong, L. Lim, J. Y. W. Tan, P. Gupta, D. Lim, E. H. Tew, A. K. S. Han, and Y. Z. Lim, Adapting automatic speech recognition for accented air traffic control communications. Proceedings of the International Conference on Military Communications and Information Systems (ICMCIS 2025), Oeiras, Portugal, May 2025, pg. 1–10., https://doi.org/10.1109/ICMCIS64378.2025.11047661.

22 J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences. Vol. 114, no. 13, pg. 3521–3526, Mar. 2017., https://doi.org/10.1073/pnas.1611835114.

23 Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, Parameter-efficient fine-tuning for large models: A comprehensive survey. Transactions on Machine Learning Research, 2024., https://openreview.net/forum?id=IIsCS8b6zj.

24 E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, LoRA: Low-rank adaptation of large language models. Proceedings of the International Conference on Learning Representations (ICLR), 2022., https://openreview.net/forum?id=nZeVKeeFYf9.

25 T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, Dec. 2023, pg. 10088–10115., https://doi.org/10.52202/075280-0441.

26 H. Chen and P. N. Garner, Bayesian parameter-efficient fine-tuning for overcoming catastrophic forgetting. IEEE/ACM Transactions on Audio, Speech, and Language Processing. Vol. 32, pg. 4253–4262, 2024., https://doi.org/10.1109/TASLP.2024.3463395.

27 D. Biderman, J. Portes, J. J. G. Ortiz, M. Paul, P. Greengard, C. Jennings, D. King, S. Havens, V. Chiley, J. Frankle, C. Blakeney, and J. P. Cunningham, LoRA learns less and forgets less. Transactions on Machine Learning Research, Aug. 2024, ISSN: 2835-8856., https://openreview.net/forum?id=aloEru2qCG.

28 D. Rathore, V. Kumar, C. Bansal, and A. Moitra, How much is too much? Exploring LoRA rank trade-offs for retaining knowledge and domain robustness. Proceedings of the 14th International Joint Conference on Natural Language Processing and 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2025), Mumbai, India, Dec. 2025, pg. 1003–1013., https://doi.org/10.18653/v1/2025.findings-ijcnlp.58.

29 Z. Song, J. Zhuo, Y. Yang, Z. Ma, S. Zhang, and X. Chen, LoRA-Whisper: Parameter-efficient and extensible multilingual ASR. Proceedings of Interspeech 2024, Kos, Greece, Sep. 2024, pg. 3934–3938., https://doi.org/10.21437/Interspeech.2024-892.

30 A. Prasad, S. Madikeri, D. Khalil, P. Motlicek, and C. Schuepbach, Speech and language recognition with low-rank adaptation of pretrained models. Proceedings of Interspeech 2024, Kos, Greece, Sep. 2024, pg. 2825–2829., https://doi.org/10.21437/Interspeech.2024-2187.

31 S. Mirzaei, J. B. Arzate, and Y. Vijay, Enhancing aviation communication transcription: Fine-tuning Distil-Whisper with LoRA. Journal of Aerospace Information Systems. Vol. 23, no. 7, pg. 590–596, July 2026., https://doi.org/10.2514/1.I011656.

32 K. K. R. Nareddy, S. Ternus, and J. Niebling, Analyzing and fine-tuning Whisper models for multilingual pilot speech transcription in the cockpit. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2025), Nashville, TN, USA, Jun. 2025., https://arxiv.org/abs/2506.21990.

33 L. Šmídl, J. Švec, D. Tihelka, J. Matoušek, J. Romportl, and P. Ircing, Air traffic control communication (ATCC) speech corpora and their use for ASR and TTS development. Language Resources and Evaluation. Vol. 53, no. 3, pg. 449–464, Sep. 2019., https://doi.org/10.1007/s10579-019-09449-5.

34 J. P. Zuluaga-Gomez, K. Veselý, I. Szöke, A. Blatt, P. Motlicek, M. Kocour, M. Rigault, K. Choukri, A. Prasad, S. S. Sarfjoo, I. Nigmatulina, C. Cevenini, P. Kolčárek, A. Tart, J. Černocký, and D. Klakow, ATCO2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications. Data-centric Machine Learning Research. Vol. 2, pg. 1–45, 2024.

35 J. Tol, ATC dataset. Hugging Face, 2024., https://huggingface.co/datasets/jacktol/atc-dataset.

36 V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, LibriSpeech: An ASR corpus based on public domain audio books. Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, Apr. 2015, pg. 5206–5210., https://doi.org/10.1109/ICASSP.2015.7178964.

37 J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.