Towards Understanding Data Requirements for Developing Automatic Speech Recognition Systems for Air Traffic Control

Stephen S. B. Clarke*, David Nielsen†, Charles I. Cutler‡, Aida Sharif-Rohani§, Krishna M. Kalyanam¶
NASA Ames Research Center, Moffett Field, CA, 94035, USA

In recent years, the application of automatic speech recognition has gained popularity across diverse industries, including aviation. Given the many applications focusing on transcribing air traffic control and management communication, this paper explores the training of OpenAI's Whisper model across multiple existing public and private air traffic control voice datasets in an effort to improve robustness. Combining roughly 60+ hours of various air traffic datasets, our goal is to train a unified Whisper model and expect an average word error rate reduction across testing datasets. Furthermore, this work aims to understand the data quantity requirements for achieving state-of-the art results by comprehensively training Whisper on varying dataset sizes. This work has the potential to improve automatic speech recognition performance across the domain, improve understanding of the quantity of data required by an aviation speech recognition system, and lastly provide metrics to compare and improve upon in future research.

I. Nomenclature

ASR	=	automatic speech recognition
$ATCo$	=	air traffic control
$ATCSCC$	=	Air Traffic Control System Command Center
ATM	=	air traffic management
NAS	=	National Airspace System
$PERTI$	=	Plan, Execute, Review, Train and Improve
WER	=	word error rate

II. Extended Abstract

Voice communication forms the backbone of commercial aviation operations. From national planning to controlling individual flights, speech is used to convey information quickly and accurately. Similarly, computer systems are relied upon in aviation to monitor safety and ensure efficient operation of the National Airspace System (NAS). As the computer systems in the aviation industry evolve, the integration of automatic speech recognition (ASR) technologies holds the promise of revolutionizing decision-making processes and contributing to the overall safety and efficiency of air traffic control (ATCo) and management (ATM).

The goal of this paper is twofold. Our first is to understand how the combination of ATCo voice from multiple sources, such as different world regions and airports, could improve robustness of ATCo ASR models. As shown by the recently developed Whisper model by OpenAI [1], combining ASR data between domains and languages can improve overall robustness of an ASR model. Furthermore, pooling data in low-resource languages has shown to improve the fine-tuning of Whisper models [2]. Given the limited quantity of high quality transcribed speech data in aviation research, this pooling has potential to improve performance and open opportunities to apply ASR in other sub-domains of aviation that have even less data available.

Secondly we want to understand how much data is required to develop a robust ASR system with the current state-of-the-art models. Although state-of-the-art ASR models have been trained on upwards of hundreds of thousands

*Aerospace Engineer, NASA Ames Research Center, stephen.s.clarke@nasa.gov, AIAA Member
†Senior Aerospace Research Engineer, KBR Inc., david.l.nielsen@nasa.gov
‡AI/ML Software Engineer, Metis Technology Solutions, charles.i.cutler@nasa.gov
§Aerospace Research Engineer, NASA Ames Research Center, aida.sharifrohani@nasa.gov
¶Senior Aerospace Research Engineer, NASA Ames Research Center, AIAA Associate Fellow

[1] to millions of hours of training data [3], the power of fine-tuning general-domain models like Whisper has shown success in adapting to low-resource languages like Greek [2] and Turkish [4]. And although the aviation industry uses English as a backbone, it contains specific terminology and phraseology which separates it from general conversation. Therefore it is a great candidate for fine-tuning from a generic English model.

ASR research in aviation is a quickly growing topic especially with recent trends in deep learning. Benchmarks on public ATCo datasets have been created using deep neural networks [5], as well as transformer models like Wav2Vec 2.0 [6]. More recently Whisper has been used to benchmark the public ATCo datasets and further fine-tuned achieving the state-of-the-art performance of 1.17% word error rate (WER) on a subset of the ATCOSIM dataset [7] through training on ATCOSIM data. However, with a WER of 13.46% on the ATCO2 dataset [7], there are still improvements to be made across the domain. This paper aims to extend the work done previously with Whisper to further understand its capability and how we can work towards even better ASR for the industry with a long term goal of 5% or lower WER across most datasets.

The remainder of this extended abstract is as follows: Section III introduces the benchmark datasets collected and used in this research, Section IV outlines the processing and training setup of the model as well as discuss the experiments to be conducted for the final paper, Section V shows some preliminary results of our training pipeline, and Section VI concludes the abstract with discussion of the results and work to be done for the final paper submission.

III. Data

Since the primary goal of this study is to train a unified ASR model for ATCo, our focus has been to collect as much audio and transcription data as is accessible to us. Throughout collection, we have categorized the data as public and private. A brief summary of all the collected data sources is provided in Table 1.

The few publicly available transcribed ATCo data sources include ATCOSIM [8], ATCO2-Test [9], and UWB-ATCC [10]. These data have been published throughout Europe and have served as benchmarks for most published research in ASR for ATCo [5] [6] [7].

For the private datasets, NASA has internally collected audio data and text transcriptions from various regions in the United States such as Tampa, Florida (KTPA) and Salt Lake City, Utah (KSLC). These airports give a wide sample of regional air traffic control accents as well as airport-specific phraseology. The audio was recorded from ground frequencies near or at the airports resulting in high signal-to-noise radio recordings. Afterwards, internal subject matter experts were tasked with creating ground-truth transcriptions from the recordings using a tool called Prodigy*. Prodigy simply provides an interface for the audio to be played in tandem with an input text box to create transcriptions.

In addition to ATCo audio transcriptions, NASA has also collected data from other aviation sub-domains notably ATM. We include two datasets collected from planning meetings at the Air Traffic Control System Command Center (ATCSCC†). More information on these is listed below. These recordings are more conversational in nature than structured ATCo commands, but we include them nonetheless to increase data quantity. Since both data sources have phraseology related to aviation, we experiment to see if the addition of the tangentially related data could improve ASR performance for ATCo speech and vice versa.

Lastly, there are also additional public data which are released under paid licenses and not included when training our model but are worth mentioning. These include the LDC-ATCC dataset with 26 hours of American English accented recordings [11], the full ATCO2 dataset which includes 5381 hours of training and 4 hours of testing data [9] and the HIWIRE database including 28 hours recorded in a simulation environment with French, Greek, Itallian, and Spanish accents [12].

A. Detailed Data Descriptions

1. ATCOSIM

The ATCOSIM dataset was developed in tandem by the Graz University of Technology (TUG) and Eurocontrol Experimental Centre (EEC). It includes ten (10) hours of ATCo data by ten different non-native english speakers. The data was recorded in a simulation environment with typical ATC conditions [8].

*https://prodi.gy/

†https://www.faa.gov/about/office_org/headquarters_offices/ato/service_units/systemops/nas_ops/atcscc

Dataset	Length in Hours	Publicly Available
ATCOSIM	10	✓
UWB-ATCC	20	✓
ATCO2-Test	1	✓
KTPA	10	
KSLC	2	
Webinar	28	
PERTI	4	

Table 1 Data Quantity

2. UWB-ATCC

Collected and annotated by the Department of Cybernetics at the University of West Bohemia, the UWB-ATCC dataset contains 20 hours of transcribed ATC communications in Czech airspace. The dataset was created for both speech-to-text and text-to-speech so it includes additional information aside from text such as pronunciation and speaker role. [13]

3. ATCO2-Test

The ATCO2 dataset was created by the Idiap Research Institute [9]. While the entire dataset contains audio and other natural language processing annotations, we are focused on the ATCO2-Test subset. This subset contains four (4) hours of audio-transcription pairs, 1.1 hours of which are released for free to the public. The data was collected from LKTB, LKPR, LZIB, LSGS, LSZH, LSZB and YSSY airports then transcribed by volunteers and paid annotators [9].

4. KTPA and KSLC

This private dataset, collected at NASA Ames Research Center, includes 10 hours of audio-transcription pairs from KTPA and 4 hours from KSLC. The audio was initially transcribed by a retired air traffic controller and further manually cleaned for data quality. The data was collected for research into ground instructions so they were recorded on ground frequencies. However, depending on the airport and time of day, the ground frequency was combined with the tower so the transcriptions include multiple types of instructions.

5. Air Route Traffic System Command Center

Although not ATCo data, this audio and transcription data was collected from planning telecons and PERTI (Plan, Execute, Review, Train and Improve) meetings at the ATCSCC in Warrenton Virginia. The planning telecons occur every two hours and lasts around 15 minutes. It brings together air traffic managers from facilities across the entire NAS to discuss terminal constraints, en route constraints, and mitigation strategies using TMIs. PERTI meetings are higher level and discuss strategies to improve management strategies for day to day operations. The structure of the audio is much closer to conversational English but it also contains a large quantity of aviation phraseology and terminology. This data is included in the study to understand how the addition of data from several aviation sub-domains could improve ASR in ATCo. In total there is 28 hours of transcribed planning telecon data and 4 hours of PERTI data, both transcribed by a retired air traffic controller.

IV. Methodology

Provided the ATCo voice and transcription data, we now discuss the methods for reaching the objectives of the paper. First being understanding how unifying the training data could improve overall WER, and second to understand the data quantity requirements for developing at-scale ASR systems within the aviation domain. Thus, two experiments are established.

A. Experimental Setup

1. Training a Unified Model

Following our motivation for training a unified model, the first experiment will involve combining the training datasets together. Through this combination we will have a corpus of nearly 60 hours for training after splitting into training, testing, and validation sets. After training and hyperparameter tuning, the unified model will be tested on each individual testing set as well as a combined testing set. This will reveal the benefit or lack thereof of combining the data and its impact on individual dataset performance. Our prediction is that the WER, on average, will be reduced across the testing sets.

2. Exploring Data Quantity Requirements

The second experiment will be to train the individual and unified models while varying the quantity of training data at 5%, 10%, 25%, 50%, and 100% of the entire training set and tested against the model's respective testing set. We expect the models with more training data to perform better overall. Furthermore, with an understanding of the correlation between input data quantity and WER we could describe the data quantity requirements for building robust ASR models that could be used in real world systems.

B. Model Training

The model training and evaluation pipeline is heavily influenced by the methodology used in [7], and thus by the python library HuggingFace [14]. HuggingFace provides streamlined documentation and pipelines used for efficiently training ASR models such as Whisper. In addition, it can be used to quickly compare our model results with those previously published to the HuggingFace Hub such as the models published by [6] and [7]. Furthermore, HuggingFace is useful in creating reproducible results which is essential in keeping training results between experiments consistent. The HuggingFace Trainer can easily be modified to change a single feature such as training data or model checkpoint without altering other aspects of the model or training pipeline.

A summary of our HuggingFace training pipeline is as follows:

1) The Whisper model processor and pretrained models are loaded from the HuggingFace Hub.
2) HuggingFace's Datasets[‡] library is used as a dataloader to ingest the audio filepaths and transcription texts from either a locally stored .tsv file or from the HuggingFace Hub. Additional processing also happens at this step such as setting the sample rate of the audio to 16k Hz for input into the Whisper model.
3) The WER metric is defined from HuggingFace's Evaluate[§] library.
4) A trainer is defined with all of the hyperparameters defined by the user and then training begins.

1. Hyperparameters

Hyperparameter tuning was performed to optimize the preliminary model results. These parameters were selected using the guess and check method. The remaining hyperparameters were left as the default chosen by the HuggingFace trainer. The hyperparameters are as follows:

- pretrained_model: medium.en
- num_proc: 40
- learning_rate: 6.25e-6
- warmup: 12
- train_batchsize: 32
- eval_batchsize: 16
- dropout: 0.1
- num_epochs: 50

For the experiments conducted in the final paper, these hyperparameters will remain static across all models and training datasets. This is to ensure consistency as the objective is to understand changes to the dataset quantity.

[‡] https://huggingface.co/docs/datasets/en/index

[§] https://huggingface.co/docs/evaluate/en/index

Training Dataset	Baseline WER %	Fine-Tuned WER %	ATCO2-Test WER %
LDC-ATCC	n/a	7.6	39.2
UWB-ATCC	n/a	14.6	26.4
ATCOSIM	n/a	3.0	44.3
KTPA	32.0	9.1	39.8
Webinar	18.277	6.513	n/a
PERTI	18.128	3.785	n/a

Table 2 Preliminary Results

2. Hardware

All model training has been done using the NASA Advanced Supercomputing Cabeus Supercomputer[¶]. In particular, we have trained using their Milan-A100 nodes which contain a 64-core host processor, 4 Nvidia A100 GPUs, and 512GB of RAM.

V. Preliminary Results

Table 2 shows the preliminary results of fine-tuning Whisper on several datasets. These serve as a baseline to ensure our training pipeline is getting consistent and accurate results. The results are consistent with the WERs reported in previous research [7]. In the final paper, we plan to report results on:

1) WER for a unified model with and without private datasets, as well as with and without the ATCSCC data. These models will be tested against each individual testing dataset along with a unified testing set.
2) WER for models trained on 5%, 10%, 25%, 50%, and 100% of total training data. These will include each individual private dataset, as well as the unified models. The models will be tested on static testing datasets for comparison.

VI. Conclusion

This paper will present a comprehensive exploration of the application of Whisper in ATCo communications. With the aim of fine-tuning a robust ATC model and understanding the data requirements for building a robust system, we hope to improve the overall performance and robustness of ASR within the broader aviation domain. Additionally, we will provide an updated benchmark of ASR performance metrics for public ATCo datasets using Whisper which can be used by future research to compare and improve upon.

References

[1] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I., "Robust speech recognition via large-scale weak supervision," *International conference on machine learning*, PMLR, 2023, pp. 28492–28518.

[2] Paraskevopoulos, G., Tsoukala, C., Katsamanis, A., and Katsouros, V., "The Greek podcast corpus: Competitive speech models for low-resourced languages with weakly supervised data," *arXiv preprint arXiv:2406.15284*, 2024.

[3] Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., Jansen, A., Xu, Y., Huang, Y., Wang, S., et al., "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 16, No. 6, 2022, pp. 1519–1532.

[4] Oyucu, S., "Comparing The Fine-Tuning and Performance of Whisper Pre-Trained Models for Turkish Speech Recognition Task," *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, IEEE, 2023, pp. 1–4.

[5] Zuluaga-Gomez, J., Motlicek, P., Zhan, Q., Vesely, K., and Braun, R., "Automatic speech recognition benchmark for air-traffic communications," *arXiv preprint arXiv:2006.10304*, 2020.

[¶] `https://www.nas.nasa.gov/hecc/resources/cabeus.html`

[6] Zuluaga-Gomez, J., Prasad, A., Nigmatulina, I., Sarfjoo, S. S., Motlicek, P., Kleinert, M., Helmke, H., Ohneiser, O., and Zhan, Q., "How does pre-trained wav2vec 2.0 perform on domain-shifted asr? an extensive benchmark on air traffic control communications," *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 205–212.

[7] van Doorn, J., Sun, J., Hoekstra, J., Jonk, P., and de Vries, V., "Whisper-ATC: Open Models for Air Traffic Control Automatic Speech Recognition with Accuracy," *International Conference on Research in Air Transportation*, 2024, pp. ICRAT–2024.

[8] Hofbauer, K., Petrik, S., and Hering, H., "The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech." *LREC*, Citeseer, 2008.

[9] Zuluaga-Gomez, J., Veselỳ, K., Szöke, I., Blatt, A., Motlicek, P., Kocour, M., Rigault, M., Choukri, K., Prasad, A., Sarfjoo, S. S., 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.

[10] Šmídl, L., "Air Traffic Control Communication," , 2011. URL `http://hdl.handle.net/11858/00-097C-0000-0001-CCA1-0`, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[11] Godfrey, J., "Air Traffic Control Complete LDC94S14A," , 1994.

[12] Segura, J., Ehrette, T., Potamianos, A., Fohr, D., Illina, I., Breton, P., Clot, V., Gemello, R., Matassoni, M., and Maragos, P., "The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication," *Online. http://www. hiwire. org*, 2007.

[13] Šmídl, L., Švec, J., Tihelka, D., Matoušek, J., Romportl, J., and Ircing, P., "Air traffic control communication (ATCC) speech corpora and their use for ASR and TTS development," *Language Resources and Evaluation*, Vol. 53, 2019, pp. 449–464.

[14] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M., "Transformers: State-of-the-Art Natural Language Processing," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.