Journal of Aviation/Aerospace Education & Research

Volume 33
Number 4 *Special Issue: Artificial Intelligence in Aviation*

Article 8

2024

Low-Resource Automatic Speech Recognition Domain Adaptation – A Case-Study in Aviation Maintenance

Nadine Amin M.S.
Purdue University, amin37@purdue.edu

Tracy L. Yother Ph.D.
Purdue University, tyother@purdue.edu

Julia Rayz Ph.D.
Purdue University, jtaylor1@purdue.edu

JAAER

Low-Resource Automatic Speech Recognition Domain Adaptation – A Case-Study in Aviation Maintenance

Nadine Amin [1a]**,Tracy L. Yother** [1b]**, Julia Rayz** [1c]
[1]Purdue University, IN 47907 USA
[a]amin37@purdue.edu, [b]tyother@purdue.edu, [c]jtaylor1@purdue.edu

Abstract

With timeliness and efficiency being critical in the aviation maintenance industry, the need has been growing for smart technological solutions that help in optimizing and streamlining the different underlying tasks (Bergkvist & Sabbagh, 2021). One such task is the technical documentation of the performed maintenance operations (Siyaev & Jo, 2021a). Instead of paper-based documentation, voice tools that transcribe spoken logbook entries allow technicians to document their work right away in a hands-free and time efficient manner. However, an accurate automatic speech recognition (ASR) model requires large training corpora (Siyaev & Jo, 2021a), which are lacking in the domain of aviation maintenance. In addition, ASR models which are trained on huge corpora in standard English, perform poorly in such a technical domain with non-standard terminology (Siyaev & Jo, 2021b). Hence, this study investigates the extent to which fine-tuning an ASR model, pre-trained on standard English corpora, on limited in-domain data improves its recognition performance in the technical domain of aviation maintenance. We present a case study on one such pre-trained ASR model, wav2vec 2.0 (Baevski et al., 2020). Results show that fine-tuning the model on a limited anonymized dataset of maintenance logbook entries significantly reduces its error rates when tested on not only an anonymized in-domain dataset, but also a non-anonymized one. This suggests that any available aviation maintenance logbooks, even if anonymized for privacy, can be used to fine-tune general-purpose ASR models, and enhance their in-domain performance. Lastly, an analysis on the influence of voice characteristics on model performance stresses the need for balanced datasets representative of the population of aviation maintenance technicians.

Keywords: *Automatic Speech Recognition, Aviation Maintenance Logbooks, Domain Adaptation*

Introduction

As serious safety risks are associated with aircraft part failures, efficiency and timeliness are crucial in the aviation maintenance, repair, and overhaul (MRO) industry (Bergkvist & Sabbagh, 2021). With the manual-in-nature work of aviation maintenance technicians (AMTs), their reliance on pen and paper (Amin et al., 2022) or even portable digital devices (Latib et al., 2023) for work procedures renders underlying tasks time-consuming and inefficient (Bergkvist & Sabbagh, 2021). To optimize MRO tasks, there has been a growing need for smart technological solutions (Bergkvist & Sabbagh, 2021). One such task is the technical documentation of executed operations (Chandola et al., 2022) in the form of maintenance logbook entries. Therefore, a voice tool that transcribes spoken logbook entries from AMTs would allow seamless, hands-free documentation while efficiently carrying out maintenance operations.

A voice transcription tool requires an accurate automatic speech recognition (ASR) model. To train such a model, large speech corpora are needed (Kleinert et al., 2018; Siyaev & Jo, 2021a; Srinivasamurthy et al., 2017). Yet, aviation is a low-resource domain lacking labeled speech corpora (Fan et al., 2023; Lin, Yang, Li, et al., 2021; Pellegrini et al., 2018). In addition, aviation maintenance logbooks contain technical, non-standard terminology (Akhbardeh, 2022; Siyaev & Jo, 2021a, 2021b). Thus, ASR models trained on large standard English corpora exhibit poor in-domain performance (Siyaev & Jo, 2021a, 2021b). To respond to these domain-specific challenges, we present a case study on the wav2vec 2.0 (Baevski et al., 2020) ASR model, pre-trained on out-of-domain data, and investigate the extent of its performance enhancement upon fine-tuning it on a limited speech dataset synthesized from textual maintenance logbook entries.

The rest of the article is organized as follows: we review the relevant literature, outline the adopted methodology, demonstrate and analyze the results, provide a general discussion of the findings and implications, and lastly summarize our conclusions and ideas for future work.

Literature Review

We present how ASR has been used in aviation, synthesizing literature from three main perspectives: the

subdomain in which ASR has been applied, the type of ASR model used, and the technique adopted to enhance the model performance in the low-resource domain of aviation.

Aviation Subdomains Employing ASR

Air traffic control (ATC) is the aviation subdomain with the most extensive ASR research (Badrinath & Balakrishnan, 2022; Fan et al., 2023; Kleinert et al., 2021; Kocour, Veselỳ, Szke, et al., 2021; Lin et al., 2019; Nigmatulina et al., 2022; Oualil et al., 2017). Communications between pilots and air traffic controllers are mainly verbal in nature (Badrinath & Balakrishnan, 2022; Lin, Yang, Li, et al., 2021). Thus, research has focused on accurately transcribing the exchanged messages (Lin, Li, et al., 2021; Lin, Yang, Li, et al., 2021; Srinivasamurthy et al., 2017; Zuluaga-Gomez et al., 2023), enhancing callsigns recognition (Guo et al., 2021; Kasttet et al., 2023; Kocour, Veselỳ, Blatt, et al., 2021; Nigmatulina et al., 2021; Zuluaga-Gomez et al., 2021, 2020), or realizing voice communications in training simulators (Cheng et al., 2015; Prasad et al., 2022). Fewer works are concerned with ASR in the subdomain of aviation MRO. They mainly use ASR systems in speech interaction modules in immersive maintenance simulators (He et al., 2017) or mixed reality training on maintenance operations (Siyaev Jo, 2021a, 2021b). Speech data, needed for training ASR models, is less abundant in aviation MRO than in ATC. This is evident in the relatively limited ASR research in the MRO subdomain.

Types of ASR Models Used

Different ASR model types are being adopted in aviation. Several works (Cheng et al., 2015; Zietsman & Malekian, 2022) still rely on hidden Markov models with no incorporation of deep neural networks (Cheng et al., 2015; Zietsman & Malekian, 2022). Many more adopt hybrid speech recognizers (Kocour, Veselỳ, Blatt, et al., 2021; Kocour, Veselỳ, Szke, et al., 2021; Nigmatulina et al., 2022; Ohneiser et al., 2021; Zuluaga-Gomez et al., 2021), and several others use end-to-end (E2E) ASR models (Fan et al., 2023; Lin, Li, et al., 2021; Lin, Yang, Guo, & Fan, 2021; Lin et al., 2019; Siyaev & Jo, 2021a; Zuluaga-Gomez et al., 2023). While E2E ASR models achieve state-of-the-art accuracies on various benchmarks and have different advantages over hybrid ones, they still struggle with domain adaptation, customization, and low-resource settings (**Li2022**). This hinders their default adoption in the aviation domain.

Adopted Low-Resource ASR Techniques

Besides being a technical domain with non-standard terminology (Akhbardeh et al., 2022; Siyaev & Jo, 2021a, 2021b), aviation is a low-resource domain lacking labeled speech corpora needed to train ASR models (Fan et al., 2023; Lin, Li, et al., 2021; Pellegrini et al., 2018). We discuss three main approaches explored in the literature to overcome these challenges: using (a) unlabeled in-domain speech data, (b) in-domain text data, and (c) out-of-domain data.

Especially in the ATC subdomain, speech data can be available but just not labeled (Šmídl et al., 2018; Srinivasamurthy et al., 2017, 2018). Thus, one approach is to leverage unlabeled in-domain speech data and adopt semi-supervised (Kleinert et al., 2018; Nigmatulina et al., 2022; Ohneiser et al., 2021; Srinivasamurthy et al., 2017; Zuluaga-Gomez et al., 2021), self-supervised (Lin, Yang, Guo, & Fan, 2021), or unsupervised (Lin, Li, et al., 2021; Lin, Yang, Guo, & Fan, 2021) techniques. Another approach is to use any available in-domain text data to (a) synthesize speech and train or adapt ASR models on the resulting speech-text pair (Siyaev & Jo, 2021a, 2021b), (b) train language models (LMs) incorporated into the ASR model (Kocour, Veselỳ, Szke, et al., 2021; Nigmatulina et al., 2021, 2022; Zuluaga-Gomez et al., 2020, 2021), and/or (c) customize ASR models with contextual information (**Guo2021**; Helmke et al., 2023; Kocour, Veselỳ, Blatt, et al., 2021; Kocour, Veselỳ, Szke, et al., 2021; Nigmatulina et al., 2021; Oualil et al., 2017; Zuluaga-Gomez et al., 2020, 2021). The third approach is to leverage the vast out-of-domain corpora through adapting pre-trained ASR models to the target domain of aviation. Fine-tuning E2E ASR models (Badrinath & Balakrishnan, 2022; Kleinert et al., 2021; Siyaev & Jo, 2021a; Zuluaga-Gomez et al., 2023) or acoustic models of hybrid ones (Oualil et al., 2017; Srinivasamurthy et al., 2017, 2018) has been explored. Some works have also adapted LMs using out-of-domain text data (Oualil et al., 2017; Srinivasamurthy et al., 2017).

Summary

Most of ASR research in aviation is in the ATC subdomain with only limited research in aviation MRO, the subdomain with which our research is concerned. E2E ASR models are the current state-of-the-art, but they face challenges in such a low-resource setting, which is what our research addresses. Guided by the low-resource ASR techniques adopted in aviation, this study opts for synthesizing speech from a limited text dataset and investigating the effect of fine-tuning an ASR model, pre-trained on out-of-domain data, on such a limited, synthetic speech corpus.

Methodology

We introduce the datasets used in fine-tuning and testing, the model architecture and variants, the adopted

performance evaluation metrics, as well as the overall experimental design.

Datasets

To the best of our knowledge, there are no publicly available speech datasets of aviation maintenance logbook entries. Thus, we opted for synthesizing speech datasets from text ones.

Text Datasets

Since maintenance logbooks are proprietary, publicly available aviation maintenance logbooks are limited (Akhbardeh, 2022). Consequently, we used a small text dataset of 6,169 anonymized instances from the University of North Dakota aviation program, open-sourced by the MaintNet library (Akhbardeh et al., 2020). Key names and numbers, which are inherently challenging to transcribe, had been removed during anonymization. Hence, for more realistic model testing, we also manually put together a small custom dataset of 45 non-anonymized instances from Purdue University Aviation Maintenance. The domain-specific terminology and natural language structure in the datasets were assumed to be representative of those in the aviation MRO domain. See Appendix A for example instances from both datasets.

Text Datasets Preprocessing

Maintenance logbook entries are free brief texts heavy on abbreviations and misspellings (Akhbardeh et al., 2022). Some instances are also cut off mid-word or mid-phrase. To appropriately preprocess the datasets, we had one of the authors who is a domain expert record herself while reading instances out loud (a) naturally expanding abbreviations, (b) correctly pronouncing misspelled words, and (c) either discarding or appropriately completing cutoff words and phrases. The instances were then accordingly cleaned. We make the cleaned MaintNet dataset as well as associated lists of abbreviations and misspellings publicly available1 to the research community. Lastly, all text was converted into uppercase, all punctuation marks except for apostrophes were removed, and numbers were converted into their spoken word form.

Speech Datasets

We used the Google Cloud Text-to-Speech (TTS) API to convert text into synthesized human-sounding speech. As of April 2023, the API supported 80 English language voices with multiple accents, genders, as well as types pertaining to the technology used to synthesize that output voice (see Table 1). There are four voice types: (a) Standard, synthesized using parametric speech synthesis (Gutkin, 2015), (b) WaveNet, a higher quality voice synthesized using DeepMind's generative model (Aharon, 2018), (c) Neural2, the highest quality voice generated using the same technology the API uses to create custom voices (Google Cloud, n.d.), and (d) Studio, a voice type for longer texts (Google Cloud, n.d.). In our experiments, the speaking speed randomly varied between 0.75 and 1.25, with 1 resembling normal speaking speed. The default pitch and volume of voices were used, and the sampling rate was set to 16 kHz.

For each of the MaintNet text instances, one of the 80 voices was chosen at random to synthesize the corresponding speech instance. After synthesis, 362 instances were discarded for they included words that were unknown to the Google Cloud TTS API and were hence spoken out one letter at a time. Of the remaining 5,807 MaintNet speech instances, 4,659 ($\approx$ 80%; $\approx$ 8.1 hrs; train-MNet) were used for fine-tuning the ASR model, and 1,148 ($\approx$ 20%; test-MNet) were used for testing it. For each text instance in the much smaller custom dataset, 27 speech instances were synthesized using different voices chosen at random from the 80 available ones. Hence, the final speech dataset contained 1,215 instances that were used to test the ASR model (test-cus).

Model

The wav2vec 2.0 model (Baevski et al., 2020) was chosen as it is one of the state-of-the-art E2E ASR models (Gandhi et al., 2022) that had also shown potential in the aviation domain (Siyaev & Jo, 2021a; Zuluaga-Gomez et al., 2023). It is trained to learn speech representations from large unlabeled speech corpora in a self-supervised framework and then fine-tuned on a smaller labeled speech corpus (Baevski et al., 2020) using a connectionist temporal classification (CTC) loss (Graves et al., 2006). We used two variants of wav2vec 2.0 that we introduce below.

Table 1

Number of Google Cloud Text-to-Speech English Language Voices per Type, Accent, and Gender (F: Female, M: Male) as of April 2023

Accent	Standard	WaveNet	Neural2	Studio	Total
American	10 (F: 5, M: 5)	14 (F: 7, M: 7)	9 (F: 5, M: 4)	2 (F: 1, M: 1)	35
Australian	4 (F: 2, M: 2)	7 (F: 4, M: 3)	4 (F: 2, M: 2)	-	15
British	5 (F: 3, M: 2)	12 (F: 6, M: 6)	5 (F: 3, M: 2)	-	22
Indian	4 (F: 2, M: 2)	4 (F: 2, M: 2)	-	-	8
Total	23	37	18	2	80

w2v2-base

This model served as a baseline trained on out-of-domain data but not further fine-tuned on an in-domain dataset. It is a 95M-parameters model, pre-trained and fine-tuned on the 960 hrs of 16 kHz standard English speech of the LibriSpeech training dataset (Panayotov et

al., 2015). We loaded this trained version of the wav2vec 2.0 model from the Hugging Face platform.

w2v2-avMRO

This is the model that we fine-tuned on in-domain data. We started with the *w2v2-base* model and then fine-tuned it on the train-MNet dataset. We froze the parameters of the feature encoder and fine-tuned the model for 1,800 steps (≈ 12 epochs). The learning rate was linearly increased from 0 to 1e–4 for a warm-up phase of 1,000 steps after which it linearly decayed. The AdamW optimizer (Loshchilov & Hutter, 2019) was used with default hyperparameter values. SpecAugment (Park et al., 2019) data augmentation was applied to the feature encoder outputs with default hyperparameter values. A vocabulary of size 32 was used: 26 tokens for the English alphabet, one for apostrophes, one for word boundaries, two for the start and end of sentences, one unknown token, and one padding token corresponding to CTC's blank token.

Evaluation Metrics

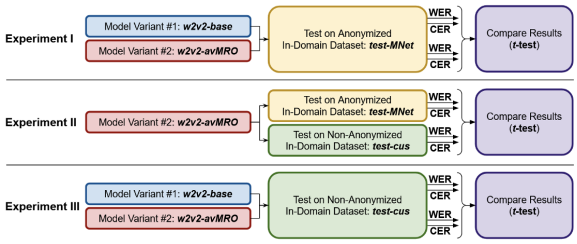
We adopted two of the most commonly used ASR performance evaluation metrics: word error rate (WER) (Badrinath & Balakrishnan, 2022; Kleinert et al., 2021; Lin et al., 2019; Siyaev & Jo, 2021a; Srinivasamurthy et al., 2017, 2018; Zuluaga-Gomez et al., 2023) and character error rate (CER) (Fan et al., 2023; Lin, Li, et al., 2021; Oualil et al., 2017; Siyaev & Jo, 2021a) . WER and CER are the percentages of words and characters, respectively, that are incorrectly transcribed by the model; the lower the rates, the more accurate the model is.

Experimental Design

To investigate the effect of fine-tuning the wav2vec 2.0 model on limited in-domain data, we designed three experiments (see Figure 1). Experiment I focused on the effect of fine-tuning alone, by comparing both models' performance on the test-MNet dataset. Since the fine-tuning dataset is anonymized, this choice of similarly anonymized testing dataset enabled isolating and assessing the effect of fine-tuning on performance. However, realistic maintenance entries are non-anonymized. Thus, by comparing the performance of *w2v2-avMRO* on the test-MNet and test-cus datasets, Experiment II assessed the effect on the fine-tuned model's performance when tested on an anonymized versus a non-anonymized dataset. Experiment III compared both models' performance on the test-cus dataset. Hence, it specifically analyzed the overall effect on performance when fine-tuning on an anonymized dataset but testing on a non-anonymized one. Next, we analyzed whether characteristics of the used synthetic voices influenced the fine-tuned model's performance on the test-cus dataset. We considered three characteristics: gender (Female, Male), accent (American, Australian, British, Indian), and speaking rate (low: at or below 0.9, normal: above 0.9 and below 1.1, high: at or above 1.1). Such analyses assessed model robustness across variations that would naturally appear in the population of AMTs.

Figure 1

Overview of Main Experiments



Note. WER = word error rate; CER = character error rate.

Results

We begin with the results of the three main experiments, summarized in Table 2. We then follow with the results of the analysis on voice characteristics and model performance.

Experiment I: Effect of Fine-Tuning (Anonymized Testing Dataset)

The WER ($t = 33.38$, $p < .001$, $r = .70$) and CER ($t = 26.53$, $p < .001$, $r = .62$) were significantly lower for the fine-tuned model. Table 3 shows example MaintNet instances correctly transcribed by *w2v2-avMRO* but incorrectly transcribed by *w2v2-base*. It is evident that without in-domain fine-tuning, the pre-trained model poorly predicted domain-specific terminology, like *rocker, gasket,* and *baffle*, as well as abbreviations, like *RPM, CHT,* and *EGT*.

Experiment II: Anonymized vs. Non-Anonymized Testing Dataset

The fine-tuned model exhibited significantly higher WER ($t = -33.33$, $p < .001$, $r = .62$) and CER ($t = -34.87$, $p < .001$, $r = .65$) for the non-anonymized test-cus dataset than the anonymized test-*MNet* one. Table 4 shows examples where *w2v2-avMRO* incorrectly transcribed custom instances. As is evident, the model could not properly transcribe the tube name, *Michelin*, and the reference manual name, *Hartzell*, since it had not seen such utterances during its fine-tuning. While it correctly predicted several part and serial numbers, the combination of numbers and

Table 2

Mean Performance of Models Across Testing Datasets

Experiment	Model	Testing Dataset	WER (95% CI)	CER (95% CI)
Experiment I	*w2v2-basew2v2-avMRO*	*test-MNet*	17.88% ± .98%1.22% ± .23%	4.02% ± .28% .26% ± .05%
Experiment II	*w2v2-avMRO*	*test-MNettest-cus*	1.22% ± .23% 9.94% ± .46%	.26% ± .05% 2.54% ± .12%
Experiment III	*w2v2-basew2v2-avMRO*	*test-cus*	20.52% ± .80%9.94% ± .46%	4.50% ± .21%2.54% ± .12%

Note. WER = word error rate; CER = character error rate; CI = confidence interval.

letters sometimes led the model to mistaken utterances like see for *C* and *H* for *eigh*. Domain-specific terminology, such as *gascolator*, and some abbreviations, like *AMM*, that did not appear in the fine-tuning dataset were also challenging for the model to correctly transcribe.

Experiment III: Effect of Fine-Tuning (Non-Anonymized Testing Dataset)

WER (t = 31.33, p < .001, r = .67) and CER (t = 21.17, p < .001, r = .52) were significantly lower for the fine-tuned model. Table 5 shows example custom instances that were correctly transcribed by *w2v2-avMRO* but incorrectly transcribed by *w2v2-base*. Both models were challenged by the custom instances being non-anonymized (refer to Table 4). However, the enhanced performance of the fine-tuned model was evident in its more accurate predictions when it comes to domain-specific terminology, such as *outboard*, *muffler*, and *camber shims*.

Voice Characteristics and Performance

Table 6 reports the performance of the fine-tuned model on the non-anonymized dataset per gender, accent, and speaking rate of the voice synthesized using the Google Cloud TTS API.

Gender

The WER and CER of the *w2v2-avMRO* model when transcribing female voices were, on average, higher than when transcribing male ones. However, these differences in WER (t = .89, p = .38, r = .03) and CER (t = .53, p = .60, r = .02) were not statistically significant. With the distribution of voice types being similar in female and male voices (see Appendix B), it is suggested that voice types did not influence the reported results of statistically insignificant differences in model performance across genders.

Table 3

Example MaintNet Instances (Correctly Transcribed by w2v2-avMRO) and their Corresponding Incorrect Transcriptions by w2v2-base

Instance	Transcription by *w2v2-base*
... SPEED SIXTEEN HUNDRED AND FIFTY **RPM** ADJUSTED **IDLE**	... SPEED SIXTEEN HUNDRED AND FIFTY **R P M** THE JUSTED **IDOL**
... **ROCKER COVER GASKET** LEAKS REMOVED AND REPLACED **GASKET**	... **ROCKA COVO GASCOT** LEAGUES REMOVED AND REPLACED **GASCET**
RIVET LOOSE AT **BAFFLE** AFT CENTER REPLACED **RIVET** WITH NEW	**RIVERT** LOSE AT **BAFFAL** AFT CENTRE REPLACED **RIVERT** WITH NEW
NUMBER TWO AND NUMBER FOUR **CHT EGT** WIRE ANCHOR BROKEN ...	NUMBER TWO IN NUMBER FOUR **C H T E G T** WIRE ANCHOR BROKEN ...

Accent

Results showed a statistically significant mean difference in WER (F(3, 1211) = 24.79, p < .001, ω = .23) and CER (F(3, 1211) = 26.65, p < .001, ω = .24) of *w2v2-avMRO* based on the accent of the transcribed voices. Post hoc tests showed that the average WER and CER were significantly higher when transcribed voices were of Indian accents compared to the other accents (Bonferroni, Hochberg, Games-Howell, p < .001; for all cases). Also, the average WER and CER (Bonferroni, Hochberg, Games-Howell, p < .01; for all cases) were significantly lower for American accents compared to Australian accents. The average WER was significantly lower for American accents compared to British ones (Bonferroni, Hochberg, Games-Howell, p < .05), but the mean difference in CER between them was statistically insignificant. There was also no statistically significant mean difference in error rates between Australian and British accents.

Table 4

Example Custom Instances and their Corresponding Incorrect Transcriptions by w2v2-avMRO

Instance	Transcription by *w2v2-avMRO*
INSTALLED NEW **MICHELIN** TUBE …	INSTALLED NEW **MISSIALAND** TUBE …
… SERIAL NUMBER **H** DASH K DASH ONE … IN ACCORDANCE WITH **HARTZELL** ENGINE TECHNOLOGIES … PART NUMBER OE DASH **A** TWO **SEE** COMPONENT RECORD FOR DETAILS	… SERIAL NUMBER **EIGH** DASH K DASH ONE … IN ACCORDANCE WITH **HARTSEL** ENGINE TECHNOLOGIES … PART NUMBER OE DASH TWO **C** COMPONENT RECORD FOR DETAILS
… IN ACCORDANCE WITH SR TWENTY **AMM** ZERO FIVE DASH THREE ZERO	… IN ACCORDANCE WITH SR TWENTY **AM M** ZERO FIVE DASH THREE ZERO
… CLEANED AND REINSTALLED FUEL **GASCOLATOR** SCREEN	… CLEANED AND REINSTALLED FUEL **GASKALATOR** SCREEN

Table 5

Example Custom Instances (Correctly Transcribed by w2v2-avMRO) and their Corresponding Incorrect Transcriptions by w2v2-base

Instance	Transcription by *w2v2-base*
REPLACED LEFT AND RIGHT **OUTBOARD** FUEL DIALS …	REPLACED LEFT AND RIGHT **UPBOARD** FUEL DIALS …
REPLACED RIGHT HAND **MUFFLER** HEAT EXCHANGER WITH NEW …	REPLACED RIGHT HAND **MUFFALER** HEAT EXCHANGER WITH NEW …
… MAIN LANDING GEAR NEGATIVE **CAMBER SHIMS** PART NUMBER …	… MAIN LANDING GEAR NEGATIVE **CAMBERSHIM'S** PART NUMBER …
REPLACED UPPER **PUCK PAN** ASSEMBLY PART NUMBER ONE …	REPLACED UPPER **PUCKPAN** ASSEMBLY PART NUMBER ONE …

As evident in Table 1 (also see Appendix B), all accents but the Indian one have voices of the Neural2 type category; the technology resulting in the highest quality voices (Google Cloud, n.d.). This contributed in part to the lower model performance reported for the Indian accent. In addition, the performance differences between accents could be partly attributed to the fact that the total number of voices provided by the Google Cloud TTS API was different across accents (see Table 1). Accordingly, accents with a greater number of voices appeared more in the model's fine-tuning dataset, which resulted in the model being better trained at transcribing them.

Table 6

Mean Performance of the Fine-Tuned w2v2-avMRO Model for the Non-Anonymized test-cus Dataset Across Characteristics of the Voices Synthesized Using Google Cloud Text-to-Speech

Characteristic	Level	*n*	WER (95% CI)	CER (95% CI)
Gender	Female	652	10.13% ± .62%	2.57% ± .15%
	Male	563	9.72% ± .69%	2.51% ± .18%
Accent*	American	539	8.33% ± .62%	2.20% ± .15%
	Australian	216	11.30% ± 1.16%	2.79% ± .31%
	British	344	9.93% ± .89%	2.44% ± .22%
	Indian	116	14.93% ± 1.46%	4.00% ± .44%
Speaking Rate*	Low	379	9.43% ± .79%	2.37% ± .19%
	Normal	451	9.50% ± .72%	2.43% ± .19%
	High	385	10.96% ± .88%	2.85% ± .23%

Note. WER = word error rate; CER = character error rate; CI = confidence interval. *p <.05.

Speaking Rate

There was a statistically significant mean difference in WER ($F(2, 1212) = 4.40$, $p = .01$, $\omega = .08$) and CER ($F(2, 1212) = 6.03$, $p < .01$, $\omega = .09$) of *w2v2-avMRO* based on the speaking rate of transcribed voices. Post hoc analyses suggested that mean error rates were significantly higher when voices had high speaking rates (Bonferroni, Hochberg, Games-Howell, $p < .05$; for all cases). There was no statistically significant mean difference in error rates for voices with low and normal speaking rates. The distribution of voice types was similar in voices with different speaking rates (see Appendix B), suggesting that voice types did not affect reported results pertaining to differences in model performance across speaking rates.

Discussion

This study investigated the extent to which fine-tuning an ASR model on limited in-domain data improves its performance in the aviation MRO domain. Results showed that fine-tuning the model lowered its error rates when tested on the anonymized, in-domain MaintNet dataset. With the model also fine-tuned on MaintNet instances, this boost in performance might be attributed in part to the degree of similarity between the fine-tuning and testing datasets. Nevertheless, when tested on the non-

anonymized custom dataset, the model's error rates still decreased because of being fine-tuned on the anonymized data. With ASR systems of WERs below 10% deemed suitable for industry (Urban et al., 2023), the fine-tuned model's performance seemed promising. Results suggest that an ASR model that accurately transcribes spoken aviation maintenance logbook entries could be attained by leveraging general-purpose ASR models, pretrained on vast non-technical out-of-domain data, and further fine-tuning them on the limited in-domain data available.

Non-anonymized logbook entries are inherently more challenging for an ASR model to transcribe than anonymized ones. Results demonstrated that the fine-tuned model still struggled with transcribing names of used tubes and reference manuals, part and serial numbers, as well as domain-specific terminology and abbreviations that it had not seen during fine-tuning. Nevertheless, results showed that the model, fine-tuned on only anonymized logbook entries, was still better at transcribing non-anonymized ones than a model that had only been trained on out-of-domain data. With the low-resource nature of the domain and maintenance logbooks being proprietary (Akhbardeh et al., 2022), this suggests that any available logbook data, even if anonymized for privacy purposes, could be used to fine-tune off-the-shelf ASR models, and enhance their in-domain performance.

The gender of the transcribed voice was not found to significantly affect the model's performance. However, the performance significantly degraded when the speaking speed was high. The voice's accent also significantly affected the performance. This was nevertheless related to the technologies used to synthesize voices of different accents, and the number of voices per accent provided by Google Cloud TTS. This sheds light on the importance of having a more balanced fine-tuning dataset that is representative of accents in the population of AMTs.

Conclusion

This study responds to the growing demand for aviation-specific corpora and language processing tools (Amin et al., 2022) by providing a gold standard dataset with lists of abbreviations and misspellings based on MaintNet's aviation maintenance dataset https://github.com/nadine-amin/Cleaned-MaintNet-Aviation-Maintenance-Dataset (Akhbardeh et al., 2020). Also, to the best of our knowledge, this study is the first to assess an ASR model's performance in transcribing spoken aviation maintenance logbook entries. It adds to the limited literature concerned with enhancing ASR in the technical, low-resource domain of aviation MRO. Accordingly, this study paves the way for several future research directions, including:

- **Leveraging Out-of-Domain Maintenance Data:** Inspired by Akhbardeh et al. (2022), we suggest assessing whether fine-tuning an ASR model on maintenance logbooks from other domains, such as the automotive MRO domain, allows the model to better learn common maintenance-related terminology and enhances its performance in aviation MRO.
- **Leveraging Institute-Specific Contextual Information:** An institute can use its own lists of reference manuals, part numbers, and so forth to enhance the ASR model performance in transcribing institute-specific non-anonymized logbook entries. Such lists can be used to either train an external LM or a context-aware encoder (Guo et al., 2021)
- **More Realistic Model Testing:** Further tests are needed to assess whether fine-tuning the model on a synthetic speech dataset would still enhance its performance on a real speech dataset. The model should also be tested in the presence of possible noise in the real environment of AMTs.

References

Aharon, D. (2018). Introducing cloud text-to-speech powered by deepmind wavenet technology [Google Cloud]. https://cloud.google.com/blog/products/ai-machine-learning/introducing-cloud-text-to-speech-powered-by-deepmind-wavenet-technology

Akhbardeh, F. (2022). *Nlp and ml methods for preprocessing, clustering and classification of technical logbook datasets* [Doctoral dissertation, Rochester Institute of Technology] [[Unpublished doctoral dissertation]].

Akhbardeh, F., Desell, T., & Zampieri, M. (2020). Maintnet: A collaborative open-source library for predictive maintenance language resources. *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*. https://doi.org/10.18653/v1/2020.coling-demos.2

Akhbardeh, F., Zampieri, M., Alm, C. O., & Desell, T. (2022). Transfer learning methods for domain adaptation in technical logbook datasets. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Amin, N., Yother, T., Johnson, M., & Rayz, J. (2022). Exploration of natural language processing (nlp) applications in aviation. *The Collegiate Aviation Review International*, *40*(1), 203–216. https://doi.org/10.22488/okstate.22.100211

Badrinath, S., & Balakrishnan, H. (2022). Automatic speech recognition for air traffic control com-

munications. *Transportation Research Record*, *2676*(1), 798–810. https://doi.org/10.1177/03611981211036359

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 12449–12460. https://doi.org/10.48550/arXiv.2006.11477

Bergkvist, E., & Sabbagh, T. (2021). *Smart future solutions for maintenance of aircraft: Enhancing aircraft maintenance at saab ab* [Master's thesis, Linköping University]. http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-176561

Chandola, D. C., Jaiswal, K., Verma, S., & Singh, B. (2022). Aviation mro: A comprehensive review of factors affecting productivity of aircraft maintenance organization. *Proceedings of the 2022 Advances in Science and Engineering Technology International Conferences (ASET)*, 1–7. https://doi.org/10.1109/ASET53988.2022.9734808

Cheng, V. H. C., Nguyen, J. N., Ballinger, D. B., Cowart, S. E. C., Fong, A. F., Jones, S. J., & Lu, H.-L. L. (2015). A speech-enabled simulation interface agent for airspace system assessments. *Proceedings of the AIAA Modeling and Simulation Technologies Conference*. https://doi.org/10.2514/6.2015-0148

Fan, P., Hua, X., Lin, Y., Yang, B., Zhang, J., Ge, W., & Guo, D. (2023). Speech recognition for air traffic control via feature learning and end-to-end training. *IEICE Transactions on Information and Systems*, *E106.D*(4), 538–544. https://doi.org/10.1587/transinf.2022EDP7151

Gandhi, S., Von Platen, P., & Rush, A. M. (2022). Esb: A benchmark for multi-domain end-to-end speech recognition [arXiv Preprint]. https://doi.org/10.48550/arXiv.2210.13352

Google Cloud. (n.d.). Standard, wavenet, neural2, and studio voices — cloud text-to-speech documentation [Retrieved February 27, 2023]. https://cloud.google.com/text-to-speech/docs/wavenet

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning*, 369–376. https://doi.org/10.1145/1143844.1143891

Guo, D., Zhang, Z., Fan, P., Zhang, J., & Yang, B. (2021). A context-aware language model to improve the speech recognition in air traffic control. *Aerospace*, *8*(11), 348. https://doi.org/10.3390/aerospace8110348

Gutkin, A. (2015). Text-to-speech for low-resource languages (episode 2): Building a parametric voice [Google AI Blog]. https://ai.googleblog.com/2015/12/text-to-speech-for-low-resource.html

Helmke, H., Kleinert, M., Linß, A., Motlicek, P., Wiese, H., Klamert, L., Harfmann, J., Cebola, N., Arilíusson, H., & Simiganoschi, T. (2023). The haawaii framework for automatic speech understanding of air traffic communication. *SESAR Innovation Days 2023*, 1–9. https://doi.org/10.61009/SID.2023.1.04

Kleinert, M., Helmke, H., Siol, G., Ehr, H., Cerna, A., Kern, C., Klakow, D., Motlíček, P., Oualil, Y., Singh, M., & Srinivasamurthy, A. (2018). Semi-supervised adaptation of assistant based speech recognition models for different approach areas. *Proceedings of the 2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*, 1–10. https://doi.org/10.1109/DASC.2018.8569879

Kleinert, M., Venkatarathinam, N., Helmke, H., Ohneiser, O., Strake, M., & Fingscheidt, T. (2021). Easy adaptation of speech recognition to different air traffic control environments using the deepspeech engine. *11th SESAR Innovation Days, Virtual*. https://elib.dlr.de/145397/

Kocour, M., Veselỳ, K., Blatt, A., Zuluaga-Gomez, J., Szke, I., Černockỳ, J., Klakow, D., & Motlíček, P. (2021). Boosting of contextual information in asr for air-traffic call-sign recognition. *Interspeech*, 3301–3305. https://doi.org/10.21437/Interspeech.2021-1619

Kocour, M., Veselỳ, K., Szke, I., Kesiraju, S., Zuluaga-Gomez, J., Blatt, A., Prasad, A., Nigmatulina, I., Motlíček, P., Klakow, D., Tart, A., Atassi, H., Kolčárek, P., Černockỳ, H., Cevenini, C., Choukri, K., Rigault, M., Landis, F., Sarfjoo, S., & Salamin, C. (2021). Automatic processing pipeline for collecting and annotating air-traffic voice communication data. *Engineering Proceedings*, *13*(1), 8. https://doi.org/10.3390/engproc2021013008

Latib, M. S. A., Balakrishnan, P., & Derus, S. R. M. (2023). Design and development of aircraft maintenance manual smart reader at politeknik banting selangor. *IV. ASC-2022/Fall Congress Hosted By - Change Shaping the Future: Proceeding Book*, 439.

Lin, Y., Deng, L., Chen, Z., Wu, X., Zhang, J., & Yang, B. (2019). A real-time atc safety monitoring framework using a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, *21*(11), 4572–4581. https://doi.org/10.1109/TITS.2019.2940992

Lin, Y., Li, Q., Yang, B., Yan, Z., Tan, H., & Chen, Z. (2021). Improving speech recognition models with small samples for air traffic control systems.

Neurocomputing, *445*, 287–297. https://doi.org/10.1016/j.neucom.2020.08.092

Lin, Y., Yang, B., Guo, D., & Fan, P. (2021). Towards multilingual end-to-end speech recognition for air traffic control. *IET Intelligent Transport Systems*, *15*(9), 1203–1214. https://doi.org/10.1049/itr2.12094

Lin, Y., Yang, B., Li, L., Guo, D., Zhang, J., Chen, H., & Zhang, Y. (2021). Atcspeechnet: A multilingual end-to-end speech recognition framework for air traffic control systems. *Applied Soft Computing*, *112*, 107847. https://doi.org/10.1016/j.asoc.2021.107847

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations*.

Nigmatulina, I., Braun, R., Zuluaga-Gomez, J., & Motlíček, P. (2021). Improving call-sign recognition with air-surveillance data in air-traffic communication [arXiv Preprint]. https://doi.org/10.48550/arXiv.2108.12156

Nigmatulina, I., Zuluaga-Gomez, J., Prasad, A., Sarfjoo, S., & Motlíček, P. (2022). A two-step approach to leverage contextual data: Speech recognition in air-traffic communications. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6282–6286. https://doi.org/10.1109/ICASSP43922.2022.9746563

Ohneiser, O., Sarfjoo, S., Helmke, H., Shetty, S., Motlíček, P., Kleinert, M., Her, H., & Murauskas, S. (2021). Robust command recognition for lithuanian air traffic control tower utterances. *Interspeech*, 3291–3295. https://doi.org/10.21437/Interspeech.2021-935

Oualil, Y., Klakow, D., Szaszák, G., Srinivasamurthy, A., Helmke, H., & Motlíček, P. (2017). A context-aware speech recognition and understanding system for air traffic control domain. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 404–408. https://doi.org/10.1109/ASRU.2017.8268964

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech*. https://doi.org/10.21437/Interspeech.2019-2680

Pellegrini, T., Farinas, J., Delpech, E., & Lancelot, F. (2018). The airbus air traffic control speech recognition 2018 challenge: Towards atc automatic transcription and call sign detection. *Interspeech*. https://doi.org/10.21437/Interspeech.2019-1962

Siyaev, A., & Jo, G.-S. (2021a). Neuro-symbolic speech understanding in aircraft maintenance metaverse. *IEEE Access*, *9*, 154484–154499. https://doi.org/10.1109/ACCESS.2021.3128616

Siyaev, A., & Jo, G.-S. (2021b). Towards aircraft maintenance metaverse using speech interactions with virtual objects in mixed reality. *Sensors*, *21*(6), 2066. https://doi.org/10.3390/s21062066

Šmídl, L., Švec, J., Pražák, A., & Trmal, J. (2018). Semi-supervised training of dnn-based acoustic model for atc speech recognition. *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018*, 646–655. https://doi.org/10.1007/978-3-319-99579-3_66

Srinivasamurthy, A., Motlíček, P., Himawan, I., Szaszak, G., Oualil, Y., & Helmke, H. (2017). Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control. *Interspeech*. https://doi.org/10.21437/Interspeech.2017-1446

Srinivasamurthy, A., Motlíček, P., Singh, M., Oualil, Y., Kleinert, M., Ehr, H., & Helmke, H. (2018). Iterative learning of speech recognition models for air traffic control. *Interspeech*, 3519–3523. https://doi.org/10.21437/Interspeech.2018-1447

Urban, E., Vilaysom, S., Farley, P., Andreas, M., Alexey, L., Orlov, D., Coulter, Dempsey, J., Erhopf, & Christiani, T. (2023). Test accuracy of a custom speech model - speech service - azure cognitive services [Retrieved May 11, 2023]. https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data?pivots=speech-studio

Zietsman, G., & Malekian, R. (2022). Modelling of a speech-to-text recognition system for air traffic control and nato air command. *Journal of Internet Technology*, *23*(7), 1527–1539. https://doi.org/10.53106/160792642022122307008

Zuluaga-Gomez, J., Nigmatulina, I., Prasad, A., Motlíček, P., Veselỳ, K., Kocour, M., & Szke, I. (2021). Contextual semi-supervised learning: An approach to leverage air-surveillance and untranscribed atc data in asr systems. *Interspeech 2021*. https://doi.org/10.21437/Interspeech.2021-1373

Zuluaga-Gomez, J., Prasad, A., Nigmatulina, I., Sarfjoo, S., Motlíček, P., Kleinert, M., Helmke, H., Ohneiser, O., & Zhan, Q. (2023). 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), 205–212. https://doi.org/10.1109/SLT54892.2023.10022724

Zuluaga-Gomez, J., Veselỳ, K., Blatt, A., Motlíček, P., Klakow, D., Tart, A., Szke, I., Prasad, A., Sarfjoo, S., Kolčárek, P., Kocour, M., Černockỳ, H., Cevenini, C., Choukri, K., Rigault, M., & Landis, F. (2020). Automatic call sign detection: Matching air surveillance data with air traffic spoken communications. *Proceedings of the 8th OpenSky Symposium*, *59*, 14. https://doi.org/10.3390/proceedings2020059014

A Text Datasets

This appendix provides a brief description of the used text datasets along with example instances.

MaintNet Dataset

MaintNet's aviation maintenance logbook dataset contains 6,169 instances from the University of North Dakota aviation program (Akhbardeh, 2022) collected over the years 2012 to 2017. Owing to the confidentiality and sensitivity of information in the logbook (Akhbardeh et al., 2022), the open-sourced version of the dataset is anonymized (Akhbardeh et al., 2022). Table 7 shows examples of instances from MaintNet's aviation maintenance dataset. Each instance contains a description of the maintenance problem that had happened and the corresponding action that had been taken, as recorded by either a mechanic or a pilot (Akhbardeh et al., 2022).

Custom Dataset

The custom dataset is a manually put together non-anonymized dataset containing sensitive information like part numbers, serial numbers, oil types, and names of reference manuals. It has 45 instances obtained from logbooks provided by Purdue University Aviation Maintenance. Table 8 shows example instances from this custom dataset.

Table 7

Example Instances from MaintNet's Aviation Maintenance Dataset

ID	Instance (Problem + Action)
100101	OIL FOUND ON ENTIRE RIGHT SIDE OF FUSELAGE. REMOVED COWLING, CLEANED ENGINE, PERFORMED ENGINE RUN U
100102	R/H ENGINE #4 CYL BAFFLE BOLT MISSING. INSTALLED NEW BOLT.
100103	R/H ENGINE #1 & 4 ROCKER COVERS LEAKING. REMOVED & REPLACED R/H ROCKER COVER GASKETS, 1 & 4.
100104	L/H ENGINE #1 & 3 ROCKER COVERS LEAKING. REMOVED & REPLACED L/H ENGINE ROCKER COVER GASKETS, 1 &
100105	ENGINE DIED DURING MAG CK. STARTED A/C USING FLOODED START PROCEDURE & RAN A/C TO

Table 8

Example Instances from Our Custom Dataset

Instance
Removed tube from nose wheel. Installed new Michelin tube P/N 092-308-0 Production lot number 20/09. Balanced nose wheel assy and installed.
Complied with ELT Inspection IAW FAR 91.207(d). ELT Battery due 5/2016. Performed Operational/Functional check IAW SR20 AMM 05-30.
Alternator #2 (p/n: 653344, s/n: H-K-171023) 500-hour inspection complied with IAW Hartzell Engine Technologies Aircraft Alternators & Starters Overhaul Manual p/n: OE-A2 (see component record for details).
Replaced all o-rings P/N M83461/1-222 in both brake calipers. Installed new temperature indicators P/N 51698-001 and 51698-003 on each caliper. Bled brakes. Operational check satisfactory.

B Synthetic Voice Types and Model Performance

This appendix reports the performance of the fine-tuned *w2v2-avMRO* model for the non-anonymized in-domain testing dataset (*test-cus*) across types of the synthetic voices generated using the Google Cloud TTS API. Results are broken down according to the gender (Table 9; Figure 2), accent (Table 10; Figure 3), and speaking rate (Table 11; Figure 4) of the synthetic voices.

Table 9

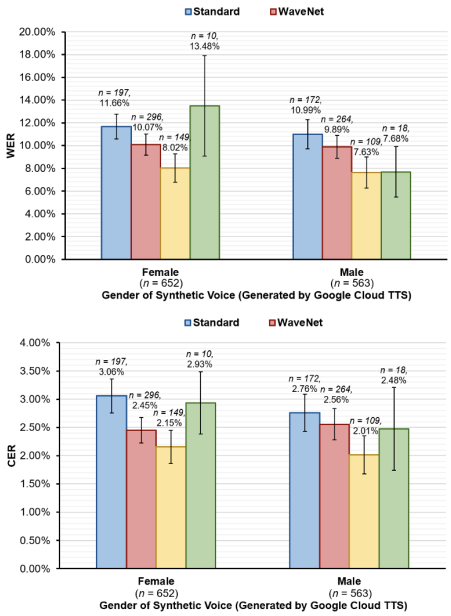
Mean Performance of the Fine-Tuned w2v2-avMRO Model for the Non-Anonymized test-cus Dataset Across Types and Genders of the Voices Synthesized Using Google Cloud Text-to-Speech

Gender	Type			
	Standard	**WaveNet**	**Neural2**	**Studio**
	n			
Female	197	296	149	10
Male	172	264	109	18
WER (95% CI)				
Female	11.66% ± 1.09%	10.07% ± 0.94%	8.02% ± 1.25%	13.48% ± 4.41%
Male	10.99% ± 1.28%	9.89% ± 1.03%	7.63% ± 1.36%	7.68% ± 2.22%
CER (95% CI)				
Female	3.06% ± 0.30%	2.45% ± 0.22%	2.15% ± 0.29%	2.93% ± 0.55%
Male	2.76% ± 0.33%	2.56% ± 0.27%	2.01% ± 0.34%	2.48% ± 0.73%

Note. WER = word error rate; CER = character error rate; CI = confidence interval.

Figure 2

Mean Performance of the Fine-Tuned w2v2-avMRO Model for the Non-Anonymized test-cus Dataset Across Types and Genders of the Voices Synthesized Using Google Cloud Text-to-Speech (TTS)



Note. Top panel: WER (95% CI). Bottom panel: CER (95% CI). WER = word error rate; CER = character error rate; CI = confidence interval.

Table 10

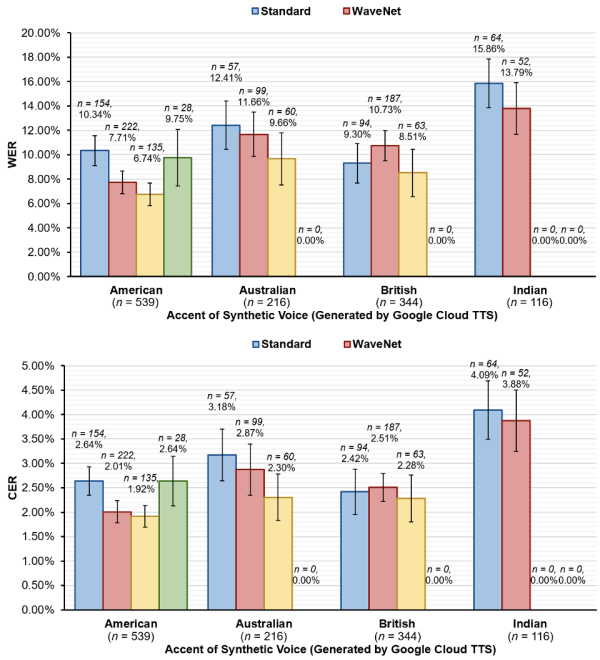
Mean Performance of the Fine-Tuned w2v2-avMRO Model for the Non-Anonymized test-cus Dataset Across Types and Accents of the Voices Synthesized Using Google Cloud Text-to-Speech

Accent	Type			
	Standard	**WaveNet**	**Neural2**	**Studio**
American	154	222	135	28
Australian	57	99	60	-
British	94	187	63	-
Indian	64	52	-	-
WER (95% CI)				
American	10.34% ± 1.23%	7.71% ± 0.93%	6.74% ± 0.93%	9.75% ± 2.33%
Australian	12.41% ± 1.98%	11.66% ± 1.83%	9.66% ± 2.14%	-
British	9.30% ± 1.62%	10.73% ± 1.25%	8.51% ± 1.94%	-
Indian	15.86% ± 2.00%	13.79% ± 2.13%	-	-
CER (95% CI)				
American	2.64% ± 0.29%	2.01% ± 0.22%	1.92% ± 0.22%	2.64% ± 0.51%
Australian	3.18% ± 0.53%	2.87% ± 0.53%	2.30% ± 0.48%	-
British	2.42% ± 0.47%	2.51% ± 0.29%	2.28% ± 0.48%	-
Indian	4.09% ± 0.60%	3.88% ± 0.63%	-	-

Note. WER = word error rate; CER = character error rate; CI = confidence interval.

Figure 3

Mean Performance of the Fine-Tuned w2v2-avMRO Model for the Non-Anonymized test-cus Dataset Across Types and Accents of the Voices Synthesized Using Google Cloud Text-to-Speech (TTS)



Note. Top panel: WER (95% CI). Bottom panel: CER (95% CI). WER = word error rate; CER = character error rate; CI = confidence interval.

Table 11

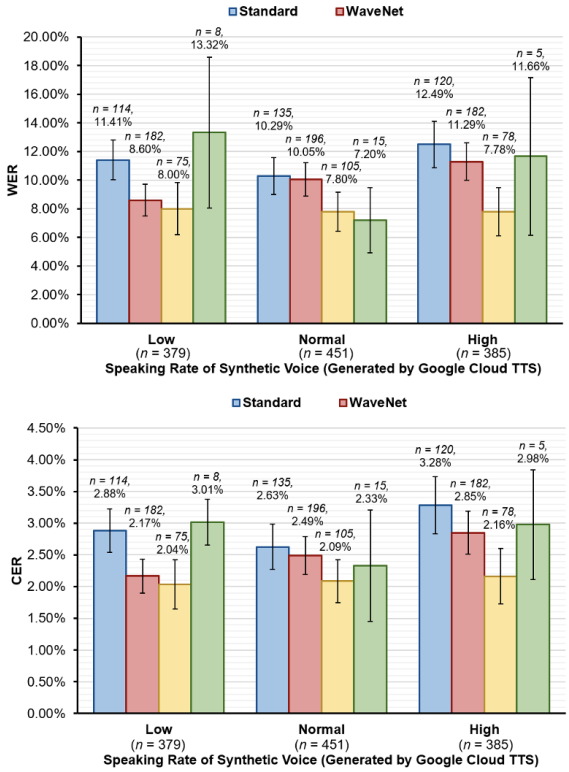
Mean Performance of the Fine-Tuned w2v2-avMRO Model for the Non-Anonymized test-cus Dataset Across Types and Accents of the Voices Synthesized Using Google Cloud Text-to-Speech (TTS)

Speaking Rate	Type			
	Standard	WaveNet	Neural2	Studio
Low	114	182	75	8
Normal	135	196	105	15
High	120	182	78	5
WER (95% CI)				
Low	11.41% ± 1.39%	8.60% ± 1.10%	8.00% ± 1.83%	13.32% ± 5.29%
Normal	10.29% ± 1.29%	10.05% ± 1.17%	7.80% ± 1.37%	7.20% ± 2.27%
High	12.49% ± 1.63%	11.29% ± 1.30%	7.78% ± 1.68%	11.66% ± 5.51%
CER (95% CI)				
Low	2.88% ± 0.34%	2.17% ± 0.27%	2.04% ± 0.39%	3.01% ± 0.36%
Normal	2.63% ± 0.35%	2.49% ± 0.30%	2.09% ± 0.34%	2.33% ± 0.88%
High	3.28% ± 0.45%	2.85% ± 0.34%	2.16% ± 0.43%	2.98% ± 0.87%

Note. WER = word error rate; CER = character error rate; CI = confidence interval.

Figure 4

Mean Performance of the Fine-Tuned w2v2-avMRO Model for the Non-Anonymized test-cus Dataset Across Types and Speaking Rates of the Voices Synthesized Using Google Cloud Text-to-Speech (TTS)



Note. Top panel: WER (95% CI). Bottom panel: CER (95% CI). WER = word error rate; CER = character error rate; CI = confidence interval.