HOW DOES PRE-TRAINED WAV2VEC 2.0 PERFORM ON DOMAIN-SHIFTED ASR? AN EXTENSIVE BENCHMARK ON AIR TRAFFIC CONTROL COMMUNICATIONS

Juan Zuluaga-Gomez [⋆,†,‡], *Amrutha Prasad* [†,¶], *Iuliia Nigmatulina* [†], *Seyyed Saeed Sarfjoo*[†], *Petr Motlicek*[†,¶], *Matthias Kleinert*[§], *Hartmut Helmke*[§], *Oliver Ohneiser*[§], *Qingran Zhan*[‖]

[†] Idiap Research Institute, Martigny, Switzerland
[‡] Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland
[¶] Brno University of Technology, Brno, Czech Republic
[§] German Aerospace Center (DLR), Institute of Flight Guidance, Braunschweig, Germany
[‖]School of Information and Electronics, Beijing Institute of Technology, Beijing, China

ABSTRACT

Recent work on self-supervised pre-training focus on leveraging large-scale unlabeled speech data to build robust end-to-end (E2E) acoustic models (AM) that can be later fine-tuned on downstream tasks e.g., automatic speech recognition (ASR). Yet, few works investigated the impact on performance when the data properties substantially differ between the pre-training and fine-tuning phases, termed domain shift. We target this scenario by analyzing the robustness of Wav2Vec 2.0 and XLS-R models on downstream ASR for a completely unseen domain, air traffic control (ATC) communications. We benchmark these two models on several open-source and challenging ATC databases with signal-to-noise ratio between 5 to 20 dB. Relative word error rate (WER) reductions between 20% to 40% are obtained in comparison to hybrid-based ASR baselines by only fine-tuning E2E acoustic models with a smaller fraction of labeled data. We analyze WERs on the low-resource scenario and gender bias carried by one ATC dataset.

Index Terms— Automatic speech recognition, Wav2Vec 2.0, self-supervised pre-training, air traffic control communications.

1. INTRODUCTION

A lot of recent work on end-to-end (E2E) acoustic modeling including automatic speech recognition (ASR) exploits self-supervised learning (SSL) of speech representations [1] including autoregressive models [2, 3] and bidirectional models [4, 5]. Self-supervised learning is a training technique capable of leveraging large-scale unlabeled speech to develop robust acoustic models [4, 6]. In fact, [7] explores a way to perform ASR without any labeled data in a complete unsupervised fashion. In a standard setup, E2E models trained by SSL are later fine-tuned on downstream tasks with much fewer labeled samples compared to standard supervised learning. By applying SSL, these systems have dramatically improved ASR performances on English speech datasets [4], such as LibriSpeech [8]. Similarly, performance on cross-lingual speech recognition is largely improved by SSL [9, 10]. It can be inferred that SSL-based pre-training allows models to capture a good representation of acoustics, which can be leveraged across different languages for ASR.

This work reviews the robustness of two well-known E2E acoustic models trained by SSL (i.e., Wav2Vec 2.0 and XLS-R) on a completely unseen domain: air traffic control (ATC) communications. ATC deals with aircraft guidance in the air and on the ground via voice communications between air traffic controllers (ATCOs) and pilots. These communications are well-defined by grammar and vocabulary [11] that must be followed to provide a safe and reliable flow of air traffic while keeping operating costs as low as possible. Despite the interest of ASR for ATC, there is not a fully functional ASR engine on the market due to: (i) lack of performance, i.e., under 5% WER is required,[1] and (ii) lack of large-scale annotated speech (less than 50 hrs of open-source databases) and its high production cost makes it almost impractical [13].

1.1. Contribution and motivation

Only a few previous works intended to measure the effect of domain mismatch between pre-training and fine-tuning phases of E2E models [14]. However, we can still categorize all databases as either read, spontaneous or conversational speech. On the contrary, ATC speech does not fit in any of these three categories due to its uniqueness, e.g., ruled by a very well-defined grammar. Our contributions[2] cover the domain mismatch scenario by answering the three questions below.

(i) How robust pre-trained E2E models are on new domains like ATC? Our results (see Table 2) ratified that E2E models pre-trained by SSL (e.g., Wav2Vec 2.0) learn a strong representation of speech. Fine-tuning on a downstream task (e.g., ASR) is computationally less expensive than flat-start training, and it requires less in-domain data to achieve on par WERs compared to traditional hybrid-based ASR systems. We also perform experiments with multilingual E2E models, i.e., XLS-R [10]. We hypothesize that pre-trained multilingual models perform better on ATC speech data that contains accented English (i.e., LiveATC-Test and ATCO2-Test sets). Potentially, due to the strong speech representation acquired during SSL

[⋆]Corresponding author: juan-pablo.zuluaga@idiap.ch

[1] There is a clear threshold between enhancing ATCOs' productivity and delaying them in their tasks due to poor ASR, see [12].

[2] Our code is stored in the following public GitHub repository: `https://github.com/idiap/w2v2-air-traffic`

978-1-6654-7189-3/22/$31.00 ©2023 IEEE

phase, which translates into a more accent-agnostic AM.

(ii) How much in-domain ATC labeled data is needed to fine-tune an E2E model that reaches on-par performance with regard to hybrid-based models? We perform a comparative study ranging from 5 minutes (few-shot learning) to ∼15 hrs of labeled speech (i.e., from 100 to 15k utterances). In addition, we investigate the impact of integrating an in-domain language model with beam search rather than using simple greedy decoding. With the aim of open science and fostering research in ATC, we also introduce baselines[3] on two well-known public ATC corpora.

(ii) How robust are E2E models on speakers with different gender? E2E models such as Wav2Vec 2.0 have experienced exponential interest in the research community. However, little investigation has been carried out about estimating the WERs gap produced by gender disparities. To name a few [15, 16, 17]. In this work, we perform an analysis by fine-tuning E2E models with ATC audio from different genders. We study this on a free and public ATC corpus where gender labels are provided.

We believe this work is impactful because the ASR field is advancing in an outpacing manner, where each month many large-scale speech models (LSSM) are poured into the research field with outstanding performances in well-known corpora, e.g., LibriSpeech [8]. Nonetheless, little has been examined in many other domains, such as ATC communications. Thus, it is of particular interest to evaluate and assess the performance of these LSSM on *'lagged'* fields.

2. RELATED WORK

Robust ASR stands as a promising tool for aiding ATCOs and ATC, in several ways. For example, reducing ATCOs' workload [18] by automatizing several of their daily tasks.[4] Another by-product of introducing ASR tools in ATC is the increase in airspace safeness and reduction of environmental impact caused by ATC operations. This is why the European Union (EU) has been funding different projects intended to bring closer speech and text-based technologies to ATC. MALORCA project concluded that ATCOs's workload can be reduced significantly by integrating ASR systems, while boosting their efficiency [19]. ATCO2[5]is also a well-known project that developed a pipeline [20] for automatic collection and pre-processing of large quantities of ATC audio data. Their main focus covered downstream task such as ASR [21], named-entity recognition [22], and acoustic and text-based speaker role recognition [23, 24]. HAAWAII[6] project develops a reliable and adaptable solution to automatically transcribe voice utterances issued by both ATCOs and pilots. Still, all previous research only investigates either standard supervised or semi-supervised [25, 26, 27] hybrid-based ASR systems.

Information uttered in ATC communications contain a diverse set of special entities, also called named entities. Some examples are callsigns, values and units. The most important and critical is the callsign, composed of an *airline designator*, a set of numbers and letters, e.g., `TVS12AB` spelled as *"SKYTRAVEL ONE TWO ALFA BRAVO*. The correct recognition of such key entities is crucial, as further it is used to extract target information from the conversations to assist ATCOs. Thus, it is almost mandatory for ASR engines to provide considerable low WERs. Additionally, the communications are mostly carried over noisy audio channels, usually below 15 dB SNR, which is the default in ATC environments. Taking into account these considerations, the *'ideal ASR engine'* should aim at preventing error propagation to the fullest extent, which can turns into misleading information passed to sub-systems at the next stages.

We redirect the reader to a general overview on spoken instruction understanding for ATC in [28] and latest work on hybrid-based ASR for ATC in [20, 22]. In [13] contextual information (also known as contextual biasing) via n-grams composition (in the HCLG graph[7]) is merged with semi-supervised learning techniques to further decrease word error rates (WER) on an ASR designed for ATC. Boosting of contextual knowledge during and after decoding has also been explored in [29, 30, 31, 32], where a set of target n-grams are added to further decrease WERs.

Despite the recent success of mixing SSL acoustic pre-training on E2E architectures for ASR, there has not been a comparative study between traditional hybrid-based and E2E acoustic modeling targeted to ATC. First, hybrid-based ASR modeling is based on a disjoint optimization of separate models i.e., AM, LM and a lexicon (e.g., phoneme-based). State-of-the-art (SOTA) models are trained with lattice-free maximum mutual information (LF-MMI) loss [33] which relies on alignments produced by a previously trained HMM-GMM model [33]. Second, E2E systems model AM and LM jointly, and they are mostly trained with connectionist temporal classification (CTC) loss [34] (enabling alignment-free training). In [35], it is compared CTC and LF-MMI adaptation of pre-trained models. Recently, attention-based (e.g., Transformers) have become the *de facto* choice for AM [4, 10, 36]. However, only few studies focused on domain shift during pre-training and fine-tuning or the impact of noisy speech on AM [37]. For instance, [14, 38, 39] perform experiments similar to ours, addressing the domain-shift scenario between pre-training and fine-tuning phases. Yet, these databases still fall into read, spontaneous or conversational speech.

3. DATASETS AND EXPERIMENTAL SETUP

This research experiments with seven datasets in the English language with various accents, speech rate, and data quality. With the aim of encouraging open research on ATC,[8] we experimented with four public databases,[9] as referenced in Table 1. To the author's knowledge, this is the first work that open sources code in the field of robust ASR targeted to ATC.

3.1. Private databases

NATS and ISAVIA: the audio data is collected and annotated by air navigation service providers (ANSPs) for HAAWAII project. The two datasets are, (i) London approach (NATS) and (ii) Icelandic en-route (ISAVIA). In total, there are 32 hrs of manually transcribed data for training and 2 hrs for testing. Both datasets are cataloged as good quality speech sampled at 8 kHz. Further details in Table 1.

[3] Our code is stored in the following public GitHub repository: `https://github.com/idiap/w2v2-air-traffic`

[4] One example is to detect named entities from voice communications. These entities are later parsed into ATCOs workstations. Using ASR reduces the overall latency of this simple, yet important procedure.

[5] `https://www.atco2.org/`

[6] `https://www.haawaii.de`

[7] In hybrid-based ASR, the different knowledge sources are represented via weighted finite-state transducers (WFST) and then merged in a final decoding graph i.e., *HCLG*. H, C, L and G are the hidden Markov models, context dependency, lexicon, and LM or grammar, correspondingly.

[8] General research on ASR for ATC has lagged behind due to privacy clauses and contracts. Ongoing and former projects prohibit code release.

[9] See the GitHub repository for further information and baselines.

LiveATC-Test: the test set is gathered from LiveATC[10] data recorded from publicly accessible VHF radio channels, as a part of ATCO2 project [13, 32], and includes pilot and ATCO recordings with accented English from airports located in U.S., Czech Republic, Ireland, Netherlands, and Switzerland. We consider LiveATC-Test as low quality speech data set i.e., signal-to-noise (SNR) ratios goes from 5 to 15 dB [22]. Audio is sampled at 16 kHz.

3.2. Public databases

ATCO2-Test: evaluation set released at Interspeech 2021 by [13, 30]. The dataset contains a mix of noisy and heavily English accented recordings from seven different airports located in Australia, Czech Republic, Slovakia, and Switzerland. The first version of the ATCO2-Test set contains 1.1 hrs of speech, it is open-source and can be accessed for free in `https://www.atco2.org/data`. The full corpus is available for purchase through ELDA in `http://www.elra.info/en/catalogues/`. We only use the open-source version for reproducibility. The recordings of both corpus are mono-channel sampled at 16kHz and 16-bit PCM. This is the first study that evaluates E2E ASR for ATCO2-Test. The WERs listed in this paper could be adopted as baselines for future research.

LDC-ATCC: public ATC corpus gathered from three different airports.[11] LDC-ATCC corpus comprises recorded speech that aims to support research in robust ASR. The recordings contain several speakers and gathered over noisy channels. The dataset is formatted in NIST Sphere format, where full transcripts, start and end times of each transmission are provided. The audio files are sampled at 8 kHz, 16-bit PCM [40].

UWB-ATCC: free public ATC corpus containing recordings of communication between ATCOs and pilots. The speech is manually transcribed and labeled with speaker roles. The audio data is single channel sampled at 8 kHz. This dataset can be downloaded for free in their website[12] [41].

ATCOSIM: free public database for research on ATC communciations. It consists of 10 hrs of speech data recorded during ATC real-time simulations using a close-talk headset microphone. The utterances are in English language and pronounced by ten non-native speakers. The database includes orthographic transcriptions and additional information about speakers and recording sessions. This dataset can be downloaded for free in their website[13] [42].

For all of our experiments, we up-sample all recordings to 16 kHz. Additionally, there are not any official train/dev/test splits for LDC-ATCC, UWB-ATCC and ATCOSIM databases. Therefore, we split them following the proportions in Table 1. We also make sure that there is no speaker or utterance overlaps between each subset.

3.3. Automatic speech recognition

Our experimental setup is split into three parts, which aims to answer each of the questions raised in the Section 1. Initially, we assess WERs of several E2E models when fine-tuned with ATC audio. We define two training datasets, i) 32 hrs of annotated data from NATS and ISAVIA database and ii) 132 hrs of ATC speech data from different projects (including all the training data from Table 1), and we redirect the reader to [13] for further details. For now on, we refer to these datasets as *32 hrs* and *132 hrs* 'fine-tuning sets'. Later, we evaluate the low-resource scenario by fine-tuning E2E models with different amount of data, for this, we use NATS and ISAVIA as private databases, and LDC-ATCC and UWB-ATCC as public databases. Finally, we evaluate the performance shift by fine-tuning E2E models with audio data from different genders. Here, we employ the free and open-source ATCOSIM database. Finally, we release training scripts to replicate the results of the last two experiments (only for the public databases).

Table 1. Characteristics of public and private databases in the domain of air traffic control communications. We only list databases used in the experiments, see [13] for a more exhaustive list. †baseline performance of our state-of-the-art hybrid-based ASR model for ATC communications.

	Characteristics		
Dataset	Train / Test	SNR [dB]	WER [%]†
Private databases			
NATS	18h / 0.9h	≥20	7.7
ISAVIA	14h / 1h	15-20	12.5
LiveATC-Test	- / 1.8h	5-15	35.8
Public databases			
ATCO2-Test	- / 1.1h	10-15	24.7
LDC-ATCC	23h / 2.6h	10-15	-
UWB-ATCC	10.4h / 2.6h	≥20	-
ATCOSIM	8h / 2.4h	≥20	-

Baseline hybrid-based ASR: all experiments are conducted with Kaldi toolkit [43]. The baseline models are composed of six convolution layers and 15 factorized time-delay neural network (∼31M trainable parameters). We follow the standard Kaldi's chain LF-MMI training recipe [33]. The input features are high-resolution MFCCs with online cepstral mean normalization. The features are extended with i-vectors. We use 3-gram ARPA LM during decoding. The model is trained for 5 epochs on 132 hrs of ATC speech (that includes NATS and ISAVIA). Further information and baseline performances can be found in our previous work [13, 21, 22]. SOTA WERs are listed in the last column of Table 1.

End-to-end ASR: we report results on four configurations of Wav2Vec 2.0/XLS-R models. From now on, we refer to these models with the following tags: i) *w2v2-B:* BASE model (95M parameters, pre-trained on train-set 960 hrs LibriSpeech [8]); ii) *w2v2-L:* LARGE-960h model (317M parameters pre-trained and then fine-tuned with LibrSpeech 960 hrs train-set); iii) *w2v2-L-60K:* LARGE-960h-LV60K model (same as w2v2-L but uses LibriSpeech + 60k hrs from LibriVox project i.e., Libri-Light [44] during the pre-training phase); iv) *w2v2-XLS-R:* XLS-R model (300M parameters pre-trained on 436k hrs of publicly available data in 128 languages [10]). We fetched all models' checkpoints from HuggingFace platform [45, 46]. Later, we perform standard fine-tuning with ATC speech data.

Hyperparameters end-to-end ASR: all experiments use the same set of hyperparameters. The feature extractor is frozen throughout the fine-tuning phase. We fine-tune each model for 10 k steps, with a 500-step warm-up phase (∼5% of total updates). Learning rate is increased linearly until $\gamma = 1\text{e}{-4}$ during

[10] Streaming audio platform that gathers VHF ATC communications.

[11] `https://catalog.ldc.upenn.edu/LDC94S14A`

[12] `https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0001-CCA1-0.`

[13] `https://www.spsc.tugraz.at/databases-and-tools`

warm-up, then it linearly decays. We use CTC loss function [34]. Dropout [47] is set to $dp = 0.1$ for the attention and hidden layers. We use GELU activation function [48] and AdamW [49] optimizer (β_1=0.9, β_2=0.999, ϵ=1e$-$8). We fine-tune each model on an NVIDIA GeForce RTX 3090 with an effective batch size of 72 (batch size of 24, gradient accumulation of 3). All the models use a character-based lexicon, i.e., we concatenate the English alphabet with some symbols and the blank symbol (standard in CTC-based AM). In total, our vocabulary is composed of 32 characters, thus, we append a linear layer of this dimension on top of the pre-trained w2v2/XLS-R AMs to generate logits at each time step. We use greedy decoding after applying Softmax to obtain the most likely character at each time step.

Language model (LM): we concatenate all text transcripts and train 2/3/4-gram ARPA LMs. The LMs are integrated by shallow fusion with a Python based CTC decoder, `PyCTCDecode`. [14] 4-gram LMs performed systematically better ($\sim$2% relative WER reduction) compared to 2-gram LMs in all test sets. We report results only with 4-gram LM as in [4]. We set $\alpha = 0.5$ and $\beta = 1.5$, which corresponds to the LM and length normalization weights. We set the beam size to 100.

Data augmentation: we apply a data augmentation strategy similar to SpecAugment [50] is applied. We mask the input sequence with a probability $p = 0.075$, and $M = 12$ consecutive frames. These hyperparameters follow closely the original Wav2Vec 2.0 implementation [4].

3.4. Incremental training

With the recent success of SSL pre-trained E2E models, it has become of particular interest to quantify how much data is actually needed to perform effectively on a downstream task. It is also important for low-resource tasks, such as ATC, where few tens of hours of labeled data are available for training or fine-tuning. In most ATC cases, data from one airport does not generalize well to other airports (for instance, see Table 3) due to a considerable AM domain-shift (accent, speaker rates and audio quality), as well as a LM domain-shift (dominance of different vocabulary). We analyze model performance versus different fine-tuning data sizes. We experimented with four *few-shot learning* scenarios with less than one hour ($\sim$1k utterances) of fine-tuning data. We split the experiments in two. First, we fine-tuned nine models on private databases, either NATS or ISAVIA data, as depicted on the left plot of Figure 1 (x-axis refers to number of utterances used during fine-tuning in log scale). Second, with the aim of open research, we performed the same approach on public databases, i.e., LDC-ATCC and UWB-ATCC. The results are on the right plot of Figure 1.

3.5. Gender experiments

We use the free and open-source ATCOSIM database to carry the gender experiments. We obtained the gender labels for each utterance from the original ATCOSIM gold annotations (check our public GitHub repository for more details). We split the train set into increasing sizes of 1h, 2h, 3h, 3.5h, and also by gender. We aim at both, analyzing the performance in WERs caused by fine-tuning an E2E with audio from different gender, and to measure the performance gain by scaling up the fine-tuning data. We trained four models for each gender (using the same hyperparameters as the ones described in Section 3.3 and with the same model: *w2v2-L-60k*) and report the results in Table 4.

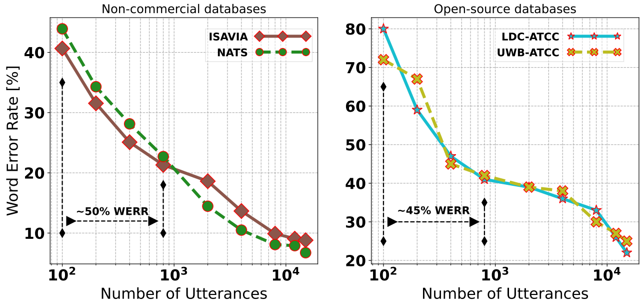

Fig. 1. Word error rates (WER) in percentages (%) for models fine-tuned with different amounts of data (x-axis). The left plot covers the results for private databases, while the right plot the public ones. Each data point corresponds to a train/test subset from the same dataset. 100, 1k and 10k utterances are roughly 5 min (few-shot), 1 h, and 10 hrs, respectively. All the evaluations are reported with *w2v2-L-60K* model and without explicit language model. We also list the Word Error Rate Reduction (WERR) by scaling up the fine-tuning set size from 100 to 800 samples.

4. RESULTS AND DISCUSSION

We structure the discussion of the results by addressing concrete questions. Our main hypothesis is that E2E models trained by SSL learn a robust representation of speech [4] and perform well on downstream tasks, i.e., ASR or multilingual ASR [10].

Breaking the paradigm, hybrid-based or E2E ASR? Although hybrid-based ASR modeling has been the default for several years, a new wave of E2E architectures pre-trained by SSL for joint AM and LM is taking its place. We compare E2E models to our best hybrid-based ASR trained with the 132 hrs fine-tuning set on Kaldi (**Baseline**, first row, Table 2). For E2E AMs we select two models. First, *w2v2-L-60k* to evaluate NATS and ISAVIA test sets, which is only fine-tuned on the 32 hrs set, i.e., in-domain data. Second, *w2v2-XLS-R+* for ATCO2-Test and LiveATC-Test test sets, which is trained on 132 hrs of ATC speech data [13, 21]. The 132 hrs set is a more diverse set, and it was also used to train the hybrid-based baseline model. We obtained 30 and 41% relative word error rate reduction (WERR) on NATS and ISAVIA when using *w2v2-L-60k* instead of our hybrid-based ASR baseline. The improvement is considerable, even though the baseline model is trained on four times more data than *w2v2-L-60k* (see Table 2). Similarly, *w2v2-XLS-R+* (last row: Table 2) surpasses the hybrid-based model on all four test sets, but more significantly on the two most challenging, ATCO2-Test and LiveATC-Test sets. In total, 19 and 30% relative WERR on ATCO2-Test and LiveATC-Test were obtained, respectively (hybrid-based $\rightarrow$ *w2v2-XLS-R+*).

However, it is worth mentioning that hybrid-based ASR is still considered the default in many industrial applications due to some advantages over E2E models. Two examples are, hybrid-based ASR does not require high-performance computing (e.g., GPUs) to perform real-time inference, while E2E models relies heavily on GPUs for speed. Further, hybrid-based ASR can be easily adapted to streaming scenarios with minimum degradation on WERs. Yet, E2E models still involve considerable architectural modifications to reach on-par WERs [51, 52, 53].

[14] `https://github.com/kensho-technologies/pyctcdecode`

Table 2. Word error rates (WER) in percentages (%) on four ATC test sets. Each model is fine-tuned on NATS and ISAVIA data (∼32 hrs). WERs are reported with greedy decoding or beam search decoding with a 4-gram ARPA LM integrated by shallow fusion. Unlabeled data column: *LS* stands for LibriSpeech 960 hrs train-set [8], *LV* for LibriVox 60k hrs train-set [44] and *ML* for 436k hrs of multilingual speech data [10]. *reports the baseline WER of Wav2Vec 2.0 (Table 1 from [4]) and XLS-R (Table 11 from [10]) models on LibriSpeech `test-other` set when only fine-tuned on 10 hrs of labeled data (comparable to our setup). †best Kaldi hybrid-based model (see [13, 21]) trained with the same amount of data as ††. ††models fine-tuned with 132 hrs of ATC speech data (instead of 32 hrs) and twice the number of steps, i.e., 20k. Numbers in **bold** refer to top WERs for models fine-tuned with the 32 hrs set and underline with 132 hrs set.

	Unlabeled	**NATS**		**ISAVIA**		**ATCO2-Test**		**LiveATC-Test**		LS*
Model (num. params.)	**data**	Greedy	+LM	Greedy	+LM	Greedy	+LM	Greedy	+LM	-
Baseline (31M)										
Hybrid-based †	-	-	7.7	-	12.5	-	24.7	-	35.8	-
BASE (95M)										
w2v2-B	LS	10.7	8.4	12.5	10.1	45.6	40.1	48.1	42.2	7.8
LARGE (371M)										
w2v2-L	LS	9.3	7.6	11.7	9.5	44.9	40.0	47.5	41.4	6.1
w2v2-L-60k	LS+LV	**6.8**	**5.4**	**8.8**	**7.3**	34.6	31.2	39.8	34.5	4.9
w2v2-L-60k+††	LS+LV	9.3	7.4	11.2	9.1	23.3	21.2	31.1	27.2	-
XLS-R (300M)										
w2v2-XLS-R	ML	8.4	6.5	10.5	8.2	39.1	33.8	42.9	36.7	15.4
w2v2-XLS-R+††	ML	9.0	7.4	10.4	8.3	**22.8**	**19.8**	**29.7**	**24.9**	-

Does additional partly-in-domain data increases ASR performance? We answer this question by comparing models fine-tuned either on the 132 hrs or 32 hrs set. The former set is a mix of public and private databases, while the latter is only NATS + ISAVIA, thus private. Note that NATS and ISAVIA are clean in-domain ATC speech corpora, i.e., considered as in-domain on the 32 hrs set and partly-in-domain otherwise (132 hrs set). Differently, ATCO2-Test and LiveATC-Test can be considered noisy and partly out-of-domain sets, i.e., airport, acoustic, and LM mismatch.

To address this question, we only focus on *w2v2-L-60k* and *w2v2-L-60k+* models fine-tuned on the 32 hrs and 132 hrs sets, respectively.[15]. We analyze the WERs obtained by greedy decoding to focus only on joint acoustic and language ASR modeling (see Section 2). A degradation on WERs is observed for the in-domain test sets, NATS: 6.8% → 9.3% WER and ISAVIA: 8.8% → 11.2% WER. This is mainly to the addition of data that does not match NATS and ISAVIA. Contrary, there was considerable WERR on the partly out-of-domain sets, ATCO2-Test: 34.6% → 23.3% WER and LiveATC-Test 39.8% → 31.1% WER. NATS test set (ISAVIA: 1% relative WERR) was impacted by the addition of partly-in-domain data, i.e., ∼7% relative lower WERs. Nevertheless, challenging test sets improved dramatically, i.e., ATCO2-Test and LiveATC-Test 43% and 33% relative WERR.

Do multilingual pre-trained E2E models help? To answer this question we compare *w2v2-L-60k+* and *w2v2-XLS-R+* models, which use the same hyperparameters, fine-tuning setup and beam search decoding with LM. We obtain a relative WERR of 8.8%, 6.6% and 8.5% on ISAVIA, ATCO2-Test and LiveATC-Test, respectively (no improvement on NATS). Significant improvement is seen on the most challenging test sets (SNR: 5-10 dB) which contain accented English speech, i.e., ATCO2-Test and LiveATC-Test. Hence, multilingual pre-trained models bring a tiny, but noticeable boost in performance compared to single-language pre-trained E2E models. This observed behavior can be attributed to the fact that *w2v2-XLS-R* have seen considerably more multilingual and accented audio data during the pre-training phase [10] in comparison to *w2v2-L-60k* [4].

The labeling process of ATC speech data can be decreased by a factor of two by only performing pre-labeling of audio with an in domain ASR system. Following this idea, we believe that is of special interest for the research community[16] to understand and analyze how much audio data is needed to reach acceptable WERs. We validated this idea by performing experiments with different amounts of fine-tuning data, thus it is up to the interested party to define the ‘acceptable’ WER threshold for the given application (e.g., deployment or pre-labeling).

How much data do you need to fine-tune Wav2Vec 2.0 and XLS-R models? We also investigate the effect on WERs when different amounts of fine-tuning data are used during the fine-tuning phase. We divide the by public and private databases. The WERs on the private databases are in the left plot of Figure 1. All the experiments are based on the most robust E2E model from Table 2 i.e., *w2v2-L-60K*.[17] The WERs plot are obtained with greedy decoding and no LM or explicit textual information added. We fine-tune 18 models varying the training data set (either NATS or ISAVIA) and varying the amount of fine-tuning samples. We initially tested the few-shot learning scenario (‘worse-case’), where only 100 labeled utterances (∼5 min) were used for fine-tuning, and achieved WERs of 40% and 43.9% for ISAVIA and NATS. Further, ∼50% relative WERR is obtained by scaling up the fine-tuning data to 50 minutes (800 utterances). Specifically, NATS 43.9% → 22.7% WER and ISAVIA 40.6% → 21.3% WER. Lastly, if all available data (∼14 hrs) is used, we reach an 8.8% and 6.8% WER for ISAVIA and NATS, respectively. This represents an ∼80% relative WERR compared to the low-resource setup (100 utterances). With around

[15] Note that the results are still comparable for the XLS-R AM, i.e., *w2v2-XLS-R* versus *w2v2-XLS-R+*.

[16] Likewise, it is of special interest for the ATC community.

[17] We select the *best model* based on lowest WERs on out-of-domain test sets, i.e., ATCO2-Test and LiveATC-Test. See last row Table 2.

Table 3. Word error rates (WER) in percentages (%) on different test sets. Models are fine-tuned only on public databases and fixed to 11 hrs of audio data. All systems are *w2v2-L-60k* and WERs are obtained with greedy decoding and no LM. †test set split by gender (Male/Female).

	Test set			
Train set	LDC	UWB	ATCO2	ATCOSIM (M/F)†
LDC-ATCC	**25.0**	64.1	58.7	41.1 / 35.7
UWB-ATCC	54.6	**21.9**	**47.9**	**32.5 / 24.6**

8 hrs (∼8000 utterances) *w2v2-L-60K* beats the performance of our SOTA hybrid-based ASR (which uses four times more training data). We follow the same methodology to evaluate the public databases. We also train 9 models for each dataset, i.e., LDC-ATCC and UWB-ATCC. We list the WERs on the right plot of Figure 1 for both test sets. Here, we note similar behaviors, thus we reach similar conclusions. First, scaling-up the fine-tuning data from 5 to 50 minutes brought ∼45% relative WERR for both, LDC-ATCC and UWB-ATCC test sets (similar trend in private databases, NATS and ISAVIA). Not surprisingly, further gains in WERs are achieved if we increase the fine-tuning data up to 11 hrs. Previous research has not explored E2E modeling[18] in the area of ATC, thus, these WERs can be adopted as baselines.

Transferability between ATC corpora: we have stated before that E2E models fine-tuned on a specific ATC corpus might not transfer well to different ATC corpora.[19] To test this premise, we train models with different public databases and test them on 4 test sets. We fixed the model (*w2v2-L-60k*), training data size to 11 hrs, and same hyperparameters. From Table 3, we can conclude that UWB-ATCC corpus transfers better to different databases, for instance LDC-ATCC. In this case, if we fine-tune *w2v2-L-60k* with UWB-ATCC set and test it on LDC-ATCC the performance is 54% WER, whereas inversely the performance is 64%, i.e., ∼10% absolute WERR. Similarly, the model trained on UWB-ATCC fits better ATCO2 test by a large margin compared to LDC-ATCC, i.e., 10% absolute lower WER.

Gender bias on air traffic control communications: finally, it is becoming a trend to go beyond simply analyzing the performance of E2E models on the selected downstream task, e.g., ASR. For instance, examining the performance shift of a given ASR model when fine-tuned on audio from different genders [15, 16]. We analyze this bias on ATCOSIM dataset, which provide the gender labels for each utterance. The results are listed in Table 4. It is evident that the experiments with female voice performed systematically better in all training scenarios (1h to 3.5h fine-tuning set). We also wanted to rule out the possibility that the speech rate was the main cause of this behavior. In average, each female recording has a speech rate of 3.4 words per second (WPS) while male has an average of 2.9 WPS. In order to determine whether female recordings are of better quality than the male ones, or whether the E2E model have some bias acquired during the pre-training phase, we calculated the WERR when fine-tuning the model between 1h to 3.5h of audio. Following Table 4 we can see that in the female experiments the reduction on WERs is higher than on the male side by around 8% absolute when scaling from 1h to 3.5h.

Table 4. Word error rates (WER) in percentages (%) on ATCOSIM dataset with different fine-tuning set sizes. WERs are reported on 0.7 hrs of speech (only from the same gender) sampled from the original test set, i.e., we fine-tune and evaluate within the same gender. All models are trained with *w2v2-L-60k* and decoding is done without LM. †list the word error rate reduction (WERR) achieved by scaling from 1h to 3.5h of speech during the fine-tuning stage.

	Dataset size				**WERR†**
Gender	1h	2h	3h	3.5h	(1h →3.5h)
Male	36.70	31.42	29.20	**28.72**	21.74%
Female	17.62	13.91	13.46	**12.37**	29.79 %

We believe that these E2E models (e.g., Wav2Vec 2.0) might carry little but noticeable gender bias. This could be one reason why the experiments with female recordings performed better. For instance, previous work have concluded that gender unbalance might affect E2E models during the pre-training phase. However, this bias can be mitigated by adding a small amount of data from the opposite gender [15]. In conclusion, it is still prudent to perform more thorough experiments before reaching hard judgments in this regard, or at least, in ATC communications.

5. CONCLUSION

This paper evaluated the robustness of pre-trained Wav2Vec 2.0 and XLS-R models on downstream ASR for ATC on different corpora. Our experiments show large recognition improvements of Wav2Vec 2.0 and XLS-R compared to *hybrid-based* ASR baselines. Quantitatively, between 20% and 40% relative WERR was obtained on ISAVIA and NATS test sets, but also from challenging multi-accent databases i.e., ATCO2-Test and LiveATC-Test. Furthermore, we demonstrated that pre-trained Wav2Vec 2.0 models allow rapid fine-tuning with small quantities of adaptation data. For instance, ∼5 min of speech allows fine-tuning a model that yields WERs of 40% and 43.9% for ISVAIA and NATS, respectively. Moreover, we showed that at least 4 hrs of in-domain data already provide acceptable WERs of ∼10% for ISAVIA and NATS recordings and by using two times more data (i.e., 8 hrs) performance surpasses hybrid-based ASR baselines. We also performed the same analysis for public databases. Similar WERs (∼40%) are achieved with 50 min of data. Finally, the gender bias in one ATC corpus is also covered by training gender dependent ASR models. We found that large-scale speech models perform systematically better on female recordings and also more gains in WERs are achieved when scaling up the fine-tuning data, in comparison to male recordings. The strength of this research is that, this is the first research aiming at analyzing the performance of these large-scale speech models on ATC. In addition, this is the first work in the ATC area that is publicly releasing a GitHub repository to replicate experiments.

6. ACKNOWLEDGMENTS

The work was supported by SESAR Joint Undertaking under Grant Agreement No. 884287 - HAAWAII (**H**ighly **A**utomated **A**ir traffic controller **W**orkstations with **A**rtificial **I**ntelligence **I**ntegration). The work was also partially supported by CleanSky Joint Undertaking under Grant Agreement No. 864702—ATCO2 (Automatic

[18] Training scripts to replicate the right plot of Figure 1 are public in our GitHub repository.

[19] This assumption also applies to hybrid-based ASR models.

collection and processing of voice data from air-traffic communications).

7. REFERENCES

[1] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Interspeech*, pp. 3465–3469, 2019.

[2] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[3] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[5] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[6] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv preprint arXiv:2202.03555*, 2022.

[7] Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli, "Unsupervised speech recognition," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[8] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[9] Zi-Qiang Zhang, Yan Song, Ming-Hui Wu, Xin Fang, and Li-Rong Dai, "Xlst: Cross-lingual self-training to learn multilingual representation for low resource speech recognition," *arXiv preprint arXiv:2103.08207*, 2021.

[10] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al., "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[11] International Civil Aviation Organization, "Icao phraseology reference guide," 2020.

[12] Oliver Ohneiser, Saeed Sarfjoo, Hartmut Helmke, Shruthi Shetty, Petr Motlicek, Matthias Kleinert, Heiko Ehr, and Šarūnas Murauskas, "Robust command recognition for lithuanian air traffic control tower utterances," in *Interspeech*, 2021.

[13] Juan Zuluaga-Gomez, Iuliia Nigmatulina, Amrutha Prasad, Petr Motlicek, Karel Veselỳ, Martin Kocour, and Igor Szöke, "Contextual Semi-Supervised Learning: An Approach to Leverage Air-Surveillance and Untranscribed ATC Data in ASR Systems," in *Interspeech*, 2021, pp. 3296–3300.

[14] Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, et al., "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[15] Yen Meng, Yi-Hui Chou, Andy T Liu, and Hung-yi Lee, "Don't speak too fast: The impact of data bias on self-supervised speech models," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3258–3262.

[16] Marcely Zanon Boito, Laurent Besacier, Natalia Tomashenko, and Yannick Estève, "A Study of Gender Impact in Self-supervised Models for Speech-to-Text Systems," in *Proc. Interspeech 2022*, 2022, pp. 1278–1282.

[17] Morgane Riviere, Jade Copet, and Gabriel Synnaeve, "ASR4REAL: An extended benchmark for speech models," *arXiv preprint arXiv:2110.08583*, 2021.

[18] Hartmut Helmke, Oliver Ohneiser, Thorsten Mühlhausen, and Matthias Wies, "Reducing controller workload with automatic speech recognition," in *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*. IEEE, 2016, pp. 1–10.

[19] Hartmut Helmke, Oliver Ohneiser, Jörg Buxbaum, and Chr Kern, "Increasing atm efficiency with assistant based speech recognition," in *Proc. of the 13th USA/Europe Air Traffic Management Research and Development Seminar, Seattle, USA*, 2017.

[20] Martin Kocour, Karel Veselỳ, Igor Szöke, Santosh Kesiraju, Juan Zuluaga-Gomez, Alexander Blatt, Amrutha Prasad, Iuliia Nigmatulina, Petr Motlíček, Dietrich Klakow, et al., "Automatic processing pipeline for collecting and annotating air-traffic voice communication data," *Engineering Proceedings*, vol. 13, no. 1, pp. 8, 2021.

[21] Juan Zuluaga-Gomez, Petr Motlicek, Qingran Zhan, Karel Veselỳ, and Rudolf Braun, "Automatic Speech Recognition Benchmark for Air-Traffic Communications," in *Interspeech*, 2020, pp. 2297–2301.

[22] Juan Zuluaga-Gomez, Karel Veselỳ, Alexander Blatt, Petr Motlicek, Dietrich Klakow, Allan Tart, Igor Szöke, Amrutha Prasad, Saeed Sarfjoo, Pavel Kolčárek, et al., "Automatic call sign detection: Matching air surveillance data with air traffic spoken communications," in *Multidisciplinary Digital Publishing Institute Proceedings*, 2020, vol. 59, p. 14.

[23] Amrutha Prasad, Juan Zuluaga-Gomez, Petr Motlicek, Oliver Ohneiser, Hartmut Helmke, Saeed Sarfjoo, and Iuliia Nigmatulina, "Grammar Based Identification Of Speaker Role For Improving ATCO And Pilot ASR," *arXiv preprint arXiv:2108.12175*, 2021.

[24] Juan Zuluaga-Gomez, Seyyed Saeed Sarfjoo, Amrutha Prasad, Iuliia Nigmatulina, Petr Motlicek, Oliver Ohneiser, and Hartmut Helmke, "BERTraffic: A robust BERT-based approach for speaker change detection and role identification of air-traffic communications," *IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar*, 2023.

[25] David Imseng, Blaise Potard, Petr Motlicek, Alexandre Nanchen, and Hervé Bourlard, "Exploiting un-transcribed foreign data for speech recognition in well-resourced languages," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2322–2326.

[26] Ajay Srinivasamurthy, Petr Motlicek, Ivan Himawan, Gyorgy Szaszak, Youssef Oualil, and Hartmut Helmke, "Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," in *Proc. of the 18th Annual Conference of the International Speech Communication Association*, 2017.

[27] Matthias Kleinert, Hartmut Helmke, Gerald Siol, Heiko Ehr, Aneta Cerna, Christian Kern, Dietrich Klakow, Petr Motlicek, Youssef Oualil, Mittul Singh, et al., “Semi-supervised adaptation of assistant based speech recognition models for different approach areas,” in *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*. IEEE, 2018, pp. 1–10.

[28] Yi Lin, “Spoken instruction understanding in air traffic control: Challenge, technique, and application,” *Aerospace*, vol. 8, no. 3, pp. 65, 2021.

[29] Petr Motlicek, Fabio Valente, and Philip N Garner, “English spoken term detection in multilingual recordings,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[30] Martin Kocour, Karel Veselỳ, Alexander Blatt, Juan Zuluaga Gomez, Igor Szöke, Jan Cernocky, Dietrich Klakow, and Petr Motlicek, “Boosting of Contextual Information in ASR for Air-Traffic Call-Sign Recognition,” in *Interspeech*, 2021, pp. 3301–3305.

[31] Iuliia Nigmatulina, Rudolf Braun, Juan Zuluaga-Gomez, and Petr Motlicek, “Improving callsign recognition with air-surveillance data in air-traffic communication,” *arXiv preprint arXiv:2108.12156*, 2021.

[32] Iuliia Nigmatulina, Juan Zuluaga-Gomez, Amrutha Prasad, Seyyed Saeed Sarfjoo, and Petr Motlicek, “A two-step approach to leverage contextual data: speech recognition in air-traffic communications,” in *ICASSP*, 2022.

[33] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, “Purely sequence-trained neural networks for ASR based on lattice-free MMI.,” in *Interspeech*, 2016, pp. 2751–2755.

[34] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[35] Apoorv Vyas, Srikanth Madikeri, and Herve Bourlard, “Lattice-free MMI adaptation of self-supervised pretrained acoustic models,” in *ICASSP*. IEEE, 2021, pp. 6219–6223.

[36] Alexei Baevski and Abdelrahman Mohamed, “Effectiveness of self-supervised pre-training for ASR,” in *ICASSP*. IEEE, 2020, pp. 7694–7698.

[37] Qiu-Shi Zhu, Jie Zhang, Zi-Qiang Zhang, Ming-Hui Wu, Xin Fang, and Li-Rong Dai, “A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3174–3178.

[38] Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, and Aaron van den Oord, “Learning robust and multilingual speech representations,” *arXiv preprint arXiv:2001.11128*, 2020.

[39] Apoorv Vyas, Srikanth Madikeri, and Hervé Bourlard, “Comparing CTC and LFMMI for Out-of-Domain Adaptation of wav2vec 2.0 Acoustic Model,” in *Proc. Interspeech 2021*, 2021, pp. 2861–2865.

[40] John Godfrey, “The Air Traffic Control Corpus (ATC0) - LDC94S14A,” 1994.

[41] Luboš Šmídl, Jan Švec, Daniel Tihelka, Jindřich Matoušek, Jan Romportl, and Pavel Ircing, “Air traffic control communication (ATCC) speech corpora and their use for ASR and TTS development,” *Language Resources and Evaluation*, vol. 53, no. 3, pp. 449–464, 2019.

[42] Konrad Hofbauer, Stefan Petrik, and Horst Hering, “The atcosim corpus of non-prompted clean air traffic control speech.,” in *LREC*, 2008.

[43] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., “The kaldi speech recognition toolkit,” in *IEEE workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number CONF.

[44] Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al., “Libri-light: A benchmark for ASR with limited or no supervision,” in *ICASSP*. IEEE, 2020, pp. 7669–7673.

[45] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, and Morgan Funtowicz et al, “Transformers: State-of-the-art natural language processing,” in *EMNLP (Demos)*, 2020, pp. 38–45.

[46] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al., “Datasets: A community library for natural language processing,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2021, pp. 175–184.

[47] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[48] Dan Hendrycks and Kevin Gimpel, “Gaussian error linear units (GELUs),” *arXiv preprint arXiv:1606.08415*, 2016.

[49] Ilya Loshchilov and Frank Hutter, “Decoupled Weight Decay Regularization,” in *International Conference on Learning Representations*, 2019.

[50] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.

[51] Niko Moritz, Takaaki Hori, and Jonathan Le Roux, “Streaming end-to-end speech recognition with joint CTC-attention based models,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 936–943.

[52] Arun Narayanan, Tara N Sainath, Ruoming Pang, Jiahui Yu, Chung-Cheng Chiu, Rohit Prabhavalkar, Ehsan Variani, and Trevor Strohman, “Cascaded encoders for unifying streaming and non-streaming ASR,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5629–5633.

[53] Liang Lu, Jinyu Li, and Yifan Gong, “Endpoint Detection for Streaming End-to-End Multi-Talker ASR,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7312–7316.