

How Does Pre-trained Wav2Vec2.0 Perform on Domain-Shifted ASR? An Extensive Benchmark on Air Traffic Control Communications

Juan Zuluaga-Gomez^{†,1,2}, Amrutha Prasad^{1,3}, Iuliia Nigmatulina¹, Saeed Sarfjoo¹, Petr Motlicek^{1,3},
Matthias Kleinert⁴, Hartmut Helmke⁴, Oliver Ohneiser⁴, Qingran Zhan⁵

¹ Idiap Research Institute, Martigny, Switzerland

² Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland

³ Brno University of Technology, Brno, Czech Republic

⁴ German Aerospace Center (DLR), Institute of Flight Guidance, Braunschweig, Germany

⁵ School of Information and Electronics, Beijing Institute of Technology, Beijing, China

juan-pablo.zuluaga@idiap.ch

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 substantially differs between the pre-training and downstream fine-tuning phases (i.e., domain shift). We target this scenario by analyzing the robustness of Wav2Vec2.0 and XLS-R models on downstream ASR for a completely unseen domain, i.e., air traffic control (ATC) communications. We benchmark the proposed models on four challenging ATC test sets (signal-to-noise ratio varies between 5 to 20 dB). Relative word error rate (WER) reduction between 20% to 40% are obtained in comparison to hybrid-based state-of-the-art ASR baselines by fine-tuning E2E acoustic models with a small fraction of labeled data. We also study the impact of fine-tuning data size on WERs, going from 5 minutes (few-shot) to 15 hours.

Index Terms: Automatic speech recognition, Wav2Vec2.0, air traffic control communications, self-supervised pre-training.

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]. In fact, [6] 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 [7]. Similarly, performance on cross-lingual speech recognition largely improved by SSL [8, 9]. It can be argued that SSL-based pre-training allows models to capture a good representation of acoustics that can be leveraged across different languages for ASR.

[†] corresponding author.

This work was supported by the SESAR Joint Undertaking under Grant Agreement No. 884287, under European Union’s Horizon 2020 Research and Innovation programme.

This work reviews the robustness of two well-known E2E acoustic models trained by SSL (i.e., Wav2Vec2 and XLS-R) on a very different domain: air traffic control (ATC) communications. ATC deals with guidance of aircraft in the air and on the ground via voice communications between air traffic controllers (ATCOs) and pilots. These are ruled by a well-defined grammar and vocabulary that must be followed to provide a safe and reliable flow of air traffic while keeping operation costs as low as possible. Despite the interest in ASR for ATC, there is not a fully functional ASR engine on the market due to: (i) lack of performance i.e., usually under 5% WER (enhancing the ATCOs productivity rather than delaying them in their tasks, see [10]), and (ii) lack of large-scale annotated speech data (less than 50 h of open-source speech data) and its high production cost makes it almost impractical [11].

1.1. Contribution and motivation

Only few past works intended to measure the effect of domain mismatch between pre-training and fine-tuning phases of E2E models [12]. However, we can still categorize all databases as either read, spontaneous or conversational speech. 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 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., Wav2Vec2) learn a strong representation of speech. Fine-tuning on a downstream task (e.g., ASR) is computationally less expensive than training from scratch, and it requires less in-domain data to achieve comparable results to hybrid-based ASR baselines. We also address the hypothesis that multilingual E2E models such as XLS-R [9] perform better on ATC speech data that contains accented English (i.e., LiveATC-Test and ATCO2-Test sets) because of the general speech representation learned during SSL.

(ii) how much ATC labeled data (both, acoustic and textual) is needed in the fine-tuning phase to reach comparable performance to hybrid-based models? We perform a comparative study ranging from 5 minutes (few-shot learning) to ~15 h of labeled speech (i.e., from 100 to 15k utterances). In addition, we investigate the performance boost obtained by decoding with beam search using an in-domain language model (LM) instead of baseline greedy decoding.

(iii) even though Wav2Vec2 and XLS-R models are not streaming by design, can such E2E models be used in real-time applications e.g., ATC? Many real-life applications (e.g., ATC) demand streaming ASR engines. We evaluate the latency of the forward pass and decoding of both E2E models, i.e., Wav2Vec2 and XLS-R during inference.

2. Related Work

With the aim of increasing airspace safeness, reducing ATCOs workload and decreasing the environmental impact caused by ATC operations, the European Union (EU) funded different projects that intend to bring closer speech and text-based technologies to ATC. MALORCA project concluded that ATCOs workload can be reduced by integrating ASR, while increasing their efficiency [13]. Then, ATCO2¹ project developed a pipeline [14] to automatically collect and pre-process large quantities of ATC speech data covering downstream task such as, ASR [15], named-entity recognition [16] and, text-based diarization [17, 18]. Ongoing HAAWAI² project develops a reliable and adaptable solution to automatically transcribe voice utterances issued by both ATCOs and pilots. Still, all previous research only investigates standard supervised and semi-supervised [19] hybrid-based ASR systems.

ATC communications are rich in named entities, e.g., call-signs, values and units. The most important and critical is the call-sign, 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. That is why, it is important for ASR engines to provide considerable high performance in order to avoid possible error propagation, which can be misleading for the sub-systems at the next stages. We redirect the reader to a general overview on spoken instruction understanding for ATC by Lin [20] and latest work on hybrid-based ASR for ATC in [14, 16]. In [11] unlabeled ATC speech is employed in semi-supervised learning to decrease word error rates. Boosting of contextual knowledge during and after decoding has also been explored in [21, 22, 23].

Despite the recent success of mixing SSL acoustic pre-training on E2E architectures for ASR, there has not been yet a comparative study between standard 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 [24] which relies on alignments produced by a previously trained HMM-GMM model [24]. Second, E2E systems model AM and LM jointly, and they are mostly trained with connectionist temporal classification (CTC) loss [25] (enabling alignment-free training). [26] compares CTC and LF-MMI adaption of pre-trained models. Recently, attention-based (e.g., Transformers) have become *de facto* choice for AM [4, 27, 9]. However, only few studies focused on domain mismatch or domain shift during pre-training and fine-tuning. For instance, [12, 28, 29] perform experiments similar to ours, addressing the domain-shift scenario between pre-training and fine-tuning phases.

Table 1: Train and test sets characteristics. † baseline performance of our state-of-the-art hybrid-based ASR model for ATC.

Dataset	Characteristics		
	Train / Test	SNR [dB]	WER [%] [†]
NATS	18h / 0.9h	≥20	7.7
ISAVIA	14h / 1h	15-20	12.5
ATCO2-Test	- / 1h	10-15	24.7
LiveATC-Test	- / 1.8h	5-15	35.8

3. Datasets and Experimental Setup

We experiment on two train sets and four test sets in English with various accents (see Table 1). The collection of ATC data is challenging and costly due to noise conditions, data privacy, rate of speech and language accent.

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

ATCO2-Test: development and evaluation set available as open-source and presented at Interspeech 2021 [11, 21]. The data consists of ATC communications from different airports located in Australia, Czech Republic, Slovakia and, Switzerland (see ATCO2 website³). ATCO2-Test contains a mix of noisy and heavily English accented recordings. This is the first study that evaluates E2E ASR for ATCO2-Test i.e., the WERs listed here could be adopted as baselines for future research.

LiveATC-Test: the test set is gathered from LiveATC⁴ data recorded from publicly accessible VHF radio channels, as a part of ATCO2 project [11, 23], 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 [16].

3.1. Automatic speech recognition

Our experimental setup is based on two fine-tuning data sets. First, we use 32 h of annotated data from NATS and ISAVIA, listing its characteristics in Table 1. Second, we use 132 h of ATC speech data from different projects, and we redirect the reader to [11] for further details. For now on, we refer to these fine-tuning sets as *32 h* and *132 h* fine-tuning sets.

Hybrid-based ASR: all experiments are conducted with Kaldi toolkit [30]. The baseline models are composed of six convolution layers and 15 factorized time-delay neural network (around 31M trainable parameters). We follow the standard Kaldi’s chain LF-MMI training recipe [24]. The input features are high-resolution MFCCs with online Cepstral mean normalization (CMN). The features are extended with i-vectors. We use 3-gram ARPA LM during decoding. The model is trained for 5 epochs on 132 h of ATC speech (that includes NATS and ISAVIA). Further information and baseline performances can be found in our previous work [11, 15, 16]. SOTA WERs are listed in the last column of Table 1.

¹EU Horizon 2020 project: <https://www.atco2.org/>

²HAAWAI project: <https://www.hawaii.de>

³ATCO2-ASRdataset-v1_beta: <https://www.atco2.org/data>

⁴Streaming audio platform that gathers VHF ATC communications

End-to-end ASR: we report results on four configurations of Wav2Vec2/XLS-R models fetched from HuggingFace platform [31]. From now on, we tag these models as: i) *w2v2-B*: BASE model (95M parameters, pre-trained on train-set 960h LibriSpeech [7]); ii) *w2v2-L*: LARGE-960h model (317M parameters pre-trained and then fine-tuned with LibriSpeech 960h train-set); iii) *w2v2-L-60K*: LARGE-960h-LV60K model (same as *w2v2-L* but uses LibriSpeech + 60k h from LibriVox project i.e., Libri-Light [32] during the pre-training phase); iv) *w2v2-XLS-R*: XLS-R model (300M parameters pre-trained on 436k h of publicly available data in 128 languages [9]). All experiments use the same set of hyperparameters. The feature encoder is not updated (frozen) during the whole fine-tuning phase (common practice in low-resource scenarios). We fine-tune each model for 10k steps, with a 500-step warm-up phase ($\sim 5\%$ of total updates). Learning rate is increased linearly until $1e-4$ during warm-up, then it linearly decays. 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). We use a character-based vocabulary of a dimension of 32.

Data augmentation: we apply a data augmentation strategy similar to [33]. We mask the input sequence with a probability $p = 0.075$, and $M = 12$ consecutive frames. We also use an activation and attention dropout of 0.05. These hyperparameters follow the original Wav2Vec2 implementation [4].

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 decoder⁵. 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 score normalization weights. We set the beam size to 100.

3.2. Incremental training

With the recent success of E2E models pre-trained with SSL, it is of particular interest to quantify how much data a model actually needs to perform effectively on a downstream task. It is especially 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 due to a considerable AM domain-shift (accent, speaker rates), as well as a LM domain-shift (dominance of different aircraft and different commands depending on the airport). 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. In total, nine models are fine-tuned on either only NATS (red dashed line), or only ISAVIA data (blue straight line) as depicted in Figure 1 (x-axis refers to number of utterances used during fine-tuning in log scale).

3.3. Streaming evaluation

Wav2Vec2 and XLS-R models are not designed with ‘streaming’ capabilities, but it is possible to leverage GPU capabilities during inference to provide real-time decoding. To test this hypothesis, we perform the following procedure: we split recording n in chunks of incremental sizes of $\sim 300ms$ of speech. We then pass each incremental chunk to the model until consuming the whole recording. Finally, we measure the mean time needed by the network to decode all chunks of a given utterance n . We

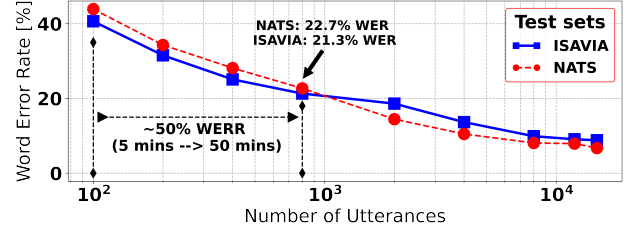


Figure 1: *Impact of fine-tuning data size on WERs. Each test set (i.e., NATS and ISAVIA) only uses its in-domain data during fine-tuning and evaluation. 100, 1k and 10k utterances are roughly 5 min (few-shot), 1 h and, 10 h, respectively.*

repeat this process on $n = 100$ random samples from the test sets, and we report the mean latency time. We do not consider the impact on WERs originated by the streaming setup.

4. Results and Discussion

In this paper, we hypothesize 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 [9]. We split our findings by answering the following questions:

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 h fine-tuning set on Kaldi (**Baseline**, first row Table 2). For E2E modeling we select: i) *w2v2-L-60k* for NATS and ISAVIA test sets, which was only fine-tuned on the 32 h set i.e., in-domain data; and ii) *w2v2-XLS-R+* for ATCO2-Test and LiveATC-Test test sets, which was trained on 132 h of ATC speech data [11, 15] i.e., more diverse data and same as the hybrid-based model. *w2v2-L-60k* yielded 30 and 41% relative word error rate reduction (WERR) on NATS and ISAVIA compared to hybrid-based 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 ATCO2-Test and LiveATC-Test, the two most challenging. In total, 19 and 30% relative WERR on ATCO2-Test and LiveATC-Test were obtained, respectively (hybrid-based \rightarrow *w2v2-XLS-R+*).

Does additional partly-in-domain data increases ASR performance? We answer this question by comparing models fine-tuned either on the 132 h or 32 h set. Note that NATS and ISAVIA are clean in-domain ATC speech corpora, i.e., considered as in-domain for the 32 h and partly-in-domain otherwise (132 h set). ATCO2-Test and LiveATC-Test can be considered as noisy and partly-in-domain sets (different airports, i.e., acoustic and LM mismatch). We focus on *w2v2-L-60k* and *w2v2-L-60k+* fine-tuned on the 32 h and 132 h sets, respectively. Note that there are comparable results between *w2v2-XLS-R* and *w2v2-XLS-R+*. We analyze WERs on greedy decoding to focus only on jointly AM+LM ASR. We noted a degradation on WERs for the in-domain test sets, NATS: 6.8% \rightarrow 9.3% WER and ISAVIA: 8.8% \rightarrow 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-in-domain sets, ATCO2-Test: 34.6% \rightarrow 23.3% WER and LiveATC-Test 39.8% \rightarrow 31.1% WER. To summarize, NATS

⁵PyCTCDecode: <https://github.com/kensho-technologies/pyctcdecode>

Table 2: WER on four proposed test sets. Each model is fine-tuned on NATS and ISAVIA data (~ 32 h). 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 h train-set [7], LV for LibriVox 60k h train-set [32] and ML for 436k h of multilingual speech data [9]. *reports the baseline WER of Wav2Vec2 (Table 1 from [4]) and XLS-R (Table 11 from [9]) models on LibriSpeech test set ‘other’ when only fine-tuned on 10 h of labeled data (comparable to our setup). \dagger best Kaldi hybrid-based model (see [11, 15]) trained with the same amount of data as $\dagger\dagger$. $\dagger\dagger$ model fine-tuned with 132 h of ATC speech data (instead of 32 h) and twice the number of steps, i.e., 20k. \P Latency covers the forward pass, decoding and detokenization (left: greedy decoding latency / right: beam search with LM latency). \S latency taken from [34]. WERs in **bold** refer to models fine-tuned to 32 h of data and underline to 132 h.

	Unlabeled	NATS		ISAVIA		ATCO2-Test		LiveATC-Test		LS*	Latency
Model (num. params.)	data	Greedy	+LM	Greedy	+LM	Greedy	+LM	Greedy	+LM	-	(ms) [¶]
Baseline (31M)											
Hybrid-based [†]	-	-	7.7	-	12.5	-	24.7	-	35.8	-	~400 [§]
BASE (95M)											
w2v2-B	LS	10.7	8.4	12.5	10.1	45.6	40.1	48.1	42.2	7.8	32/69
LARGE (371M)											
w2v2-L	LS	9.3	7.6	11.7	9.5	44.9	40.0	47.5	41.4	6.1	33/73
w2v2-L-60k	LS+LV	6.8	5.4	8.8	7.3	34.6	31.2	39.8	34.5	4.9	33/76
w2v2-L-60k+ ^{††}	LS+LV	9.3	<u>7.4</u>	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	39/76
w2v2-XLS-R+ ^{††}	ML	<u>9.0</u>	<u>7.4</u>	<u>10.4</u>	<u>8.3</u>	22.8	19.8	29.7	24.9	-	-/-

test set (ISAVIA: 1% relative WERR) was impacted by the addition of partly-in-domain data, i.e., $\sim 7\%$ relative worse WERs. Nevertheless, challenging test sets improved dramatically, i.e., ATCO2-Test and LiveATC-Test 43% and 33% relative WERR.

Does multilingual pre-trained models help? If we compare w2v2-L-60k+ and w2v2-XLS-R+ that use the same fine-tuning setup and beam search decoding with LM, a relative WERR of 8.8%, 6.6% and 8.5% is seen 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. Therefore, multilingual pre-trained models bring a small boost on performance compared to single-language pre-trained models. We can also infer that w2v2-XLS-R have seen considerably more multilingual and accented data during the pre-training phase [9] in comparison to w2v2-L-60k [4].

How much data do you need to fine-tune Wav2Vec2 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 list the WERs on Figure 1. All the experiments are based on the most robust E2E model from Table 2 i.e., w2v2-L-60K. The WERs are obtained by greedy decoding, i.e., no LM or explicit textual information is 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, $\sim 50\%$ relative WERR is obtained by scaling up the fine-tuning data to 50 minutes (800 utterances). Specifically, NATS 43.9% \rightarrow 22.7% WER and ISAVIA 40.6% \rightarrow 21.3% WER. Lastly, if all available data (~ 14 hrs) is used, we reach a 8.8% and 6.8% WER for ISAVIA and NATS, respectively. This represents a $\sim 80\%$ relative WERR compared to the low-resource setup (100 utterances). With around 8 h (~ 8000 utterances)

w2v2-L-60K beats the performance of our SOTA hybrid-based ASR (which uses four times more training data).

Is real-time ASR possible on E2E architectures, e.g., Wav2Vec2? We benchmark all six models from Table 2 on streaming mode in one mid-end NVIDIA GeForce GTX 1080 Ti GPU. The latency includes the model’s forward pass, beam search decoding (if +LM) and detokenization. Main results are reported in Table 2 (last column). Latency of the forward pass of Wav2Vec2 and XLS-R models are in overall below ~ 100 ms. For instance, w2v2-B/L/L-60k and w2v2-XLS-R models have a latency below 40 ms when performing greedy decoding. It takes roughly double if beam search decoding with a 4-gram LM is used. This research does not cover the degradation on WERs caused by using E2E models in streaming mode.

5. Conclusion

This paper evaluated the robustness of pre-trained Wav2Vec2 models on downstream ASR for ATC. Our experiments show large recognition improvements of Wav2Vec2 and XLS-R compared to *hybrid-based* ASR baselines. Quantitatively, between 20% and 40% relative WERR was obtained on test sets from Iceland Oceanic airspace (ISAVIA), London approach traffic (NATS) and from challenging multi-accent sets i.e., ATCO2-Test and LiveATC-Test. Furthermore, we demonstrated that the pre-trained Wav2Vec2 allows a rapid fine-tuning phase with small quantities of adaptation data e.g., ~ 5 min of speech allows fine-tuning a model that yields WERs of 40% and 43.9% for ISAVIA and NATS, respectively. Moreover, we showed that at least 4 h of in-domain data already provide acceptable WERs of $\sim 10\%$ for ISAVIA and NATS recordings and by using two times more data (i.e., 8 h) performance surpasses hybrid-based ASR baselines. Finally, we obtained competitive numbers on latency for Wav2Vec2 and XLS-R models on a mid-end GPU, i.e., $\sim 40/80$ ms on greedy and beam search decoding with LM.

6. References

- [1] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” *Interspeech*, pp. 3465–3469, 2019.
- [2] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [3] A. Baevski, S. Schneider, and M. Auli, “vq-wav2vec: Self-supervised learning of discrete speech representations,” *arXiv preprint arXiv:2110.05453*, 2021.
- [4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.
- [5] S. Chen, C. Wang, Z. Chen *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *arXiv preprint arXiv:2110.13900*, 2021.
- [6] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, “Unsupervised speech recognition,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *ICASSP*. IEEE, 2015, pp. 5206–5210.
- [8] Z.-Q. Zhang, Y. Song, M.-H. Wu, X. Fang, and L.-R. Dai, “Xlst: Cross-lingual self-training to learn multilingual representation for low resource speech recognition,” *arXiv preprint arXiv:2103.08207*, 2021.
- [9] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, “Xls-r: Self-supervised cross-lingual speech representation learning at scale,” *arXiv preprint arXiv:2111.09296*, 2021.
- [10] O. Ohneiser, S. Sarfjoo, H. Helmke, S. Shetty, P. Motlicek, M. Kleinert, H. Ehr, and Š. Murauskas, “Robust command recognition for lithuanian air traffic control tower utterances,” in *Interspeech*, 2021.
- [11] J. Zuluaga-Gomez, I. Nigmatulina, A. Prasad, P. Motlicek, K. Vesely, M. Kocour, and I. Szöke, “Contextual Semi-Supervised Learning: An Approach to Leverage Air-Surveillance and Untranscribed ATC Data in ASR Systems,” in *Interspeech*, 2021, pp. 3296–3300.
- [12] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve *et al.*, “Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training,” *arXiv preprint arXiv:2104.01027*, 2021.
- [13] H. Helmke, O. Ohneiser, J. Buxbaum, and C. 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.
- [14] M. Kocour, K. Vesely, I. Szöke, S. Kesiraju, J. Zuluaga-Gomez, A. Blatt, A. Prasad, I. Nigmatulina, P. Motlicek, D. Klakow *et al.*, “Automatic processing pipeline for collecting and annotating air-traffic voice communication data,” *Engineering Proceedings*, vol. 13, no. 1, p. 8, 2021.
- [15] J. Zuluaga-Gomez, P. Motlicek, Q. Zhan, K. Vesely, and R. Braun, “Automatic Speech Recognition Benchmark for Air-Traffic Communications,” in *Interspeech*, 2020, pp. 2297–2301.
- [16] J. Zuluaga-Gomez, K. Vesely, A. Blatt, P. Motlicek, D. Klakow, A. Tart, I. Szöke, A. Prasad, S. Sarfjoo, P. Kolčárek *et al.*, “Automatic call sign detection: Matching air surveillance data with air traffic spoken communications,” in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 59, no. 1, 2020, p. 14.
- [17] A. Prasad, J. Zuluaga-Gomez, P. Motlicek, O. Ohneiser, H. Helmke, S. Sarfjoo, and I. Nigmatulina, “Grammar based identification of speaker role for improving atco and pilot asr,” *arXiv preprint arXiv:2108.12175*, 2021.
- [18] J. Zuluaga-Gomez, S. S. Sarfjoo, A. Prasad, I. Nigmatulina, P. Motlicek, O. Ohneiser, and H. Helmke, “Bertraffic: A robust bert-based approach for speaker change detection and role identification of air-traffic communications,” *arXiv preprint arXiv:2110.05781*, 2021.
- [19] A. Srinivasamurthy, P. Motlicek, I. Himawan, G. Szaszak, Y. Oualil, and H. 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.
- [20] Y. Lin, “Spoken instruction understanding in air traffic control: Challenge, technique, and application,” *Aerospace*, vol. 8, no. 3, p. 65, 2021.
- [21] M. Kocour, K. Vesely, A. Blatt, J. Z. Gomez, I. Szöke, J. Cernocky, D. Klakow, and P. Motlicek, “Boosting of Contextual Information in ASR for Air-Traffic Call-Sign Recognition,” in *Interspeech*, 2021, pp. 3301–3305.
- [22] I. Nigmatulina, R. Braun, J. Zuluaga-Gomez, and P. Motlicek, “Improving callsign recognition with air-surveillance data in air-traffic communication,” *arXiv preprint arXiv:2108.12156*, 2021.
- [23] I. Nigmatulina, J. Zuluaga-Gomez, A. Prasad, S. S. Sarfjoo, and P. Motlicek, “A two-step approach to leverage contextual data: speech recognition in air-traffic communications,” in *ICASSP*, 2022.
- [24] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, “Purely sequence-trained neural networks for asr based on lattice-free mmi,” in *Interspeech*, 2016, pp. 2751–2755.
- [25] A. Graves, S. Fernández, F. Gomez, and J. 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.
- [26] A. Vyas, S. Madikeri, and H. Bourlard, “Lattice-free mmi adaptation of self-supervised pretrained acoustic models,” in *ICASSP*. IEEE, 2021, pp. 6219–6223.
- [27] A. Baevski and A. Mohamed, “Effectiveness of self-supervised pre-training for asr,” in *ICASSP*. IEEE, 2020, pp. 7694–7698.
- [28] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. v. d. Oord, “Learning robust and multilingual speech representations,” *arXiv preprint arXiv:2001.11128*, 2020.
- [29] A. Vyas *et al.*, “Comparing ctc and lfmmi for out-of-domain adaptation of wav2vec 2.0 acoustic model,” 2021.
- [30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit,” in *IEEE workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.
- [31] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, and M. F. *et al.*, “Transformers: State-of-the-art natural language processing,” in *EMNLP (Demos)*, 2020, pp. 38–45.
- [32] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, “Libri-light: A benchmark for asr with limited or no supervision,” in *ICASSP*. IEEE, 2020, pp. 7669–7673.
- [33] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.
- [34] C. S. Leow, T. Hayakawa, H. Nishizaki, and N. Kitaoka, “Development of a low-latency and real-time automatic speech recognition system,” in *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2020, pp. 925–928.