

BEST-RQ-BASED SELF-SUPERVISED LEARNING FOR WHISPER DOMAIN ADAPTATION

Raphaël Bagat, Irina Illina, Emmanuel Vincent

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

ABSTRACT

Automatic Speech Recognition (ASR) systems, despite large multilingual training, struggle in out-of-domain and low-resource scenarios where labeled data is scarce. We propose BEARD (BEST-RQ Encoder Adaptation with Re-training and Distillation), a novel framework designed to adapt Whisper’s encoder using unlabeled data. Unlike traditional self-supervised learning methods, BEARD uniquely combines a BEST-RQ objective with knowledge distillation from a frozen teacher encoder, ensuring the encoder’s complementarity with the pre-trained decoder. Our experiments focus on the ATCO2 corpus from the challenging Air Traffic Control (ATC) communications domain, characterized by non-native speech, noise, and specialized phraseology. Using about 5,000 hours of untranscribed speech for BEARD and 2 hours of transcribed speech for fine-tuning, the proposed approach significantly outperforms previous baseline and fine-tuned model, achieving a relative improvement of 12% compared to the fine-tuned model. To the best of our knowledge, this is the first work to use a self-supervised learning objective for domain adaptation of Whisper.

Index Terms— Automatic speech recognition, Whisper, self-supervised learning, domain adaptation

1. INTRODUCTION

Automatic speech recognition (ASR) has reached near-human accuracy in many domains [1]. The arrival of large-scale end-to-end models made these models easier to use out-of-the-box. However, despite being trained on massive multilingual datasets, these models still struggle with out-of-domain scenarios, like out-of-vocabulary words, spontaneous speech, noisy speech, etc. To adapt models to these new domains, supervised, self-supervised or self- trainings can be explored. Supervised training needs transcribed audio, which is costly and time-consuming to obtain.

A way to improve a speech recognition system to a new domain with little transcribed speech and a larger amount of untranscribed audio, is self-training [2]. A teacher model is first trained using the transcribed audio, and then used to create pseudo-labels for the untranscribed speech. A student model is then trained using the pseudo-labels in a supervised way. Though this method has shown to be successful [3], pseudo-labeling large amounts of data is computationally expensive and sensitive to the teacher’s model accuracy.

Self-supervised learning (SSL) methods were introduced to perform model training in the absence of annotations. SSL methods enable leveraging large amounts of unlabeled data to learn informative representations through representation learning. A decoder can then be added on top of the SSL encoder and trained in a supervised

This work was funded by the DeepMAUVES project supported by DGA of french MoD and CNRS, and granted access to the HPC resources of IDRIS under the allocation 2024-AD011015024 made by GENCI.

manner for ASR. wav2vec 2.0 [4] proposes using contrastive learning to predict representations of masked parts. This method, while proven effective across diverse domains [5, 6], is computationally expensive. Instead of using contrastive learning, HuBERT [7] employs k-means clustering to learn a quantizer that maps speech signals into discrete labels. It then performs BERT-style [8] prediction pre-training, by masking regions of the input. The prediction loss is applied over the masked regions. Both wav2vec 2.0 and HuBERT have been utilized to improve speech recognition for low-resource domains [9]. More recently, BERT-based Speech pre-Training with Random-projection Quantizer (BEST-RQ) [10] was introduced as a lighter SSL alternative for training speech encoders. The approach learns a model to predict the masked speech signals, in the form of discrete labels generated with a random-projection quantizer. The quantizer projects speech inputs with a randomly initialized matrix, and does a nearest-neighbor lookup in a randomly initialized codebook. As the quantizer is kept frozen and independent from the ASR model, this strategy remains adaptable and suitable for integration with any speech recognition framework. BEST-RQ has showed speech recognition improvements in different domains [11, 12].

In the context of a project funded by the French Directorate General of Armaments, we focus on the Air Traffic Control (ATC) communications low-resource domain. The ATC domain presents unique challenges for ASR: non-native speech, noisy conditions, and specialized phraseology. Transcribed data from this domain are scarce, whereas large amounts of untranscribed ATC speech are readily available [13]. In order to improve ATC speech recognition, different works have been proposed over the years, on simulated data [14], using a complex cascaded architecture [15], through a challenge [16], and using diverse datasets [17]. XLS-R and Whisper were studied on ATC data and showed promising results [18, 19]. All these previous works are based on the use of labeled ATC data only. Seeing how Whisper already provides a strong base for ATC ASR, we propose to leverage a large amount of unlabeled ATC speech to further improve Whisper’s speech recognition.

In this paper, we propose BEARD (BEST-RQ Encoder Adaptation with Re-training and Distillation), a framework that adapts Whisper’s encoder with unlabeled speech to improve ASR performance on new domains. Unlike traditional SSL methods, which are designed to pre-train speech encoders from scratch, BEARD makes use of Whisper’s already powerful encoder. By combining self-supervised learning and distillation, BEARD re-trains Whisper’s encoder on untranscribed speech while preserving its complementarity with the decoder. We show that BEARD gives a stronger base to fine-tune on low-resource labeled data from the same domain. Using BEARD decreases significantly the word error rate on the ATCO2 corpus compared to previous works [18, 19] and a model fine-tuning using only labeled data. Compared to previous studies, we apply an SSL objective to a middle-layer of the speech encoder and use distillation at the same time. To the best of our knowledge, this is the first work that leverages self-supervised learning to improve Whisper.

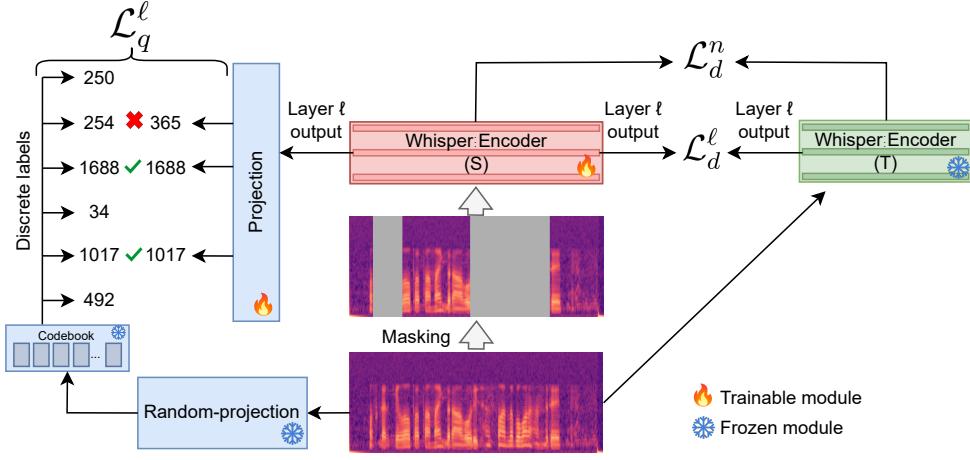


Fig. 1: Architecture of the proposed BEARD framework. On the left side, we use BEST-RQ’s objective (\mathcal{L}_q^ℓ). It is applied to the output of the ℓ -th Transformer layer. On the right, we use two distillation losses: \mathcal{L}_d^ℓ , \mathcal{L}_d^n . They are computed at two different layers, the ℓ -th layer and the output layer, respectively, by leveraging a frozen teacher encoder.

2. PROPOSED METHODOLOGY

2.1. Whisper

Whisper, end-to-end encoder-decoder Transformer, is a state-of-the-art model for automatic speech recognition [20]. It has been trained on 680,000 hours of transcribed multilingual data. Whisper’s encoder is mostly focused on acoustic features, while its decoder is mostly focused on linguistic features. Whisper differs from self-supervised learning models in that it is trained exclusively on labeled data, using a supervised objective.

2.2. BEST-RQ

BERT-based Speech pre-Training with Random-projection Quantizer (BEST-RQ) [10] is a self-supervised learning approach which converts input speech signal into discrete labels. This is done with a random-projection quantizer, which projects the input signal and finds the nearest vector in a codebook. The index of that vector becomes the discrete label. During pre-training, portions of the input log-mel spectrogram are masked and fed into a speech encoder. The speech encoder is trained to predict the discrete labels of the masked parts. The random-projection quantizer is kept frozen throughout the pre-training stage. After pre-training the speech encoder, a decoder can be added on top and trained for the ASR task. BEST-RQ had first been proposed as a pre-training approach for Conformer architectures [21]. It was demonstrated that BEST-RQ can also be applied to convolution-free Transformer architectures [22].

2.3. BEST-RQ Encoder Adaptation with Re-training and Distillation (BEARD)

SSL approaches, such as BEST-RQ, cannot be directly applied to models like Whisper. These methods are designed to train speech encoders from scratch, without a pre-existing decoder. In Whisper’s case, the encoder was trained along the decoder. Applying SSL to the encoder would leave the decoder unchanged, leading to a mismatch between them.

To address this issue, we propose BEARD, a framework that leverages BEST-RQ self-supervision to adapt Whisper’s encoder with untranscribed data, combined with knowledge distillation to maintain complementarity with its decoder. As shown in Fig. 1, we take two copies of Whisper’s pre-trained encoder, the decoder is not used. One of the encoders (referred to as S, for student) will be modified using unlabeled data, we call this entire stage *re-training*. The other encoder (referred to as T, for teacher) is frozen and will serve as the teacher for the distillation.

To re-train S with unlabeled data, we apply BEST-RQ’s quantization mechanism, whose loss is noted as \mathcal{L}_q^ℓ . Following Chiu et al. [10], a frozen random-projection quantizer is used, and by using codebooks, assigns discrete labels to the input. Instead of applying the prediction loss to the output of S, as originally proposed [10], we propose applying it to the output of its ℓ -th layer, out of n layers. This allows the upper layers to maintain representational alignment with the decoder, preserving the complementarity. We add a projection layer that takes the output of the ℓ -th layer as input, and is trained to predict the discrete labels of masked regions, as in BEST-RQ.

Using only \mathcal{L}_q^ℓ would lead to changes in the model up to the ℓ -th layer, leaving the following layers untouched. In order to adapt all layers, and, at the same time, preserve the complementarity with the decoder, we introduce two distillation losses using the frozen teacher encoder (T). \mathcal{L}_d^n is computed between the last layer outputs of S and T. By making sure S’s output space remains close to T’s, we assure the complementarity with the decoder is preserved. \mathcal{L}_d^ℓ is computed between the outputs of the ℓ -th layer of S and the ℓ -th layer of T, which helps the distillation process. Both \mathcal{L}_d^n and \mathcal{L}_d^ℓ use cosine similarity, which should be maximized. We use cosine similarity over L1 or mean squared error, because it is less constraining, allowing for the encoder to adapt to the new domain. Furthermore, as using SSL requires masking parts of S’s input, the distillation losses are computed exclusively on the unmasked regions.

The overall training objective \mathcal{L} is defined as

$$\mathcal{L} = \mathcal{L}_q^\ell + \lambda \mathcal{L}_d^\ell + \beta \lambda \mathcal{L}_d^n \quad (1)$$

where λ and β are weighting coefficients.

Table 1: WER (%) obtained with different fine-tuning methods on ATCO2. The Adaptation method column indicates the method used to adapt the ASR model. The ATCO2 for FT column indicates the quantity of ATCO2 audio used for fine-tuning. Bold numbers indicate the best result and those results which are statistically equivalent to it. * indicates results reported in the literature.

Adaptation method	# trained params.	ATCO2 for FT	Layer ℓ	Distillation weight λ	WER (%)
XLS-R (FT ATC) + LM [18]	300M	0 min			19.80*
Whisper-small, FT [19]	244M	52 min			22.79*
Whisper-small, No FT	0	0 min			63.32
Whisper-small, FT	244M	2 h 24 min			19.54
Whisper-small, BEARD + FT	244M	2 h 24 min	8	0.5 1.0	18.40 18.27
			7	0.5 1.0	17.87 19.68
Whisper-small, BEARD + FT	244M	2 h 24 min	6	0.5 1.0	17.17 18.05
			5	0.5 1.0	17.72 18.06
Whisper-small, BEARD + FT	244M	2 h 24 min	4	0.5 1.0	18.01 17.58

Following Chiu et al. [10], we apply normalization to prevent the random projection from collapsing to a limited subset of codes. The normalization is applied using LayerNorm at the input of both the random-projection quantizer and the projection layer. It normalizes the vectors to have 0 mean and standard deviation of 1.

After re-training S using BEARD, we propose adding back Whisper’s decoder on top of S. We then fine-tune them both, without BEARD, using labeled speech from the same domain as the untranscribed speech.

3. EXPERIMENTAL SETTINGS

3.1. Dataset

We conducted our experiments using the ATCO2 dataset¹ [13]. It contains air traffic control communications between pilots and air traffic controllers from various airports. The speech is non-native, with high speech rate and noisy, with signal-to-noise ratios (SNR) varying from -10dB to 40dB, estimated using WADA-SNR [23]. The corpus consists of 2 parts: 5,381 hours of audio without transcription, and 4 hours of audio with human transcriptions². We do not have any information about the number of speakers and their native language. For the self-supervised re-training stage, we employ the untranscribed part of the dataset. For the fine-tuning stage, we use the transcribed speech part. We conduct 4-fold cross-validation. For a given fold, 2 h 24 min (25,000 words) are used for the fine-tuning, 36 min (5,300 words) for validation, and 1 h (10,000 words) for testing. All the audio files of ATCO2 are sampled at 16kHz, which matches Whisper’s input requirements.

3.2. General parameters

Our experiments were conducted with the Whisper-small model [20] comprising 244M parameters. Its compact size makes it suitable for

¹ATCO2 project data, ELRA catalog (<http://catalog.elra.info>), ISLRN : 589-403-577-685-7, ELRA ID : ELRA-S0484

²Only one hour of transcribed audio is freely available.

deployment on a wide range of devices, including embedded systems, while remaining a high-performance ASR model. Its parameter scale (244M) is comparable to the XLS-R model [5] (300M) employed as the baseline for ATCO2 [18]. Moreover, both Whisper-small and XLS-R are end-to-end ASR systems trained on a wide variety of languages. For BEARD re-training stage, models are re-trained for a single epoch over all untranscribed data available, with a batch size of 32. We chose to run only one epoch since over 5,000 hours is a significant amount of data. We set the learning rate to 1e-5 for Whisper’s encoder and 5e-4 for the projection layer. We use a masking span of 4 frames [24]. We reduce the masking probability from 0.15 to 0.10 to ensure that a sufficient number of unmasked frames remain for the distillation. For the random-projection quantizer, we use a codebook size of 2048, which obtained the best results in our preliminary experiments. We set β to 0.1 to down-weight \mathcal{L}_d^n . Higher values led to worse results. On 8 NVIDIA V100 GPUs, one epoch of BEARD re-training takes around 7 hours on the 5,381 hours of untranscribed speech. This is much shorter than creating pseudo-labels. For the fine-tuning stage, models are trained until convergence, with a batch size of 16. The learning rate is set to 1e-5. For decoding, we use greedy search for computational reasons. Our code is publicly available³.

The results are reported in terms of the word error rate (WER). Early stopping is made using the WER on the validation set. The statistical significance of the results has been validated using the matched pair sentence segment test with SCTK [25] ($p = 0.001$).

4. RESULTS AND DISCUSSIONS

4.1. Baselines

We consider four baselines. The first two are prior works on the ATCO2 dataset. An XLS-R model fine-tuned on 132 hours of ATC speech from diverse corpora (referred to as XLS-R FT) [18], ATCO2 was only used for testing. This is the first baseline that was presented

³<https://gitlab.inria.fr/rbagat/beard>

Table 2: WER (%) on ATCO2 when BEARD is applied without distillation losses, with \mathcal{L}_d^ℓ only, with \mathcal{L}_d^n only, or with both. In any case, the \mathcal{L}_q^ℓ loss is used. Bold numbers indicate the best result.

Layer ℓ / λ	Using \mathcal{L}_d^ℓ	Using \mathcal{L}_d^n	WER (%)
6 / 0.5	No	No	80.98
	Yes	No	37.28
	No	Yes	20.44
	Yes	Yes	17.17

on ATCO2, the authors employed an external language model (LM) during decoding. Van Doorn et al. used Whisper-small and fine-tuned it on 52 min of ATCO2 audio (Whisper-small FT) [19]. We evaluate two baselines on Whisper-small: the model without fine-tuning (No FT), and a fine-tuned version (FT) fine-tuned on 2 h 24 min of ATCO2 audio and evaluated with our cross-validation setup.

The reported four baselines are shown in the first four rows of Table 1. Among these, No FT performs worst, with a WER of 63.32%. XLS-R FT reaches 19.80%, outperforming Whisper-small FT (22.79%). FT (2 h 24 min), using the same base model as Whisper-small FT, but fine-tuned with more data, achieves a WER of 19.54%, which is comparable to XLS-R FT.

4.2. BEARD results

For every experiment, we take the pre-trained Whisper-small model and we re-train its 12-layer encoder using BEARD. Then, we fine-tune both encoder and decoder on transcribed speech in a cross-validation setup. We evaluated BEARD under multiple conditions with different parameter settings: encoder layer $\ell \in \{4, 5, 6, 7, 8\}$ and weighting factor $\lambda \in \{0.5, 1\}$. These values were chosen to evaluate the sensibility of the method to the choice of the layer ℓ , and to see whether the distillation should be more balanced or not.

Effect of ℓ — Results in Table 1 show that applying BEARD at layers $\ell \in \{4, 5, 6, 8\}$ and using $\lambda = 0.5$ consistently significantly outperforms baseline results. Our best configuration, with $\ell = 6$ and $\lambda = 0.5$, achieves a WER of 17.17%, which represents a *significant improvement* of 12% (relative) compared to FT. Applying BEARD at layers 4, 5, and 6 yields slightly better results than at layer 7 and 8, suggesting that middle-level encoder layers are more effective for this type of adaptation. In additional experiments not reported in the table, applying BEARD at layer $\ell = 9$ with $\lambda = 0.5$, led to a WER of 18.64%. This further confirms that performance tends to degrade as BEARD is applied closer to the output of the encoder.

Effect of λ — To assess the influence of the distillation weight λ , we evaluated two values, $\lambda \in \{0.5, 1.0\}$, for each of the encoder layer ℓ to which BEARD was applied. As shown in Table 1, lower value of λ leads to better performance across most layers. For layers $\ell = 5, 6$, and 7 , using $\lambda = 0.5$ resulted in significantly lower WERs compared to $\lambda = 1.0$. In all cases, except when BEARD is applied at layer 7 with $\lambda = 1.0$, the results significantly outperform FT’s results.

SNR analysis — As previously mentioned, one of the main challenges of the ATC domain is the noise. Our ATCO2 unlabeled data contains different SNR level indications. We expect the proposed methodology will perform better across different SNRs. Fig. 2 compares the WER obtained by our best BEARD configuration ($\ell = 6, \lambda = 0.5$) and the FT model at different SNR levels. The SNRs were computed using WADA-SNR [23]. As it can be seen, the model re-trained with BEARD outperforms FT across all SNR

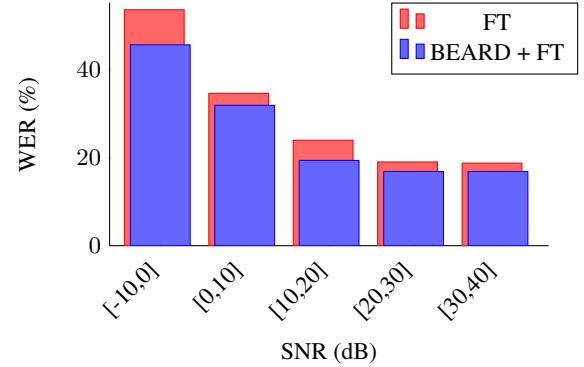


Fig. 2: Comparison of WER across SNR bins for our best BEARD configuration ($\ell = 6, \lambda = 0.5$) and FT. SNR was estimated using WADA-SNR [23].

levels. The largest improvements are within the [10,20]dB SNR range, with a relative improvement of 19%. Notably, in the negative SNRs, BEARD reaches a relative improvement of 15% over FT.

Ablation study — We conducted an ablation study to assess the contribution of the distillation losses. Four variants were compared: a version of BEARD without any distillation, one using only \mathcal{L}_d^ℓ , one using only \mathcal{L}_d^n , and the proposed BEARD combining both losses. For all variants, the \mathcal{L}_q^ℓ loss was used during re-training on unlabeled data, and the models were fine-tuned on labeled data after. All experiments were performed by applying BEARD at layer $\ell = 6$ with $\lambda = 0.5$, which is our best configuration. Results can be found in Table 2. It can be seen that removing distillation entirely leads to heavy degradation, with a WER of 80.98%. It confirms that BEST-RQ’s loss should not be applied to Whisper’s encoder without assuring complementarity with the decoder. Using distillation with only \mathcal{L}_d^ℓ , still results in degraded performance, with a WER of 37.28%. By using only \mathcal{L}_d^n , the results improve over using only \mathcal{L}_d^ℓ and lead to a WER of 20.44%, which is comparable to the FT model. This exhibits that when the complementarity with the decoder is not sufficiently assured, applying BEARD has less effect on the ASR performance. The best result is obtained when using both distillation losses, reaching 17.17% WER. This shows the importance of having distillation within the encoder, along the self-supervised objective.

5. CONCLUSION

In this work, we investigated whether self-supervised learning can help Whisper adapt to a new domain. We introduced BEARD, a framework that combines self-supervised learning and distillation to adapt Whisper’s encoder using unlabeled speech. The modified encoder is then fine-tuned with the decoder using a limited amount of labeled data. On the ATCO2 corpus, the best BEARD configuration significantly reduced the word error rate, with a relative improvement of 12% over fine-tuning with transcribed speech only. BEARD showed WER improvements across all SNR ranges. The ablation study demonstrated that distillation is important to keep the encoder-decoder complementarity. These results confirm that large-scale unlabeled data can be effectively exploited through self-supervision to adapt Whisper ASR model. The proposed approach can be used to adapt any encoder-decoder models to new domains. Future work will study alternative SSL objectives, use different quantity of unlabeled data (not currently tested due to computational limitations), filter the unlabeled data, and adapt larger Whisper variants.

6. REFERENCES

- [1] W. Xiong, J. Droppo, X. Huang, F. Seide, M.L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, “Toward human parity in conversational speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, 2017.
- [2] G. Zavaliagkos and T. Colthurst, “Utilizing untranscribed training data to improve performance,” in *DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne*, 1998.
- [3] J. Kahn, A. Lee, and A. Hannun, “Self-training for end-to-end speech recognition,” in *Proc. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7084–7088.
- [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. 12449–12460, 2020.
- [5] 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,” in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.
- [6] R. Huang and B. Mak, “wav2vec 2.0 ASR for cantonese-speaking older adults in a clinical setting,” in *Proc. Interspeech 2023*, 2023, pp. 4958–4962.
- [7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [8] J. Devlin, M-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, June 2019, pp. 4171–4186.
- [9] J. Zhao and W-Q. Zhang, “Improving automatic speech recognition performance for low-resource languages with self-supervised models,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1227–1241, 2022.
- [10] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, “Self-supervised learning with random-projection quantizer for speech recognition,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 3915–3924.
- [11] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang, et al., “Google USM: Scaling automatic speech recognition beyond 100 languages,” *arXiv preprint arXiv:2303.01037*, 2023.
- [12] H. Huang, T. Park, K. Dhawan, I. Medennikov, K.C. Puvvada, N.R. Koluguri, W. Wang, J. Balam, and B. Ginsburg, “NEST: Self-supervised fast conformer as all-purpose seasoning to speech processing tasks,” in *Proc. 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [13] J. Zuluaga-Gomez, K. Veselý, I. Szöke, P. Motlicek, et al., “ATCO2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications,” *arXiv preprint arXiv:2211.04054*, 2022.
- [14] J. Ferreiros, JM. Pardo, R. De Córdoba, J. Macias-Guarasa, JM. Montero, F. Fernández, V. Sama, G. González, et al., “A speech interface for air traffic control terminals,” *Aerospace Science and Technology*, vol. 21, no. 1, pp. 7–15, 2012.
- [15] Y. Lin, D. Guo, J. Zhang, Z. Chen, and B. Yang, “A unified framework for multilingual speech recognition in air traffic control systems,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3608–3620, 2021.
- [16] T. Pellegrini, J. Farinas, E. Delpech, and F. Lancelot, “The airbus air traffic control speech recognition 2018 challenge: Towards ATC automatic transcription and call sign detection,” in *Proc. Interspeech 2019*, 2019, pp. 2993–2997.
- [17] S. Badrinath and H. Balakrishnan, “Automatic speech recognition for air traffic control communications,” *Transportation research record*, vol. 2676, no. 1, pp. 798–810, 2022.
- [18] J. Zuluaga-Gomez, A. Prasad, I. Nigmatulina, S.S. Sarfjoo, P. Motlicek, M. Kleinert, H. Helmke, O. Ohneiser, and Q. Zhan, “How does pre-trained wav2vec 2.0 perform on domain-shifted ASR? an extensive benchmark on air traffic control communications,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 205–212.
- [19] J. van Doorn, J. Sun, J. Hoekstra, P. Jonk, and V. de Vries, “Whisper-ATC open models for air traffic control automatic speech recognition with accuracy,” in *Proc. Int. Conf. Res. Air Transp. (ICRAT)*, 2024.
- [20] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.
- [21] A. Gulati, J. Qin, C-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al., “Conformer: Convolution-augmented transformer for speech recognition,” in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.
- [22] Z. Hou, G. Huybrechts, A. Bhatia, D. Garcia-Romero, K.J. Han, and K. Kirchhoff, “Revisiting convolution-free transformer for speech recognition,” in *Proc. Interspeech 2024*, 2024, pp. 4568–4572.
- [23] C. Kim and R.M. Stern, “Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis,” in *Proc. Interspeech 2008*, 2008, pp. 2598–2601.
- [24] R. Whetten, T. Parcollet, M. Dinarelli, and Y. Estève, “Open implementation and study of BEST-RQ for speech processing,” in *Proc. 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 460–464.
- [25] NIST, “SCTK,” <https://github.com/usnistgov/SCTK.git>, 2024.