Interspeech 2024
1-5 September 2024, Kos, Greece

Whisper-PMFA: Partial Multi-Scale Feature Aggregation for Speaker Verification using Whisper Models

Yiyang Zhao[1], *Shuai Wang*[2], *Guangzhi Sun*[3], *Zehua Chen*[1], *Chao Zhang*[1], *Mingxing Xu*[1], *Thomas Fang Zheng*[1*]

[1]Tsinghua University, China, [2]Shenzhen Research Institute of Big Data, China
[3]University of Cambridge, United Kingdom

zhaoyy22@mails.tsinghua.edu.cn, wangshuai@cuhk.edu.cn, gs534@cam.ac.uk,
{zhc23,cz277,xumx,fzheng}@tsinghua.edu.cn

Abstract

In this paper, Whisper, a large-scale pre-trained model for automatic speech recognition, is proposed to apply to speaker verification. A partial multi-scale feature aggregation (PMFA) approach is proposed based on a subset of Whisper encoder blocks to derive highly discriminative speaker embeddings. Experimental results demonstrate that using the middle to later blocks of the Whisper encoder keeps more speaker information. On the VoxCeleb1 and CN-Celeb1 datasets, our system achieves 1.42% and 8.23% equal error rates (EERs) respectively, receiving 0.58% and 1.81% absolute EER reductions over the ECAPA-TDNN baseline, and 0.46% and 0.97% over the ResNet34 baseline. Furthermore, our results indicate that using Whisper models trained on multilingual data can effectively enhance the model's robustness across languages. Finally, the low-rank adaptation approach is evaluated, which reduces the trainable model parameters by approximately 45 times while only slightly increasing EER by 0.2%.[1]

Index Terms: speaker verification, Whisper, LoRA, speaker recognition, multilingual

1. Introduction

Speaker verification (SV) is crucial for biometric authentication, aiming to confirm a person's identity based on their voice characteristics. In the past decade, the advent of deep learning has led to significant advancements in speaker verification technology [1–3]. Models such as the convolutional neural network-based Residual Network (ResNet) [4], the time-delay neural network-based ECAPA-TDNN [5], and their diverse variants [6–8], alongside multi-scale feature fusion models like MFA-Conformer [9], have significantly contributed to the development of the field. Meanwhile, novel training methods, including loss functions [10–12], strategic training approaches [13], and score normalization techniques [14], have also considerably enhanced the performance of speaker verification systems.

However, the increasing complexity of model architectures intensifies the demand for training data [15]. Acquiring this data, particularly labelled datasets, is far from trivial, this challenge is further exacerbated in the realm of low-resource languages. To address these challenges, researchers have proposed numerous solutions, a widely used strategy involves leveraging large pre-trained models trained on extensive corpora(e.g. Whisper [16], HuBERT [17], WavLM [18]). The integration of pre-trained models provides a robust foundation for feature extraction and representation learning, thereby alleviating some of the constraints imposed by data scarcity. Berns et al. implemented a speaker change detection task on Whisper and Wav2vec2 by innovatively adding speaker change labels to the training data [19]. Further, Cai et al. explored the feasibility of applying automatic speech recognition (ASR) models to speaker verification tasks by testing an ASR-pretrained Conformer model in speaker verification scenarios [20, 21].

As a large pre-trained ASR model trained on extensive corpora, Whisper has been extensively trained on multilingual and diverse situational audio datasets [16]. This extensive training grants it impressive performance and robust cross-linguistic features. Accordingly, we adapted it as a pre-trained model for speaker recognition tasks.

However, due to the fine-tuning paradigm's intensive memory and computational resource requirements, leveraging large-scale models trained on extensive datasets introduces a significant challenge. Addressing this challenge, Sang et al. introduce an adapter approach for adapting self-supervised speech models to speaker verification tasks [22]. Concurrently, Peng et al. explore parameter-efficient transfer learning methods for adapting pre-trained Transformer models to speaker verification tasks, aiming to reduce the computational burden [23].

In this paper, we propose an effective Whisper-based partial multi-scale feature aggregation model (Whisper-PMFA). Compared to the widely used MFA architecture [9], our approach benefits from partial layer selection, effectively reducing the performance degradation caused by the integration of excessive irrelevant information. This selective process also mitigates the computational and memory overhead caused by full concatenation. Additionally, to address the issue of computational overhead, we explored the use of the low-rank adaptation (LoRA) approach [24] as an alternative to fine-tuning. The primary contributions of this paper can be summarized as follows:

- We developed a partial multi-scale feature aggregation module, adapting the Whisper model for the speaker verification task through selective aggregation of Whisper layers.
- In our enhanced model, by incorporating LoRA [24] in place of comprehensive fine-tuning, we significantly reduced the model's trainable parameters by approximately 45 times while only incurring a marginal increase in EER of 0.2%.
- Our experiments confirmed Whisper's potential for cross-lingual speaker recognition applications.
- Our evaluations on the VoxCeleb1 and CN-Celeb1 datasets have conclusively demonstrated that the proposed model achieves significant improvements over the baseline models.

2. Whisper-PMFA

In this section, the Whisper-PMFA approach is introduced, aimed at using the rich and diverse speech knowledge obtained from a large amount of training data and embedded in a pre-

*Corresponding author

[1]Our source code will be released in Wespeaker.

10.21437/Interspeech.2024-1515

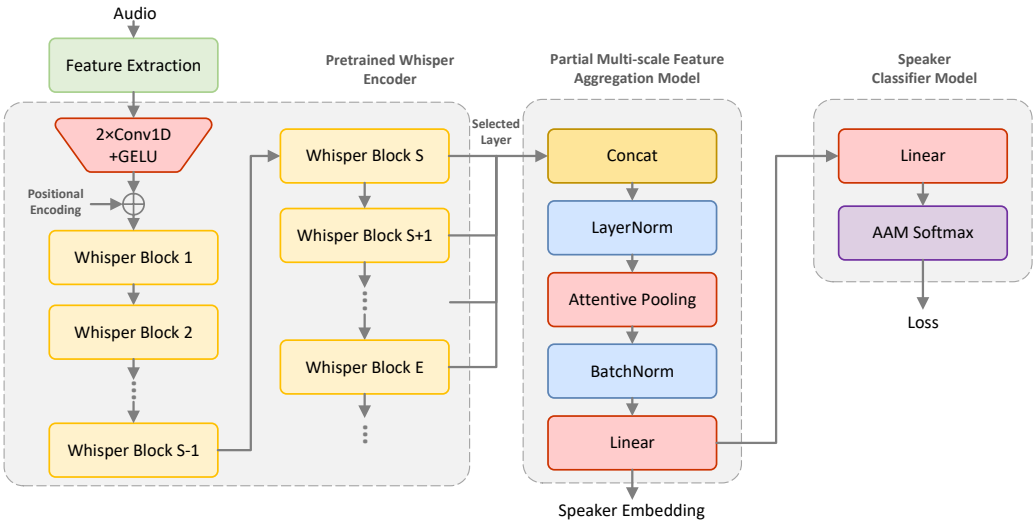

Figure 1: *The overall architecture of Whisper-PMFA, where S denotes the index of the initial Whisper block selected for feature aggregation, and E represents the index of the final Whisper block selected.*

trained ASR model to assist in the enhancement of speaker verification tasks. It combines the Whisper pre-trained model and aggregation model, with architectural details in Figure 1.

2.1. Preliminary of Whisper

Whisper [16] is a sophisticated, multilingual speech recognition model trained on a vast corpus of 680,000 hours of multilingual and diverse acoustic data. It combines a classical encoder-decoder Transformer [25] architecture, with its encoder comprising two convolutional layers, sinusoidal positional encoding, and a series of Transformer blocks, to effectively handle diverse linguistic and acoustic challenges. In this paper, we utilize the encoder from the Whisper Large-v2 as our pre-trained model, which comprises 32 Transformer blocks, each equipped with an attention mechanism that consists of 20 heads.

2.2. Partial Multi-scale Feature Aggregation Model

Multi-scale feature aggregation (MFA), involves the concatenation of output features from various frame-level modules within a speaker embedding architecture, before pooling at the utterance level. In previous research, MFA-Conformer has already demonstrated its effectiveness on Conformer models [9]. However, as the number of layers and the output size increase, concatenating the outputs of all blocks results in significant computational and memory overhead. Moreover, performing full concatenation may also introduce a substantial amount of non-speaker-related information, potentially leading to a degradation in model performance.

Therefore, unlike MFA, we replace the full concatenation operation with the partially selected layer concatenation:

$$\begin{aligned} \mathbf{H}' &= \text{Concat}(h_s, h_{s+1}, \cdots, h_e) \\ \mathbf{H} &= \text{LayerNorm}(\mathbf{H}') \end{aligned} \tag{1}$$

where s is the first Whisper block number to be selected, and e is the last Whisper block number to be selected. $\mathbf{h}_i \in \mathbb{R}^{d \times T}$ is the output of i-th Whisper block, $\mathbf{H}, \mathbf{H}' \in \mathbb{R}^{D \times T}$ with $D = k * d$, k is the sum of the chosen Whipser block numbers. Thereafter, we use attentive statistics pooling [26] to extract speaker cues from frame-level features that are helpful for the speaker verification task. Finally, the speech vector is passed through batch normalization and a fully connected layer to obtain a low-dimensional speaker embedding representation.

2.3. LoRA for model adaptation

Compared to full fine-tuning, LoRA [24] optimizes storage and computational efficiency by modulating low-rank subspace parameters. In this paper, we apply LoRA to the Q (query), K (key), V (value), and O (output) weights of the Whisper model's multi-head attention, freezing the model's remaining parameters. For the model's weight matrix $W \in \mathbb{R}^{d \times k}$, the update mechanism is delineated as follows:

$$W + \Delta W = W + BA \tag{2}$$

where $B \in \mathbb{R}^{d \times r}$ is initialised with zeros, $A \in \mathbb{R}^{r \times k}$ is initialised with random Gaussian initialisation, with the rank $r \ll \min(d, k)$.

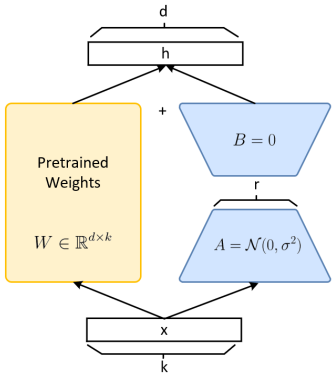

Figure 2: *LoRA. The pretrained weight parameters W are frozen, with only A and B being updated.*

3. Experimental setup

3.1. Dataset

To investigate the effectiveness of the proposed methods, we conduct experiments on the VoxCeleb1 [27] and CN-Celeb1 [28] datasets. The VoxCeleb1 dataset is a large-scale audio-

visual collection designed for speaker recognition tasks, containing over 150,000 utterances from 1,251 celebrities sourced from YouTube videos. The CN-Celeb1 dataset is a comprehensive audio dataset tailored for speaker recognition research. It consists of approximately 130,000 utterances by 1,000 Chinese celebrities from various domains, including but not limited to entertainment, sports, and politics.

During the training phase, we adhered to the official dataset partitions provided, training two distinct models on VoxCeleb1 and CN-Celeb1, respectively. In the testing phase for VoxCeleb1, we utilized the original trial list associated with VoxCeleb1 (VoxCeleb1-O) for evaluation. Similarly, for CN-Celeb1, we conducted tests using the official trial lists provided.

3.2. Model configuration

In our experiments, we selected two baseline models: ResNet34 and ECAPA-TDNN. Both baseline models, along with our proposed model, were implemented within the Wespeaker framework [29]. The specifications and parameter settings for each model are as follows:

ECAPA-TDNN [5]: ECAPA-TDNN is a modified version of the Time Delay Neural Network (TDNN), it features an enhanced Channel-wise Correlation Matrix Attention mechanism and time-domain x-vectors. Our implementation adopted the recommended parameters from Wespeaker.

ResNet34 [30]: ResNet-based r-vector is the winning system of VoxSRC 2019. It utilizes residual connections to alleviate the vanishing gradient problem during training, enabling the effective training of very deep neural networks. The parameter settings also follow recipes in Wespeaker.

Whisper-PMFA: Our proposed model Whisper-PMFA is built upon the pretrained Whisper large-v2 model, which consists of 32 Transformer blocks. For our implementation, we selected partial blocks from the Whisper large-v2 model to obtain the speaker embedding. The details of the experiments and results will be presented in Section 4.2. The dimensionality of the speaker embedding was set to 192.

3.3. Implementation details

To enhance the robustness of the systems, we applied three data augmentation techniques across all systems: additive noise augmentation from the MUSAN dataset [31], reverberation noise augmentation from the RIRs dataset [32], and speed perturbation [33] with 0.9 and 1.1 times speed changes.

During the feature extraction phase for the two baseline models, we used Wespeaker's standard method. This process entails selecting random 2-second clips from each speech sample and extracting 80-dimensional FBank features from these segments. The window length was set to 25 milliseconds with a frameshift of 10 milliseconds, and no voice activity detection was performed.

For our proposed model, named Whisper-PMFA, we utilized an 80-channel log magnitude Mel spectrogram consistent with the training of Whisper. During the training phase, we initially froze the parameters of the Whisper model and fine-tuned the remaining parameters for 4 epochs. This strategy was employed to prevent the pre-trained model from being fine-tuned in the wrong direction due to the random initialization of other parts. Subsequently, we conducted an overall fine-tuning of the entire model.

All models were trained using the AAM-Softmax loss [11, 34], with a margin of 0.2 and a scaling factor of 30. None of the models underwent large-margin fine-tuning during the training process.

3.4. Evaluation

We use cosine distance with AS-Norm for scoring. We report the system performance using two evaluation metrics: Equal Error Rate (EER) and Minimum Detection Cost Function (minDCF) with $P_{target} = 0.05$ and $C_{FA} = C_{Miss} = 1$.

4. Evaluation results and analysis

4.1. Performance evaluation and analysis

Experimental results are shown in Table 1. In the experiments, we conducted hierarchical fusion on layers 17-24, and the details regarding the selection of layers will be discussed in the next section.

As shown in Table 1, even though the two baseline models implemented within the Wespeaker framework achieved good results, our proposed model significantly outperforms the two baseline models on both the CN-Celeb1 and VoxCeleb1 datasets. On the VoxCeleb1 dataset, our model achieves an EER of 1.42%, representing a reduction of 24.3% compared to ResNet34 and a reduction of 29.0% compared to ECAPA-TDNN. On the CN-Celeb1 dataset, our model achieves an EER of 8.30%, representing a reduction of 10.4% compared to ResNet34 and a reduction of 17.9% compared to ECAPA-TDNN. MinDCF also shows corresponding improvements across all datasets.

The experimental results on multiple datasets demonstrate that although Whisper has not been optimized for speaker verification, our proposed method effectively utilizes the information learned and filtered from the pre-trained Whisper model. Whisper-PMFA leverages this information to capture distinctive speaker embeddings better, thereby significantly enhancing the effectiveness of speaker recognition tasks.

4.2. Layer selection experiment

The results of the layer selection experiment are presented in Table 2. We explored two different layer aggregation approaches: one using 8 layers and the other using 16 layers. In the 8-layer experiment, we divided the model into four parts: the front (layers 1-8), the front-middle (layers 9-16), the middle-back (layers 17-24), and the back (layers 25-32). In the 16-layer experiment, we selected adjacent parts based on the previous division, resulting in three groups of experiments: layers 1-16, layers 9-24, and layers 17-32 [2].

The experimental results for both the 8-layer and 16-layer experiments indicate that models utilizing 16 layers generally perform worse than those utilizing 8 layers. This suggests that increasing the number of layers does not necessarily improve model performance. Instead, the inclusion of excessively irrelevant information with an increased number of layers can hinder the model's pooling layers from effectively extracting relevant speaker-related cues. This ultimately leads to a degradation in performance.

Additionally, the experimental results indicate that compared to other parts, the mid-back part (layers 17-24) of the Whisper model contains more speaker-related cues. Fusion models based on these layers achieved the best performance.

[2]Due to the constraint of GPU memory, a maximum of 16 layers is supported in our experiments

Table 1: *Intra-Language and Cross-Language Performance Evaluation for Systems Trained on VoxCeleb1 and Cn-Celeb1*

Model	Score Norm	Training Dataset:VoxCeleb1				Training Dataset:CN-Celeb1			
		VoxCeleb1-O		CN-Celeb1-T		VoxCeleb1-O		CN-Celeb1-T	
		EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
ECAPA-TDNN	-	2.23	0.157	11.97	0.457	8.22	0.463	10.43	0.453
ResNet34	-	1.99	0.154	11.90	0.458	8.53	0.501	9.49	0.419
Whisper-PMFA(ours)	-	1.62	0.144	11.41	0.500	4.31	0.307	9.00	0.399
ECAPA-TDNN	AS-Norm	2.00	0.144	12.12	0.405	7.73	0.432	10.10	0.403
ResNet34	AS-Norm	1.88	0.154	11.42	**0.374**	7.87	0.454	9.27	0.385
Whisper-PMFA(ours)	AS-Norm	**1.42**	**0.121**	**11.24**	0.440	**3.91**	**0.270**	**8.30**	**0.358**

Table 2: *Layer selection experiment result*

Selected Layers	VoxCeleb1-O	
	EER(%)	minDCF
1-8	3.93	0.287
9-16	1.66	0.135
17-24	**1.42**	**0.121**
25-32	1.65	0.148
1-16	2.07	0.144
9-24	1.74	0.134
17-32	2.03	0.163

4.3. Cross-Language performance analysis

The experiments in the second and third columns of the table demonstrate the cross-linguistic performance of our proposed model compared to the baselines. When trained on Chinese (CN-Celeb1) and tested on English (VoxCeleb1), as well as the reverse scenario, our proposed model consistently outperforms both baseline models. This phenomenon is particularly notable when training in Chinese and testing in English, where our model achieves performance improvements of nearly 50% compared to the baselines. A critical factor contributing to this significant enhancement is utilizing the Whisper pre-trained model as the foundation of our proposed architecture. The Whisper model has been pre-trained on a diverse corpus that spans multiple languages, endowing it with robust cross-linguistic features. Our model, built upon this pre-trained base, inherits and optimizes these features for cross-lingual tasks. This advancement underscores the potential of pre-trained models in improving the adaptability and generalization of language processing systems across diverse linguistic contexts.

While our model achieves notable cross-lingual performance improvements in certain scenarios, its performance uplift is less marked when training in English (VoxCeleb1) and testing in Chinese (CN-Celeb1). This could be attributed to the complex nature of the CN-Celeb1 test set, which encompasses various domains such as drama, singing, and speeches. These diverse conditions likely impede the model's generalization in this context. Addressing this challenge will form the focus of our next research phase, aiming to enhance the model's adaptability and performance across varied linguistic domains.

4.4. Comparison with results in the literature

We compared our Whisper-PMFA with several state-of-the-art models, all models are trained on the VoxCeleb1 dataset. The experimental outcomes demonstrated that our proposed model achieved the highest performance among the models tested, this indicates the effectiveness of our method.

Table 3: *Comparison with published systems on VoxCeleb1*

Model	VoxCeleb1-O	
	EER(%)	minDCF
M-sv [35]	3.61	-
Inter-layer Adapter WavLM [22]	2.58	0.187
DROP-TDNN [36]	2.15	-
HuBERT-Base ECAPA-TDNN [37]	1.86	-
Whisper-PMFA(ours)	**1.42**	**0.121**

4.5. Investigation of more effective adaptation method

To address the efficiency reduction caused by the Whisper model's size, we integrated LoRA [24] as an alternative to full fine-tuning. According to the results shown in Table 4, this method significantly reduced the trainable parameters by approximately 45 times with only a minimal increase in EER by 0.2%. The Whisper-PMFA model, enhanced with LoRA, maintained high performance while achieving a parameter count comparable to other models, demonstrating a more effective adaptation method for large-scale models.

Table 4: *Comparison with the Integration of LoRA*

Model	# Params	EER(%)	minDCF
Whisper-PMFA	487.7M	1.42	0.121
Whisper-PMFA(LoRA)	10.9M	1.62	0.150

5. Conclusions

In this paper, the Whisper-PMFA framework is proposed, which leverages the rich speech knowledge embedded in the pre-trained Whisper ASR model to achieve high-quality speaker embedding extraction through selective feature aggregation. Experimental results on the widely used VoxCeleb1 and CN-Celeb1 datasets show that Whisper-PMFA can achieve notably lower EERs than the competing models and high cross-linguistic robustness. In addition, the LoRA adaptation approach is also investigated as a trial adaptation method, achieving a significant reduction in the number of trainable model parameters while maintaining competitive performance.

6. Acknowledgement

Shuai Wang is supported by Internal Project of Shenzhen Research Institute of Big Data under grant No.J00220230014.

7. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*, Calgary, 2018.

[2] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, Florence, 2014.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[4] W. Fang, Z. Yu, Y. Chen, T. Huang, T. Masquelier, and Y. Tian, "Deep residual learning in spiking neural networks," in *Proc. Neur IPS*, Montreal, 2021.

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, Shanghai, 2020.

[6] T. Zhou, Y. Zhao, and J. Wu, "Resnext and res2net structures for speaker verification," in *Proc. SLT*, virtual event, 2021.

[7] B. Gu, W. Guo, and J. Zhang, "Memory storable network based feature aggregation for speaker representation learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 643–655, 2023.

[8] Y. Tang, G. Ding, J. Huang, X. He, and B. Zhou, "Deep speaker embedding learning with multi-level pooling for text-independent speaker verification," in *Proc. ICASSP*, Brighton, 2019.

[9] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H.-y. Lee, and H. Meng, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," in *Proc. Interspeech*, Inchon, 2022.

[10] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, Long Beach, 2019.

[12] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proc. CVPR*, Hawaii, 2017.

[13] J. Thienpondt, B. Desplanques, and K. Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification," in *Proc. ICASSP*, Toronto, 2021.

[14] P. Matejka, O. Novotnỳ, O. Plchot, L. Burget, M. D. Sánchez, and J. Cernockỳ, "Analysis of score normalization in multilingual speaker recognition." in *Proc. Interspeech*, Stockholm, 2017.

[15] X. Hu, L. Chu, J. Pei, W. Liu, and J. Bian, "Model complexity of deep learning: A survey," *Knowledge and Information Systems*, vol. 63, pp. 2585–2619, 2021.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, Hawaii, 2023.

[17] 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.

[18] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[19] T. Berns, N. Vaessen, and D. A. van Leeuwen, "Speaker and language change detection using wav2vec2 and whisper," *arXiv preprint arXiv:2302.09381*, 2023.

[20] D. Cai, W. Wang, M. Li, R. Xia, and C. Huang, "Pretraining conformer with asr for speaker verification," in *Proc. ICASSP*, Rhodes, 2023.

[21] D. Liao, T. Jiang, F. Wang, L. Li, and Q. Hong, "Towards a unified conformer structure: from asr to asv task," in *Proc. ICASSP*, Rhodes, 2023.

[22] M. Sang and J. H. Hansen, "Efficient adapter tuning of pre-trained speech models for automatic speaker verification," *arXiv preprint arXiv:2403.00293*, 2024.

[23] J. Peng, T. Stafylakis, R. Gu, O. Plchot, L. Mošner, L. Burget, and J. Černockỳ, "Parameter-efficient transfer learning of pre-trained transformer models for speaker verification using adapters," in *Proc. ICASSP*, Rhodes, 2023.

[24] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *Proc. ICLR*, Vienna, 2021.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Neur IPS*, Long Beach, 2017.

[26] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech*, Hyderabad, 2018.

[27] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[28] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *Proc. ICASSP*, Barcelona, 2020.

[29] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *Proc. ICASSP*, Rhodes, 2023.

[30] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[31] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[32] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, New Orleans, 2017.

[33] S. Yang and M. Liu, "Data augmentation for speaker verification," in *Proc. EITCE*, Xiamen, 2022.

[34] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *Proc. APSIPA ASC*, Lanzhou, 2019.

[35] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in *Proc. Interspeech*, Brno, 2021.

[36] Q.-B. Hong, C.-H. Wu, and H.-M. Wang, "Decomposition and reorganization of phonetic information for speaker embedding learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1745–1757, 2023.

[37] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *Proc. ICASSP*, Singapore, 2022.