SSRFNET : STAGE-WISE SV-MIXER AND REDIMNET FUSION NETWORK FOR SPEAKER VERIFICATION

Kyowon Koo[1*], *Jungwoo Heo*[1*],*Seung-bin Kim*[1], *Hyun-seo Shin*[1], *Chan-yeong Lim*[1], *Jisoo Son*[1], *Kyung Wha Kim*[2], *and Ha-Jin Yu*[1†]

[1]University of Seoul, Republic of Korea
[2]Supreme Prosecutors' Office Republic of Korea

ABSTRACT

Speaker verification critically depends on effective speech representations. While spectrogram-based features have long dominated, recent studies show that combining them with raw waveform signals or embeddings from large self-supervised pre-trained models (PTMs) yields complementary information. However, existing dual-branch or attention-based fusion approaches often suffer from excessive complexity or limited integration capacity. In this work, we propose the Stage-wise SV-Mixer and RedimNet Fusion Network (SSRFNet), which retains spectrograms as the primary input and progressively injects PTM embeddings as auxiliary hints via a lightweight Projection-based Fusion Module(PFM). This design ensures compatibility between heterogeneous features without requiring a full-scale additional backend. By incorporating ReDimNet as an efficient spectrogram backbone and SV-Mixer as a compressed PTM, SSRFNet achieves both high accuracy and reduced model size, with experiments confirming superior verification performance and substantially improved practicality on VoxCeleb, VoxSRC, and VCMix benchmarks.

Index Terms**—** Speaker Verification, Pre-trained Model, SV-Mixer, ReDimNet

1. INTRODUCTION

Speaker verification(SV) is the task of verifying whether the speaker of a given utterance corresponds to an enrolled speaker. The effectiveness of SV systems critically depends on how speech is represented. For many years, spectrogram-based features have been the dominant choice, as they provide a stable time–frequency view of speech. Architectures such as TDNN, ECAPA-TDNN, and ReDimNet have been optimized to process spectrograms, and this alignment has contributed to steady improvements in SV performance [1–3].

To further enrich representations, earlier studies explored combining spectrograms with raw waveform features [4,5]. These works demonstrated that the two carry complementary information: spectrograms capture structured frequency–time characteristics, while waveform inputs preserve fine-grained acoustic features that may be lost in spectrogram transformations. Such hybrid designs showed that leveraging multiple representations can enhance robustness and discrimination in SV.

Following this line of research, recent advances have shifted attention from raw waveform encoders to large-scale self-supervised pre-trained models (PTMs), such as Wav2vec2.0 [6], HuBERT [7], and WavLM [8]. PTMs produce embeddings from raw waveforms, leveraging large-scale self-supervised pretraining to encode richer speaker-relevant representations than conventional features. Given the complementary characteristics of spectrograms and PTM embeddings, several works have explored their integration. For instance, Peng et al. [9] proposed a Dual-Branch ECAPA-TDNN (DBE), in which two parallel ECAPA-TDNNs respectively process spectrograms and PTM embeddings, while intermediate layers exchange auxiliary information. The approach proves effective; however, the doubled model size constrains its practical use. In contrast, Li et al. [10] introduced a Fine-Grained Fusion Module (FGFM), which employs attention-based feature fusion to combine spectrogram and PTM embeddings as input to a single ECAPA-TDNN. This lightweight design improves efficiency but, due to its simplicity, still leaves ample room for improvement compared to multi-branch structures such as DBE. These observations suggest that PTM–spectrogram fusion holds great potential, and realizing its full benefits will require continued exploration of the trade-off between performance and efficiency.

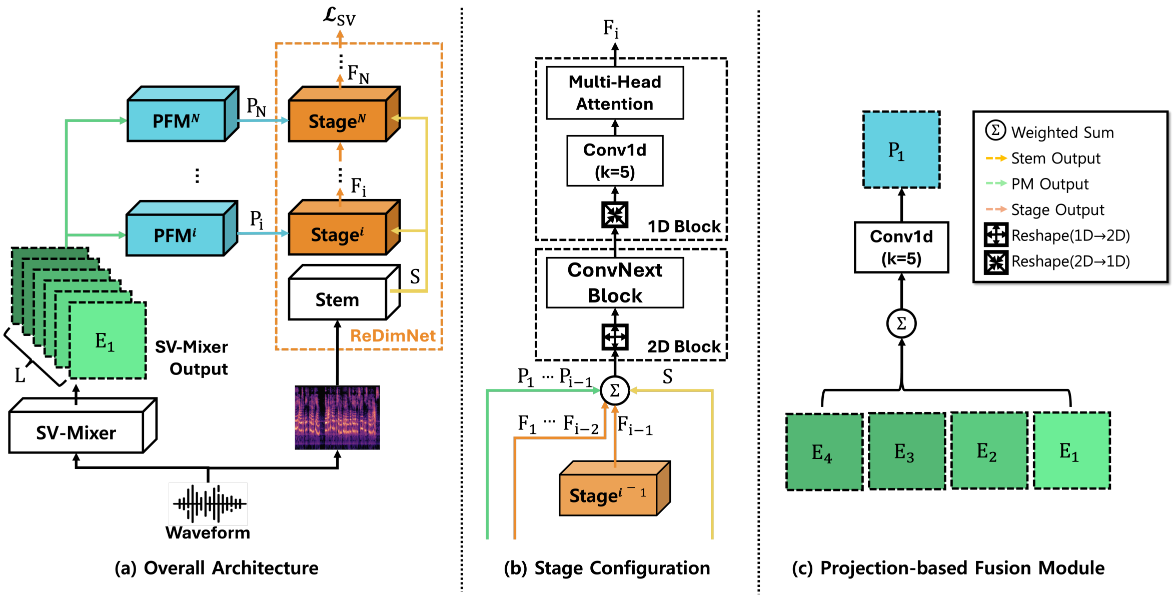
In this paper, we propose the Stage-wise SV-Mixer and RedimNet Fusion Network (SSRFNet), a framework designed to achieve both efficiency and compatibility in spectrogram–PTM fusion. The central idea of SSRFNet is to resolve the mismatch between waveform-based PTM embeddings and spectrogram-oriented backbones through a stage-wise hint injection strategy. Drawing inspiration from DBE, PTM embeddings are progressively injected into multiple backend layers as auxiliary hints, enabling their information to be gradually aligned with spectrogram processing. While DBE uses spectrogram features as hints, SSRFNet retains spectrograms as the primary input and integrates PTM embeddings as auxiliary information to enrich the representation. For this purpose, we introduce the Projection-based Fusion Module(PFM), a lightweight mechanism that fuses PTM embeddings into the backbone. Our design builds on the insight from FGFM that feature fusion can be achieved effectively without adding a separate full-scale backend. Through this progressive adaptation, PTM features are smoothly aligned with spectrogram processing, enabling an integration that achieves both strong performance and practical efficiency beyond existing fusion methods.

To further enhance the framework, SSRFNet adopts ReDimNet [3] as its spectrogram backbone. ReDimNet has recently emerged as a lightweight yet state-of-the-art architecture in speaker verification, ensuring that the proposed framework is both efficient and performance-leading. In addition, to mitigate the heavy com-

[*]Equal contribution

[†]Corresponding author

This work was supported by the Supreme Prosecutor's Office Research Grant in 2025 (research title: Development of fake voice detection technology robust in new voice generation technology and speaker recognition)



(a) Overall Architecture

(b) Stage Configuration

(c) Projection-based Fusion Module

Fig. 1: The overall structure of SSRFNet

putational cost of large PTMs, SSRFNet leverages SV-Mixer [11], a compressed model distilled from PTMs that provides competitive representations with far lower complexity. With this combination, SSRFNet delivers even higher verification accuracy while reducing model size by approximately 25% compared to prior approaches.

2. SSRFNET

The objective of this paper is to design a speaker verification system that is both lightweight and high-performing. To achieve this goal, we incorporate ReDimNet, a state-of-the-art SV backend system, together with SV-Mixer, a lightweight alternative to large-scale PTMs. We also devise an integration strategy that effectively combines these two components.

2.1. Overall Architecture

SSRFNet consists of three principal components: SV-Mixer, Projection-based Fusion Module, and ReDimNet. SV-Mixer extracts PTM hidden states directly from the waveform, providing a lightweight substitute for large-scale PTMs that achieves competitive representational quality with substantially lower complexity. ReDimNet functions as the spectrogram-based backend for speaker verification; through reshape-based 1D and 2D processing combined with residual aggregation, it delivers both efficiency and expressive capacity, making it well suited to the stage-wise fusion strategy. The Projection-based Fusion Module is designed to bridge SV-Mixer and ReDimNet.

As illustrated in Fig. 1 (a), the input waveform is converted into a spectrogram and processed by the ReDimNet stem module, which is composed of convolutional layer, to yield the primary representation $\mathbf{S} \in \mathbf{R}^{\mathbf{T \times H}}$. This representation is then refined as it passes through the six stages of ReDimNet. In parallel, SV-Mixer produces L hidden states $\mathbf{E}_1, \dots, \mathbf{E}_L \in \mathbf{R}^{T \times H}$ through CNN and multi-layer perceptron(MLP) layers, which are transformed by the PFM into projected embeddings $\mathbf{P}_i \in \mathbb{R}^{T \times H}$ and sequentially injected into the N successive ReDimNet stages as auxiliary features that provide a complementary perspective. After all stages, the spectrogram features $\mathbf{S}$, the projected embeddings $\mathbf{P}_i$, and the stage outputs $\mathbf{F}'_i$ are aggregated, and statistical pooling is applied to derive the final speaker embedding.

This architecture preserves the structural stability of spectrogram-based backbones while progressively enriching them with PTM-derived features, yielding an integration strategy that achieves both efficiency and performance.

2.2. Stage Configuration

The stage configuration module refines input representations by alternating between 2D and 1D convolutional blocks, thereby progressively extracting speaker-related features. In the original ReDimNet, each stage receives spectrogram features $\mathbf{S}$ together with the outputs of preceding stages $\mathbf{F}_1, \dots, \mathbf{F}_{i-1}$ and produces the i-th stage representation $\mathbf{F}_i \in \mathbb{R}^{T \times H}$ through weighted summation followed by 1D and 2D convolutional blocks. In our framework, this process is extended by additionally incorporating projected PTM features $\mathbf{P}_1, \dots, \mathbf{P}_{i-1}$ generated by the PFM. By augmenting the inputs in this manner, the system preserves the original processing flow of ReDimNet while injecting auxiliary features that provide a complementary perspective to spectrograms.

Formally, the input to stage i is defined as the combination of spectrogram features $\mathbf{S}$, projected embeddings $\mathbf{P}_1, \dots, \mathbf{P}_i$, and outputs of preceding stages $\mathbf{F}_1, \dots, \mathbf{F}_{i-1}$. Each stage block follows the ReDimNet design, consisting of a ConvNeXt-based 2D block that captures local time–frequency structures and a 1D block with multi-head attention and convolution to refine long-range temporal dependencies. The resulting output $\mathbf{F}_i$ is forwarded to the next stage and accumulated in the residual pool. This progressive input design encourages speaker information to be accumulated and refined across stages, ultimately enhancing representational quality and improving system performance.

2.3. Projection-based Fusion Module

Raw waveform–based PTM features complement spectrogram representations [12]. However, the two lie in distinct representational spaces, making naive fusion ineffective. Prior research on Fine-Grained Feature Modulation (FGFM) [10] demonstrated that even simple attention-based fusion at the input level can yield meaningful improvements, indicating that complex architectural modifications are not always required. Building on this insight, we introduce the Projection-based Fusion Module(PFM), a lightweight mechanism specifically designed to transform PTM embeddings into a form that is compatible with spectrogram-based processing.

To enable structured processing, the L hidden states extracted from the PTM are partitioned into N overlapping groups of size K. Let s_i denote the start index of the i-th group, obtained by evenly spacing N integers in $[1, \ldots, L{-}K]$ (e.g., $s_1 = 1$ and $s_N = L{-}K$). Each group collects $\mathbf{E}_{s_i}, \mathbf{E}_{s_i+1}, \ldots, \mathbf{E}_{s_i+K}$ and is processed by the Projection-based Fusion Module as follows:

$$\mathbf{G}_i = \sum_{j=1}^{K} \mathbf{w}_j \odot \mathbf{E}_{s_i+j}, \quad i = 1, \ldots, N, \tag{1}$$

(2)

where $\mathbf{w}_j \in \mathbb{R}^K$ are learnable parameters and $\odot$ denotes broadcasting followed by element-wise multiplication. This design allows the model to assign varying importance to different hidden states, as observed in [13], thereby enabling effective utilization of critical layer information. The aggregated vector is then projected through a 1D convolution followed by Layer Normalization to align feature dimensions:

$$\mathbf{P}_i = \text{LayerNorm}\big(\text{Conv1D}(\mathbf{G}_i)\big), \tag{3}$$

Through this process, the Projection-based Fusion Module preserves speaker-relevant information and provides a stable foundation for stage-wise fusion.

3. EXPERIMENTAL SETUP

3.1. Dataset

We train on VoxCeleb2 [14], a large-scale corpus widely used for speaker verification. To improve robustness, data augmentation is applied using additive noise and music from MUSAN [15] and reverberation simulated with RIR filters [16], each with a probability of 80%. Evaluation follows the VoxCeleb1 [17] protocols O, E, and H, with additional tests on VoxSRC2023 [18] and VCMix [19] to assess generalization.

3.2. Baselines

All baseline systems are constructed under a unified configuration to ensure fairness in comparison. In this setup, the student network is defined as a $L = 12$ layers SV-Mixer, and the teacher network is WavLM-Large [8]. The backend model is ReDimNet, implemented with the B2 version, and embeddings are obtained through attentive statistics pooling [20] and optimized using the AAM-Softmax loss [21] with a margin of 0.2 and a scale of 30. For consistency, all reported results involving ReDimNet are obtained without applying Large-Margin Fine-Tuning(LMFT) [22] or AS-Norm [23].

Based on this common configuration, we re-implemented several representative fusion architectures. Li et al. [10] proposed Fine-Grained Feature Modulation, which integrates spectrogram features and PTM embeddings through a simple attention mechanism at the input level. we reproduced this approach by applying FGFM to SV-Mixer combined with a ReDimNet backend. Peng et al. [9] introduced a dual-branch strategy, where spectrograms and PTM embeddings are processed by separate backbones and merged through an Attention Fusion Module. we replicated this design using two parallel ReDimNet backbones. In addition, we considered a simplified variant of our own framework, referred to as sum fusion. This baseline corresponds to SSRFNet without the PFM, preserving stage-wise fusion while directly combining spectrogram and PTM embeddings through element-wise addition.

Table 1: Stepwise performance improvements on VoxCeleb1-O, from the baseline ReDimNet to the full proposed framework. † denotes our implementation based on the official code, without hard-margin tuning.

Model	EER (%)	minDCF
ReDimNet(Spectrogram)†	0.744	0.052
ReDimNet(PTM only)	0.935	0.061
ReDimNet(spec + PTM, dual-branch)	1.148	0.069
ReDimNet(Spec + PTM, FGFM)	0.755	0.042
SSRFNet(w/o Projection)	0.723	0.047
SSRFNet	**0.601**	**0.041**

3.3. Implementation Details

Our framework follows SV-Mixer [11], with knowledge distillation performed under the OS-KDFT framework [24]. This scheme jointly optimizes supervised training and knowledge distillation while keeping the teacher network in inference mode.

Input features are 80-dimensional Mel spectrograms extracted with a 25 ms window and a 10 ms hop size. Training uses 3-second crops with a batch size of 128. Optimization follows a two-stage procedure: AdamW warm-up with cosine annealing for 5k steps, followed by SGD fine-tuning with Nesterov momentum (0.9) and cosine decay. The learning rate is scheduled from 1×10^{-3} to 1×10^{-4} during warm-up and from 1×10^{-1} to 1×10^{-4} during training, with weight decay set to 2×10^{-5}. Training proceeds for 304k steps, with validation every 2k steps and early stopping.

In SSRFNet, $N = 4$ denotes the number of ReDimNet stages into which PTM embeddings are injected, and $K = 4$ the number of PTM embeddings in PFM. All experiments are implemented in PyTorch on four NVIDIA RTX A5000 GPUs. Code and pretrained models will be released.[1]

4. RESULTS

4.1. Baseline vs Proposed

Table 1 summarizes the stepwise performance changes from the baseline ReDimNet to the proposed framework. The baseline ReDimNet trained on spectrogram inputs achieves an EER of 0.74% and a minDCF of 0.052. When the spectrogram input is replaced with PTM features, performance degrades substantially (0.93% EER and 0.061 minDCF), confirming that ReDimNet is not directly compatible with PTM embeddings. We also tested conventional strategies that combine spectrogram and PTM features through simple summation or dual-branch fusion, which have been commonly used in prior work, but these configurations did not yield improvements.

[1] `https://github.com/dayflys/SSLBackend-2026`

Table 2: Comparison with Existing Systems (Sorted by Model Size).

Model	Params	Vox1-O	Vox1-E	Vox1-H
Over 300M parameter of PTMs-based SV system				
wavLM Large / ECAPA [8]	322.0M	0.617	0.662	1.318
+ LMFT	322.0M	0.383	0.480	0.986
HuBERT Large / ECAPA [8]	322.0M	0.808	0.822	1.678
+ LMFT	322.0M	0.590	0.654	1.227
Wav2Vec2.0 Large / ECAPA [13]	320.0M	0.803	0.729	1.394
+ LMFT	320.0M	0.585	0.625	1.138
UnispeechSAT Large / ECAPA [13]	320.0M	0.696	0.685	1.433
+ LMFT	320.0M	0.537	0.569	1.180
wavLM large / LAP [25]	317.8M	0.370	0.500	1.010
wavLM large / MHFA [26]	316.2M	0.490	0.700	1.700
+ LMFT	316.2M	0.490	0.800	1.700
wavLM large / CA-MHFA [27]	316.3M	0.550	0.620	1.180
+ LMFT	316.3M	0.420	0.480	0.960
Under 300M parameter of PTMs-based SV system				
wavLM base+ / ECAPA [8]	100.0M	0.840	0.928	1.758
wavLM base+ / ECAPA†	100.0M	0.850	0.953	1.959
UnispeechSAT Base / ECAPA [13]	100.0M	1.005	0.933	1.866
HuBERT base / ECAPA [8]	100.0M	0.989	1.068	2.216
wav2vec2.0 base / MHFA [26]	95.4M	1.920	2.040	4.440
+ LMFT	95.4M	1.750	1.930	4.100
wavLM base+ / MHFA [26]	96.2M	0.660	0.890	1.900
+ LMFT	96.2M	0.590	0.790	1.730
wavLM base+ / CA-MHFA [27]	96.3M	0.700	0.720	1.450
+ LMFT	96.3M	0.590	0.650	1.300
wavLM base+ / LAP [25]	96.1M	0.610	0.770	1.490
SEED (2025) [28]	105.6M	0.810	0.970	1.920
SV-Mixer / ECAPA [11]	80.3M	0.776	0.950	1.842
SSRFNet	**74.2M**	**0.601**	**0.800**	**1.446**

In contrast, incorporating PTM features as auxiliary inputs while retaining spectrograms as the main input leads to a modest gain: the proposed model without the PFM reduces EER to 0.72% and minDCF to 0.047. Finally, adding the PFM yields the best performance, achieving 0.601% EER and 0.041 minDCF. This stepwise progression indicates that spectrograms continue to serve as a critical component for ReDimNet, whereas straightforward integration of PTM features remains insufficient. By contrast, the proposed projection-based fusion provides a principled means of leveraging the complementary characteristics of PTMs, thereby yielding substantial gains in overall system performance.

4.2. Comparison Across Evaluation Protocols

Table 2 presents a comparison between the proposed framework and existing PTM-based speaker verification systems under multiple evaluation protocols. Systems are grouped by parameter scale to highlight the trade-off between capacity and accuracy.

Large-scale models (over 300M parameters), such as WavLM Large and HuBERT Large combined with ECAPA or MHFA [26], achieve strong performance but require significant computational resources. Their accuracy is further improved when logit-level fusion methods such as LMFT are applied. However, these configurations demand both high parameter counts and additional inference overhead.

In the sub-300M range, various base models combined with ECAPA, MHFA, CA-MHFA [27], or LAP [25] achieve moderate accuracy, with performance generally enhanced by LMFT. Nevertheless, many of these systems remain above 90M parameters, and their EERs on VoxCeleb1-O typically exceed 0.65%.

Table 3: Comparison of generalization performance on VCMix and VoxSRC23.

Model	Params	VCMix	VoxSRC23
wavLM base+ / ECAPA†	100.0M	2.81	5.66
SEED (2025) [28]	105.6M	**2.29**	4.94
SV-Mixer / ECAPA [11]	80.3M	3.29	4.89
SSRFNet	74.2M	2.47	**4.14**

The proposed system, by contrast, uses only 74.2M parameters yet achieves an EER of 0.601% on VoxCeleb1-O, substantially outperforming other systems of comparable or larger size without relying on LMFT. On more challenging protocols such as VoxCeleb1-H, the proposed system maintains competitive performance (1.446%).

These results indicate that the proposed hybrid integration of SV-Mixer and ReDimNet provides a favorable balance between accuracy and efficiency. Compared to most prior systems, which depend on both large-scale PTMs and additional fusion strategies, our approach attains competitive performance across diverse evaluation settings with a compact and standalone design.

4.3. Comparison Across Various Datasets

Table 3 compares the proposed system with existing PTM-based approaches on VCMix and VoxSRC23, which were not included in training, to evaluate generalization to unseen and diverse domain conditions. The wavLM base+ combined with ECAPA, with 100M parameters, records 2.810% EER on VCMix and 5.660% on VoxSRC23. SEED [28], a recently introduced system with 105.6M parameters, achieves the best performance with 2.29% on VCMix and 4.94% on VoxSRC23. SV-Mixer combined with ECAPA [11] uses 90.3M parameters and records 3.29% and 4.89% EER on the two datasets, respectively.

In comparison, the proposed system uses only 74.2M parameters yet achieves 2.467% on VCMix and 4.136% on VoxSRC23. The results on VoxSRC23 demonstrate that Projection-based fusion provides strong generalization under diverse domain conditions, while the comparison with SEED highlights that our model achieves comparable accuracy on VCMix with substantially fewer parameters, underscoring its efficiency.

5. CONCLUSION

This paper introduced the Stage-wise SV-Mixer and RedimNet Fusion Network (SSRFNet), a unified framework that reconciles spectrogram-oriented backbones with waveform-based PTM embeddings through progressive projection and integration. By retaining spectrograms as the primary input and injecting PTM features via a lightweight Projection Module, SSRFNet resolves the structural mismatch between heterogeneous representations without requiring dual-branch backbones or heavy attention mechanisms. Extensive experiments on VoxCeleb, VoxSRC, and VCMix confirmed that SSRFNet surpasses both spectrogram-only and PTM-only baselines as well as conventional fusion strategies, while operating with a smaller parameter footprint.

These results highlight that stage-wise projection fusion provides a practical and scalable alternative to heavy multi-branch designs, offering an effective balance between accuracy and efficiency. As a natural extension, future work will explore extending the framework to diverse combinations of PTMs and backbones, thereby examining the generality of lightweight fusion designs for speaker verification.

6. REFERENCES

[1] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification,” in *Interspeech 2020*, 2020, pp. 3830–3834.

[2] J. Thienpondt and K. Demuynck, “Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.

[3] I. Yakovlev, R. Makarov, A. Balykin, P. Malov, A. Okhotnikov, and N. Torgashov, “Reshape dimensions network for speaker recognition,” in *Interspeech 2024*, 2024, pp. 3235–3239.

[4] D. Salvati, C. Drioli, and G. L. Foresti, “A late fusion deep neural network for robust speaker identification using raw waveforms and gammatone cepstral coefficients,” *Expert Systems with Applications*, vol. 222, p. 119750, 2023.

[5] B. Ma, C. Xu, and Y. Zhang, “A novel speech feature fusion algorithm for text-independent speaker recognition,” *Multimedia Tools and Applications*, vol. 83, no. 24, pp. 64 139–64 156, 2024.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[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] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “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.

[9] S. Peng, W. Guo, H. Wu, Z. Li, and J. Zhang, “Fine-tune pre-trained models with multi-level feature fusion for speaker verification,” in *Interspeech 2024*, 2024, pp. 2110–2114.

[10] Y. Li, W. Guan, H. Huang, S. Miao, Q. Su, L. Li, and Q. Hong, “Efficient integrated features based on pre-trained models for speaker verification,” in *Interspeech 2024*, 2024, pp. 2140–2144.

[11] J. Heo, H. seo Shin, C. yeong Lim, K. won Koo, S. bin Kim, J. Son, and H.-J. Yu, “Sv-mixer: Replacing the transformer encoder with lightweight mlps for self-supervised model compresison in speaker verification,” in *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2025, to appear.

[12] K. Zhang, Z. Hua, R. Lan, Y. Guo, Y. Zhang, and G. Xu, “Multi-view collaborative learning network for speech deepfake detection,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 1, pp. 1075–1083, Apr. 2025.

[13] 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 *ICASSP 2022*, 2022, pp. 6147–6151.

[14] J. Chung *et al.*, “Voxceleb2: Deep speaker recognition,” *arXiv preprint arXiv:1806.05622*, 2018.

[15] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” 2015.

[16] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký, “Building and evaluation of a real room impulse response dataset,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[17] A. Nagrani *et al.*, “Voxceleb: a large-scale speaker identification dataset,” in *Proc. Interspeech*, 2017.

[18] J. Huh, J. S. Chung, A. Nagrani, A. Brown, J.-w. Jung, D. Garcia-Romero, and A. Zisserman, “The voxceleb speaker recognition challenge: A retrospective,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3850–3866, 2024.

[19] H. Heo *et al.*, “Rethinking session variability: Leveraging session embeddings for session robustness in speaker verification,” in *Proc. ICASSP*, 2024.

[20] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” in *Interspeech 2018*, 2018, pp. 2252–2256.

[21] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694.

[22] Y. Liu, L. He, and J. Liu, “Large margin softmax loss for speaker verification,” in *Interspeech 2019*, 2019, pp. 2873–2877.

[23] S.-C. Yin, R. Rose, and P. Kenny, “Adaptive score normalization for progressive model adaptation in text independent speaker verification,” in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4857–4860.

[24] J. Heo, C. yeong Lim, J. ho Kim, H. seo Shin, and H.-J. Yu, “One-step knowledge distillation and fine-tuning in using large pre-trained self-supervised learning models for speaker verification,” in *Interspeech 2023*, 2023, pp. 5271–5275.

[25] Jin Sob Kim and Hyun Joon Park and Wooseok Shin and Sung Won Han, “Rethinking Leveraging Pre-Trained Multi-Layer Representations for Speaker Verification,” in *Interspeech 2025*, 2025, pp. 3713–3717.

[26] J. Peng, O. Plchot, T. Stafylakis, L. Mošner, L. Burget, and J. Černocký, “An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 555–562.

[27] J. Peng, L. Mošner, L. Zhang, O. Plchot, T. Stafylakis, L. Burget, and J. Černocký, “Ca-mhfa: A context-aware multi-head factorized attentive pooling for ssl-based speaker verification,” in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[28] K. Nam, J. Heo, J. weon Jung, G. Park, C. Jung, H.-J. Yu, and J. S. Chung, “SEED: Speaker Embedding Enhancement Diffusion Model,” in *Interspeech 2025*, 2025, pp. 3718–3722.