The following publication Z. Jin et al., "Distilling Attention Knowledge for Speaker Verification," ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2026, pp. 16447-16451 is available at https://doi.org/10.1109/ICASSP55912.2026.11464971.

DISTILLING ATTENTION KNOWLEDGE FOR SPEAKER VERIFICATION

Zezhong Jin[1], *Shujie Liu*[2], *Zhe Li*[3], *Chong-Xin Gan*[1], *Zilong Huang*[1], *Man-Wai Mak*[1], *Kong Aik Lee*[1]

[1] The Hong Kong Polytechnic University, Hong Kong SAR, China
[2] Microsoft Research Asia
[3] The University of Hong Kong, Hong Kong SAR, China

ABSTRACT

In recent years, the adoption of pre-trained speech models as feature extractors for speaker verification (SV) has surged. To reduce model complexity, knowledge from these pre-trained models has been transferred to compact student models, enabling them to achieve performance unattainable by conventional methods. While label-level and feature-level knowledge distillation (KD) have both yielded promising results in SV, attention-level KD remains under explored. Different from the other two methods, attention-level KD aims to guide the student network to focus more on the informative and important parts of the feature. To explore the attention-level KD in SV tasks, in this paper, we develop two distinct attention maps that enable the student model to learn the teacher's attention along both the frequency and temporal dimensions, named Frequency-attentive KD (FREQ-AKD) and Temporal-attentive KD (TEMPO-AKD), respectively. Experiments conducted on the VoxCeleb and CN-Celeb datasets demonstrate the effectiveness of both FREQ-AKD and TEMPO-AKD.

Index Terms— Speaker verification; knowledge distillation; attention maps

1. INTRODUCTION

Speaker verification (SV) has greatly benefited from the rapid development of deep neural networks [1–5]. In particular, the introduction of refined model architectures such as ECAPA-TDNN [6], ResNet [7], and CAM++ [8] has led to substantial improvements in SV performance. Some researchers have leveraged a large amount of unlabeled data for self-supervised learning [9, 10] and subsequently fine-tuned learned models using a small set of labeled data, thereby achieving performance that surpasses that of conventional supervised approaches. In addition, it was found that employing large self-supervised models (e.g., Wav2vec 2.0 [11], HuBERT [12], and WavLM [13]) to extract frame-based speech features, rather than relying on traditional Mel-Frequency Cepstral Coefficients (MFCCs) or filterbank features yields excellent results for speaker verification. However, the aforementioned methods rely on large-capacity models trained with extensive amounts of data, which makes deployment on edge devices challenging. One solution to this problem is Knowledge Distillation (KD) [14], which trains a lightweight student model to mimic a powerful teacher model, achieving performance close to that of the teacher model.

Depending on the type of transferred knowledge, KD methods fall into three groups: those based on transferring label [15–18], features [19,20], and attention maps [21]. Label-level KD minimizes the Kullback–Leibler (KL) divergence between the output distributions of the teacher and student networks. Feature-level KD aims to align the student's intermediate features with those of the teacher. Attention-level KD leverages teacher's attention maps for the student to learn which parts of the input should be focused on.

Although label-level and feature-level KDs have demonstrated promising performance in SV, research on attention-level KD remains limited. Compared to label-level and feature-level KDs, the principle of attention-level KD is more intuitive. The reason is that it can guide the student network to focus on the more informative and important frequencies and frames by encouraging the student network to mimic the attention maps transferred from the teacher.

As a bridge of the knowledge transfer, the selection of attention maps are crucial, which largely determines the success of attention-level KD. In this paper, we focus on attention-level KD and design two different attention maps to guide the student model in focusing on the important parts of both the frequencies and temporal dimensions. For the frequency dimension, we utilize the weights from the squeeze and excitation module to form the attention map, which represents the importance of each frequency. For the temporal dimension, we employ the weights from the attentive statistics pooling module to form an attention map, indicating the significance of each frame. Our contributions are summarized as follows:

1. We introduce attention knowledge distillation to speaker verification. Unlike prior works [22, 23] that focus on multi-head self-attention, our approach targets the internal attention mechanisms of mainstream ASV back-

THIS WORK WAS SUPPORTED BY THE RGC OF HONG KONG SAR, GRANT NOS. POLYU 15210122 AND 15228223.

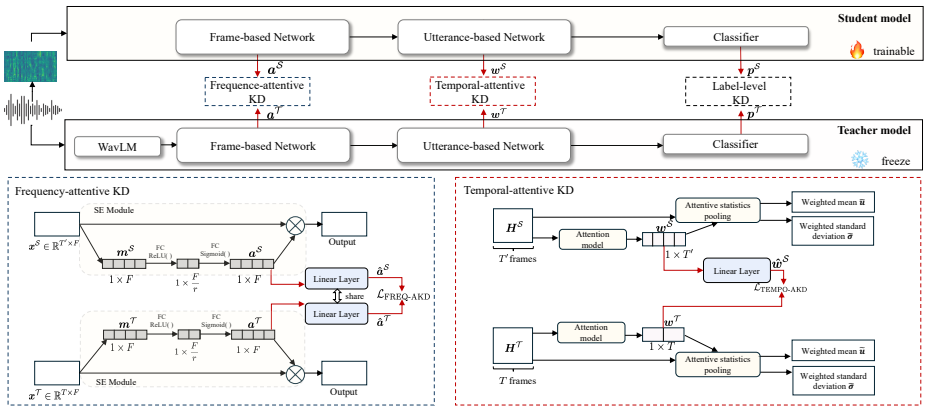

Fig. 1: The Framework of attention-based KD. Due to differences in the input features, the number of frames in the features of the teacher network T is not equal to the number of frames in the features of the student network T'.

bones, such as SE blocks and pooling layers.

2. We design two different attention maps to guide the student model to focus on the informative and important parts across different dimensions (frequency and temporal).

3. We show that attention-based KD can achieve significantly better performance compared to baselines with similar model sizes, and it can lead to students whose performance is as good as that of the teacher.

2. METHODOLOGY

2.1. Traditional Knowledge Distillation

Define a dataset as $\mathcal{D} = \{(\mathbf{s}, y)\}$, where $\mathbf{s}$ is a speech signal after data augmentation and y is the corresponding label. That is, $y \in \{1, \ldots, C\}$, where C is the number of speakers in the training set. After feeding $\mathbf{s}$ into the teacher and student models, we obtain the logits vector $\boldsymbol{q}^{\mathcal{T}} \in \mathbb{R}^C$ and $\boldsymbol{q}^{\mathcal{S}} \in \mathbb{R}^C$, where $\mathcal{T}$ and $\mathcal{S}$ represent the teacher and student models, respectively. After the softmax function, we obtain the respective speaker posterior probabilities $\boldsymbol{p} \in \mathbb{R}^C$. KD facilitates knowledge transfer from a large, fixed teacher model to a smaller student model by minimizing the Kullback-Leibler (KL) divergence between their probability distributions:

$$\mathcal{L}_{\text{KD}} = \sum_{i=1}^{C} p_i^{\mathcal{T}} \log \left(\frac{p_i^{\mathcal{T}}}{p_i^{\mathcal{S}}} \right), \tag{1}$$

where $i \in \{1, \ldots, C\}$ indexes the classes. Notably, all the losses in this paper are sample-wise.

2.2. Frequency-Attentive Knowledge Distillation

Defining the kind of information to be distilled and the construction of attention maps plays a pivotal role in the effectiveness of attention-based KD. In speech processing, both the frequency and temporal dimensions encapsulate critical information relevant to SV. To fully leverage the information in the speech features, we propose two distinct attention mechanisms: frequency-centric attention maps and temporal-centric attention maps. These maps are specifically designed to enable the student model to learn the teacher's attention distributions across the frequency and temporal dimensions, facilitating a more effective knowledge transfer.

As shown in Fig. 1, for frequency-attentive KD, we utilize the Squeeze-and-Excitation (SE) module to capture frequency-wise dependencies, which explicitly indicates the relative importance of each frequency channel. Given an input feature map $\boldsymbol{x} \in \mathbb{R}^{T \times F}$, where T denotes the number of frames and F denotes the number of frequency bins. We first perform global average pooling across the temporal dimension to produce a frequency-wise statistic:

$$\boldsymbol{m} = [m_1, m_2, \ldots, m_F], \quad m_f = \frac{1}{T} \sum_{t=1}^{T} x_{t,f}. \tag{2}$$

Next, $\boldsymbol{m}$ is processed through two fully connected (FC) layers with ReLU and sigmoid activations to obtain the frequency attention vector $\boldsymbol{a}$:

$$\boldsymbol{a} = \sigma(\boldsymbol{W}_2(\text{ReLU}(\boldsymbol{W}_1 \boldsymbol{m} + \boldsymbol{b}_1) + \boldsymbol{b}_2)), \tag{3}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{\frac{F}{r} \times F}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{F \times \frac{F}{r}}$ are the weight matrices of the two FC layers, $\boldsymbol{b}_1 \in \mathbb{R}^{\frac{F}{r}}$ and $\boldsymbol{b}_2 \in \mathbb{R}^F$ are the corresponding bias terms, σ denotes the sigmoid activation function, and r is a hyperparameter controlling the compression of frequency information (we set $r = 8$ following the standard ECAPA-TDNN configuration [6]). We extract the frequency attentive weights computed by the SE module for each model, obtaining their frequency attention maps:

$$\boldsymbol{a}^{\mathcal{T}} = [a_1^{\mathcal{T}}, a_2^{\mathcal{T}}, \ldots, a_F^{\mathcal{T}}] \quad \text{and} \quad \boldsymbol{a}^{S} = [a_1^{S}, a_2^{S}, \ldots, a_F^{S}]. \tag{4}$$

To reduce the gap between the attention maps of the student and teacher networks, we first project them into a hidden

space using $\boldsymbol{W}_3$, apply a ReLU activation, and then transform them back to the original dimension with $\boldsymbol{W}_4$:

$$\begin{aligned}\hat{\boldsymbol{a}}^{\mathcal{T}} &= \boldsymbol{W}_4(\text{ReLU}(\boldsymbol{W}_3\boldsymbol{a}^{\mathcal{T}} + \boldsymbol{b}_3) + \boldsymbol{b}_4),\\ \hat{\boldsymbol{a}}^{\mathcal{S}} &= \boldsymbol{W}_4(\text{ReLU}(\boldsymbol{W}_3\boldsymbol{a}^{\mathcal{S}} + \boldsymbol{b}_3) + \boldsymbol{b}_4).\end{aligned} \tag{5}$$

Finally, we define the frequency-attentive KD loss:

$$\mathcal{L}_{\text{FREQ-AKD}} = \mathcal{D}(\hat{\boldsymbol{a}}^{\mathcal{T}}, \hat{\boldsymbol{a}}^{S}), \tag{6}$$

where $\mathcal{D}(\cdot,\cdot)$ represents the chosen loss function (e.g., MSE or KL divergence) between the projected frequency-level attention maps $\hat{\boldsymbol{a}}^{\mathcal{T}}$ and $\hat{\boldsymbol{a}}^{S}$. When $\mathcal{D}(\cdot,\cdot)$ is the KL divergence, we apply Softmax to $\hat{\boldsymbol{a}}^{\mathcal{T}}$ and $\hat{\boldsymbol{a}}^{S}$ to normalize them into probability distributions before computing the loss.

2.3. Temporal-attentive Knowledge Distillation

In temporal-attentive KD, we utilize the attention scores from the attentive statistics pooling module as the attention map, allowing the student model to learn the temporal importance assigned by the teacher model. During training, the speech segments have fixed length. The frame-level features of the teacher and student models are respectively denoted as $\boldsymbol{H}^{\mathcal{T}} = [\boldsymbol{h}_1^{\mathcal{T}}, \boldsymbol{h}_2^{\mathcal{T}}, \ldots, \boldsymbol{h}_T^{\mathcal{T}}]$ and $\boldsymbol{H}^{\mathcal{S}} = [\boldsymbol{h}_1^{\mathcal{S}}, \boldsymbol{h}_2^{\mathcal{S}}, \ldots, \boldsymbol{h}_{T'}^{\mathcal{S}}]$, where $\boldsymbol{h}_t \in \mathbb{R}^D$ represents the frame-level features at timestep t. The sequence lengths in the two networks are different because the student uses FBank features (typically computed at a frame shift of 10ms), while the teacher processes raw waveform. At a 16kHz sampling rate, this results in a 2:1 length ratio ($T' = 2T$).

To assign importance to each frame, an attention score e_t is computed as follows:

$$e_t = \boldsymbol{v}^{\mathrm{T}} \tanh(\boldsymbol{W}_5\boldsymbol{h}_t + \boldsymbol{b}_5), \tag{7}$$

where $\boldsymbol{W}_5 \in \mathbb{R}^{d\times D}$ is a learnable weight matrix, $\boldsymbol{b}_5 \in \mathbb{R}^d$ is a bias term, $\boldsymbol{v} \in \mathbb{R}^d$ is a trainable projection vector, and the activation function $\tanh(\cdot)$ introduces non-linearity.

The attention scores are then normalized using the softmax function to obtain the frame-level attention weights:

$$w_t = \frac{\exp(e_t)}{\sum_{\tau=1}^{T}\exp(e_\tau)}, \quad t = \{1,\ldots,T\}, \tag{8}$$

where w_t represents the importance assigned to frame t.

Since $T' = 2T$, we use a linear layer to transform the temporal dimension of $\boldsymbol{w}^{\mathcal{S}} = [w_1^{\mathcal{S}}, \ldots, w_{T'}^{\mathcal{S}}]$ into T, denoted as $\hat{\boldsymbol{w}}^{\mathcal{S}}$. Note that this mapping is feasible because the input length is fixed during training. The temporal-attentive KD loss is defined as:

$$\mathcal{L}_{\text{TEMPO-AKD}} = \mathcal{D}(\boldsymbol{w}^{\mathcal{T}}, \hat{\boldsymbol{w}}^{\mathcal{S}}). \tag{9}$$

Finally, the total loss for SV is formulated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CLS}} + \mathcal{L}_{\text{KD}} + \lambda(\mathcal{L}_{\text{FREA-AKD}} + \mathcal{L}_{\text{TEMPO-AKD}}), \tag{10}$$

where $\mathcal{L}_{\text{CLS}}$ is the classification loss computed using ground-truth labels, such as the AAM-Softmax loss, and $\mathcal{L}_{\text{KD}}$ is the label-level knowledge distillation loss in (1). Additionally, λ serves as a hyperparameter that balances the contribution of attention KD losses.

3. EXPERIMENTAL SETUP

3.1. Datasets

For training, we employed the VoxCeleb2 development set [24], which includes 5,994 speakers. For evaluation, we utilized the VoxCeleb1 test sets [25] namely, Vox-O, Vox-E, and Vox-H. Additionally, to further validate our approach, we conducted experiments on the CN-Celeb dataset [26], a multi-genre Mandarin corpus compiled from Chinese open media. In this case, the development sets from CN-Celeb1 and CN-Celeb2, encompassing 2,793 speakers, served as the training data, while the CN-Celeb evaluation set (CN-eval), containing 18,000 utterances from 196 speakers, was used for testing.

We followed the data augmentation strategy from Kaldi's recipes [27]. Specifically, we enriched the training data by adding noise, music, and babble from the MUSAN dataset [28] and simulating reverberated speech using the RIR corpus [29].

3.2. Network Training

To maintain consistency with [17], we adopted WavLM Large [13] with ECAPA-TDNN [6] as our teacher model for both VoxCeleb and CN-Celeb. For the student model, we employed an ECAPA-TDNN with 512 channels, which has significantly fewer parameters than the teacher model.

For VoxCeleb, each speech signal was randomly cropped to 2 seconds and augmented with a probability of 0.6 during training, while for CN-Celeb, each audio waveform was randomly segmented into 3 seconds with an augmentation probability of 0.8. After augmentation, we extracted 80-dimensional filter bank (Fbank) features using a 25-ms frame length and a 10-ms frame-shift. These Fbank features were then used as input for the student model, whereas the augmented raw waveform was fed into the teacher network. We optimized the student model using SGD, employing a linear learning rate warmup over the first 6 epochs, which increased the learning rate from 5e-4 to 0.15. The ECAPA-TDNN was trained with a batch size of 256. We set the hyperparameter λ to 0.01, which will be further discussed in Section 4.

4. RESULTS AND DISSCUSSIONS

4.1. Main Results

Table 1 presents our results on the VoxCeleb1 test set. According to Row 4, Attention KD improves performance on Vox1-O, Vox1-E and Vox1-H. Specifically, Attention KD

Table 1: Performance of different KD methods on the VoxCeleb1 and CN-Celeb test set. "–" denotes using the classification loss only.

System	Row	# Params (M)	Distillation Method	Vox1-O		Vox1-E		Vox1-H		CN-eval	
				EER(%) ↓	minDCF ↓	EER(%) ↓	minDCF ↓	EER(%) ↓	minDCF ↓	EER (%) ↓	minDCF ↓
Teacher model WavLM-TDNN [13]	1	316.62	–	0.43	–	0.54	–	1.15	–	7.33	0.421
Student model	2	7.34	–	1.02	0.106	1.41	0.162	2.26	0.257	8.87	0.462
	3	7.34	KD [14]	0.90	0.098	0.99	0.117	1.96	0.195	8.21	0.451
	4	7.76	Attention KD	**0.76**	**0.094**	**0.91**	**0.111**	**1.91**	**0.195**	**7.70**	**0.448**

Table 2: Compared with different advanced speaker verification systems on Vox1-O.

System	# Params (M)	Vox1-O	
		EER (%) ↓	minDCF ↓
Whisper-SV [30]	5.34	1.71	0.211
ResNet34 (256) [31]	7.03	1.42	-
NEMO Small [32]	15.88	0.88	0.137
ECAPA-TDNN [6]	20.77	0.87	0.107
Attention KD (ours)	7.76	**0.76**	**0.094**

Table 3: Exploring the importance of frequency-level (FREQ) and temporal-level (TEMPO) attention in Attention KD.

Distillation Method	TEMPO-AKD	FREQ-AKD	EER (%) on Vox1-O
Attention KD	✗	✓	0.79
	✓	✗	0.85
	✓	✓	**0.76**

achieves an EER of 0.76% on Vox1-O, representing a 16% improvement over KD, which demonstrates the effectiveness of Attention KD.

To further validate the effectiveness of our method, we conducted experiments on the CN-Celeb dataset. According to Table 1, Attention KD achieves an EER of 7.70% on CN-eval. Despite the student model having significantly fewer parameters than the teacher model, it attains performance close to that of the teacher. Moreover, compared to KD, Attention KD improves performance on CN-Celeb by 6%, demonstrating its effectiveness and robustness.

Table 2 compares our method with recent state-of-the-art speaker verification systems on VoxCeleb1. All these methods were trained on VoxCeleb2. As shown in Table 2, our method achieves the best performance despite having a smaller model size, further highlighting its effectiveness.

4.2. Ablation Studies

To explore the importance of frequency and frame-level attention KD in our approach, we conducted experiments using the levels individually and both levels simultaneously. According to Table 3, the best performance is achieved when both frequency-level and frame-level attention KD are applied together. Using only frequency-level attention KD results in an EER of 0.79%, while using only frame-level attention KD yields an EER of 0.85%. This indicates that frequency-level attention KD plays a more crucial role.

We also explored different loss functions for attention KD, such as KL divergence loss and MSE loss. We found

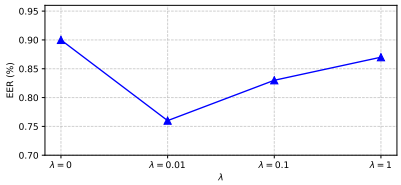

Fig. 2: Effect of λ on performance

Table 4: Exploring the impact of loss types in Attention KD.

Distillation Method	Attention KD Loss Type	EER (%) on Vox1-O
Attention KD	KL-Divergence	0.88
	MSE	**0.79**

Table 5: The role of the linear layer $\boldsymbol{W}_3$(Eq. 5) in Frequency-Level (FREQ) attention knowledge distillation

Distillation Method	MLP W_3	EER (%) on Vox1-O
Attention KD	✗	0.81
	✓	**0.76**

that using MSE loss achieved the best performance based on Table 4. Therefore, in this paper, we adopt MSE loss for $\mathcal{D}$. Additionally, we explored the impact of different values of λ on system performance, as shown in Fig. 2. Results indicate that the best performance is achieved when $\lambda = 0.01$. We also explored the effect of mapping the student and teacher attention maps to the same feature space in FREQ using $\boldsymbol{W}_3$. Table 5 shows that using $\boldsymbol{W}_3$ yields better performance than not using it. This further demonstrates the effectiveness of mapping their attention maps to the same feature space.

5. CONCLUSIONS

We have applied attention-based knowledge distillation to speaker verification. By leveraging the attention maps from the squeeze-and-excitation and attentive statistic pooling modules, important frequency and temporal information are highlighted. This enables the student model to focus on the most crucial information, thereby enhancing the effectiveness of KD. Our experiments demonstrate the effectiveness of attention KD, showing that the student model, despite having significantly fewer parameters than the teacher model, can achieve performance close to that of the teacher.

6. REFERENCES

[1] Junjie Li, Kong Aik Lee, Duc-Tuan Truong, Tianchi Liu, and Man-Wai Mak, "Xi+: Uncertainty supervision for robust speaker embedding," *arXiv preprint arXiv:2509.05993*, 2025.

[2] Zezhong Jin, Shubhang Desai, Xu Chen, Biyi Fang, Zhuoyi Huang, Zhe Li, Chong-Xin Gan, Xiao Tu, Man-Wai Mak, Yan Lu, et al., "Trink: Ink generation with transformer network," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 4857–4864.

[3] Zhe Li, Man-Wai Mak, Mert Pilanci, and Helen Meng, "Mutual information-enhanced contrastive learning with margin for maximal speaker separability," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 2961–2972, 2025.

[4] Zhe Li, Man-Wai Mak, Jen-Tzung Chien, Mert Pilanci, Zezhong Jin, and Helen Meng, "Disentangling speech representations learning with latent diffusion for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 3896–3907, 2025.

[5] Zhe Li, Man-Wai Mak, Mert Pilanci, Hung-yi Lee, Chong-Xin Gan, Jiabao Sheng, and Helen Meng, "Towards a unified view of parameter-efficient speech pretrained models for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2026.

[6] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *in Proc. Interspeech*, pp. 3830–3834, 2020.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[8] Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen, "CAM++: A fast and efficient network for speaker verification using context-aware masking," *in Proc. Interspeech, 2023,*, pp. 5301–5305.

[9] Zezhong Jin, Youzhi Tu, and Man-Wai Mak, "Self-supervised learning with multi-head multi-mode knowledge distillation for speaker verification," in *Proc. Interspeech 2024*, 2024, pp. 4723–4727.

[10] Zezhong Jin, Youzhi Tu, and Man-Wai Mak, "W-GVKT: Within-global-view knowledge transfer for speaker verification," in *Proc. Interspeech*, 2024, pp. 3779–3783.

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

[12] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 2021,*, pp. 3451–3460.

[13] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing, 16, 6, 2022,*, pp. 1505–1518.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," in *in NIPS Deep Learning and Representation Learning Workshop*, 2015.

[15] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2022, pp. 11953–11962.

[16] Chong-Xin Gan, Youzhi Tu, Zezhong Jin, Man-Wai Mak, and Kong Aik Lee, "Grouped knowledge distillation with adaptive logit softening for speaker recognition," in *Proc, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[17] Duc-Tuan Truong, Ruijie Tao, Jia Qi Yip, Kong Aik Lee, and Eng Siong Chng, "Emphasized non-target speaker knowledge in knowledge distillation for automatic speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10336–10340.

[18] Zezhong Jin, Youzhi Tu, Chong-Xin Gan, Man-Wai Mak, and Kong-Aik Lee, "Adversarially adaptive temperatures for decoupled knowledge distillation with applications to speaker verification," *Neurocomputing*, p. 129481, 2025.

[19] Zhiyuan Peng, Xuanji He, Ke Ding, Tan Lee, and Guanglu Wan, "Label-free knowledge distillation with contrastive loss for light-weight speaker recognition," in *International Symposium on Chinese Spoken Language Processing*, 2022, pp. 324–328.

[20] Zezhong Jin, Youzhi Tu, Zhe Li, Zilong Huang, Chong-Xin Gan, and Man-Wai Mak, "Denoising student features with diffusion models for knowledge distillation in speaker verification," in *Proc, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[21] Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin, "Class attention transfer based knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11868–11877.

[22] Jeong-Hwan Choi, Joon-Young Yang, and Joon-Hyuk Chang, "Efficient lightweight speaker verification with broadcasting cnn-transformer and knowledge distillation training of self-attention maps," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[23] Victoria Mingote, Antonio Miguel, Alfonso Ortega, and Eduardo Lleida, "Class token and knowledge distillation for multi-head self-attention speaker verification systems," *Digital Signal Processing, 133, 2023,*, p. 103859.

[24] J Chung, A Nagrani, and A Zisserman, "Voxceleb2: Deep speaker recognition," *in Proc. Interspeech, 2018,*, pp. 1086–1090.

[25] A Nagrani, J Chung, and A Zisserman, "Voxceleb: a large-scale speaker identification dataset," *in Proc. Interspeech, 2017,*, pp. 2616–2620.

[26] Lantian Li, Ruiqi Liu, Jiawen Kang, Yue Fan, Hao Cui, Yunqi Cai, Ravichander Vipperla, Thomas Fang Zheng, and Dong Wang, "CN-Celeb: multi-genre speaker recognition," *Speech Communication, 137, 2022,*, pp. 77–91.

[27] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[28] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[29] Marco Jeub, Magnus Schafer, and Peter Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 16th International Conference on Digital Signal Processing*, 2009.

[30] Li Zhang, Ning Jiang, Qing Wang, Yue Li, Quan Lu, and Lei Xie, "Whisper-sv: Adapting whisper for low-data-resource speaker verification," *Speech Communication*, vol. 163, pp. 103103, 2024.

[31] Joon Son Chung, Arsha Nagrani, Ernesto Coto, Weidi Xie, Mitchell McLaren, Douglas A Reynolds, and Andrew Zisserman, "Voxsrc 2019: The first voxceleb speaker recognition challenge," *arXiv preprint arXiv:1912.02522*, 2019.

[32] Danwei Cai and Ming Li, "Leveraging asr pretrained conformers for speaker verification through transfer learning and knowledge distillation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.