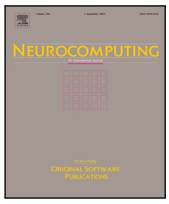
Neurocomputing 624 (2025) 129481

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Adversarially adaptive temperatures for decoupled knowledge distillation with applications to speaker verification

Zezhong Jin, Youzhi Tu, Chong-Xin Gan, Man-Wai Mak *, Kong-Aik Lee

Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region of China

ARTICLE INFO

Communicated by R.C. Guido

Keywords:
Speaker verification
Knowledge distillation
Adaptive temperature
Adversarial learning

ABSTRACT

Knowledge Distillation (KD) aims to transfer knowledge from a high-capacity teacher model to a lightweight student model, thereby enabling the student model to attain a level of performance that would be unattainable through conventional training methods. In conventional KD, the loss function's temperature that controls the smoothness of class distributions is fixed. We argue that distribution smoothness is critical to the transfer of knowledge and propose an adversarial adaptive temperature module to set the temperature dynamically during training to enhance the student's performance. Using the concept of decoupled knowledge distillation (DKD), we separate the Kullback–Leibler (KL) divergence into a target-class term and a non-target-class term. However, unlike DKD, we adversarially update the temperature coefficients of the target and non-target classes to maximize the distillation loss. We named our method **A**dversarially **A**daptive **T**emperature for **DKD** (AAT-DKD). Our approach demonstrates improvements over KD methods across three test sets of Voxceleb1 for two student models (x-vector and ECAPA-TDNN). Specifically, compared to the traditional KD and DKD, our method achieves a remarkable reduction of 17.78% and 11.90% in EER using ECAPA-TDNN speaker embedding. Moreover, our method performs well on CN-Celeb and VoxSRC21, further highlighting its robustness and effectiveness across different datasets.

1. Introduction

With the development of deep neural networks, numerous high-performance network architectures have been applied to automatic speaker verification (ASV) and have achieved excellent performance. These architecture include x-vector [1], ResNet [2], ECAPA-TDNN [3], and CAM++ [4]. Additionally, many researchers have employed large self-supervised models (Wav2vec 2.0 [5], HuBERT [6], and WavLM [7]) to extract frame-based speech features for downstream ASV and automatic speech recognition (ASR) tasks. In many cases, speaker embedding networks using features from these large models outperform the networks that use Mel-Frequency Cepstral Coefficients (MFCCs) or filterbank features [8–10]. However, powerful networks typically demand more computation and memory resources, making deployment difficult. Knowledge Distillation (KD) [11] offers a solution to this problem. KD allows a lightweight student model to mimic a more powerful teacher model through the process of knowledge distillation, enabling the student model to inherit performance close to the teacher model.

A speaker embedding neural network consists of three parts. The first part is an encoder, which transforms the input speech signal into frame-level features. This is followed by a temporal aggregation layer [12–14] that combines these frame-level features into a condensed representation of the entire input sequence. In the final stage of the neural network, a decoder is responsible for mapping the utterance-level representations to speaker classes [15–17]. The decoder is built from a series of fully connected layers, including a bottleneck layer that plays a key role in extracting the speaker embeddings.

In ASV, knowledge distillation between the teacher and student embedding networks can be done at two different levels: embedding level [18,19] and label level [20,21]. The former endeavors to align the student's intermediate features with those of the teacher in the embedding space. The latter minimizes the Kullback–Leibler (KL) divergence between the output distributions of the teacher and student networks with a fixed temperature.

In [20], the authors discovered the importance of non-target speakers in KD and used the idea of decoupled knowledge distillation (DKD) [22] to emphasize the class probabilities of non-target speakers. The authors of [21] used a similar strategy, but the class probabilities were divided into top-K and non-top-K groups rather than the target and non-target parts.

* Corresponding author.
E-mail addresses: 22123484r@connect.polyu.hk (Z. Jin), youzhi.tu@connect.polyu.hk (Y. Tu), chong-xin.gan@connect.polyu.hk (C.-X. Gan), enmwmak@polyu.edu.hk (M.-W. Mak), kong-aik.lee@polyu.edu.hk (K.-A. Lee).

https://doi.org/10.1016/j.neucom.2025.129481
Received 16 October 2024; Received in revised form 27 December 2024; Accepted 16 January 2025
Available online 23 January 2025

The methods above use a constant temperature during training, typically with a default value of 1.0 in SV. The temperature, which regulates the smoothness of the output distributions from both the teacher and student networks, plays a crucial role in determining the effectiveness of knowledge distillation [11]. Using a fixed temperature may not be suitable throughout all stages of training. Therefore, we aim to adapt the temperature so that it can adjust to different training stages.

However, when directly adapting the temperature, it quickly converges to a value that deviates far from the reasonable range, making this approach ineffective (as demonstrated in Section 5.2). In [23], the authors used adversarial training to adapt the temperature and obtained remarkable performance. Inspired by this strategy, we propose an **A**dversarially **A**daptive **T**emperature for **DKD** (AAT-DKD) to enable the model to adapt the temperatures adversarially under the DKD framework. The temperatures are adapted through reversal gradient descent during the training of the student network, aiming to maximize the distillation loss between the teacher and student.

Unlike using a fixed reversal coefficient in the Gradient Reversal Layer (GRL) [23,24], we introduce a dynamic reversal coefficient that varies with each mini-batch, depending on the overall sample quality of the mini-batch. In our method, the target probability from the teacher model is used to determine its sample quality and control the degree of contribution to adversarial learning. Therefore, using a dynamic reversal coefficient strengthens the contribution of high-quality samples and reduces the effect of low-quality ones, which benefits learning more robust speaker embeddings.

On the other hand, in DKD, the distributions of target-speaker knowledge distillation (TSKD) and non-target-speaker knowledge distillation (NSKD) are different. Therefore, using the same temperature for both KDs is unreasonable. To address this issue, we use separate temperatures for TSKD and NSKD.

Our contributions are summarized as follows:

1. We propose an Adversarially Adaptive Temperature for the DKD framework, allowing the student and teacher to adapt the temperature parameters. To the best of our knowledge, it is the first attempt to make the temperature parameters in KD learnable in ASV.
2. In the process of adversarial learning, we replace the constant reversal coefficient with a dynamic one that varies according to the quality of the mini-batch samples.
3. We integrate our method into the state-of-the-art knowledge distillation framework, DKD, and set different learnable temperatures for different parts of DKD.
4. We demonstrate the effectiveness and robustness of our method through extensive experiments on diverse datasets, including Voxceleb, VoxSrc 2021, and CN-Celeb.
5. To better demonstrate the generalizability of our method, we also validated it on the image classification task using the CIFAR-100 dataset.

2. Related works

Knowledge distillation (KD) was initially proposed in [11]. Its purpose is to transfer the knowledge from a pre-trained teacher model to a more lightweight student model. Traditional KD achieves this by matching the output distributions of the teacher and student models. Enabling a student model to learn from a high-capacity teacher model can significantly improve the performance of the student model.

In recent years, KD has been widely applied to ASV. For instance, Truong et al. [20] applied label-level DKD [22] to emphasize non-target speaker information during distillation and achieved promising results. In [25], a self-knowledge distillation framework was proposed. It utilizes an auxiliary self-teacher network to distill its own refined knowledge without the need of a pre-trained teacher network. In [26], a cross-modal knowledge distillation framework was proposed, where a more discriminative face recognition model was utilized as a teacher to guide a speech model to improve ASV performance. In [27], the authors explored a combination of Knowledge Distillation and Random Erasing data augmentation to enhance the generalization ability and robustness of text-dependent speaker verification systems.

To achieve embedding-level KD, the authors in [18] minimized the mean square errors and cosine distances between the embeddings of the same utterance from the teacher and student networks. The authors in [19] extended the idea in [18] to the contrastive loss that pulls the student's and teacher's embeddings of the same utterance together and pushes their embeddings from different utterances apart.

In self-supervised learning, there are numerous knowledge distillation-based methods, such as distillation with no label (DINO) [28], which distills information from the teacher network to the student network using positive sample pairs. In [29], the authors improved DINO's performance by introducing distillation between the same global views of each sample. In [30], the authors adopted multi-mode knowledge distillation to further improve the ASV performance by introducing multi-head projection layers. In [31], the authors proposed a prototype division strategy to iteratively refine prototypes in the projection space, addressing the fixed and insufficient prototype issue in DINO for ASV.

Despite their promising performance, the methods mentioned above set the temperature parameter to a fixed value. In [32], it was reported that a low temperature causes the distillation to focus on the teacher model's highest logit. In contrast, a high temperature flattens the distribution, making the distillation pay attention to all the logits. The study in [33] also demonstrated that altering the temperature can affect the effectiveness of knowledge distillation. Therefore, using a fixed temperature for KD is not an ideal solution. The study in [23] employed adversarial learning to allow the model to adapt the temperature parameter, achieving good results in image classification. Inspired by this, we also adopted an adversarial learning approach to make the temperature parameter adaptive. However, one difference between [23] and our proposed method is that we additionally introduce a dynamic adversarial factor to regulate the intensity of adversarial learning, accounting for the quality of individual samples. Furthermore, based on the popular DKD framework, we use separate temperatures to better adapt the distributions of different parts of DKD.

3. Methodology

3.1. Conventional knowledge distillation

Consider a C-way classification task and define the dataset as $\mathcal{D} = \{(\mathbf{x}, y)\}$, where $\mathbf{x}$ is a speech signal after data augmentation and y is the corresponding label. After feeding $\mathbf{x}$ into the teacher and student models, we obtain the logits vector $\mathbf{q}^{\mathcal{T}} \in \mathbb{R}^C$ and $\mathbf{q}^{S} \in \mathbb{R}^C$, where $\mathcal{T}$ and S represent the teacher and student models, respectively. Take the teacher as an example, the posterior probability of the i-th class is

$$p_i^{\mathcal{T}} = \frac{e^{q_i^{\mathcal{T}}/\tau}}{\sum_{j=1}^{C} e^{q_j^{\mathcal{T}}/\tau}}, \quad i = 1, 2, \ldots, C, \tag{1}$$

where τ is a temperature parameter controlling the smoothness of the posterior distribution. KD transfers knowledge from a large frozen teacher model to a small student model by minimizing the KL divergence between the probability outputs of the teacher and student models:

$$L_{\mathrm{KD}} = \sum_{i=1}^{C} p_i^{\mathcal{T}} \log\left(\frac{p_i^{\mathcal{T}}}{p_i^{S}}\right). \tag{2}$$

Notably, all the losses in this paper are sample-wise.

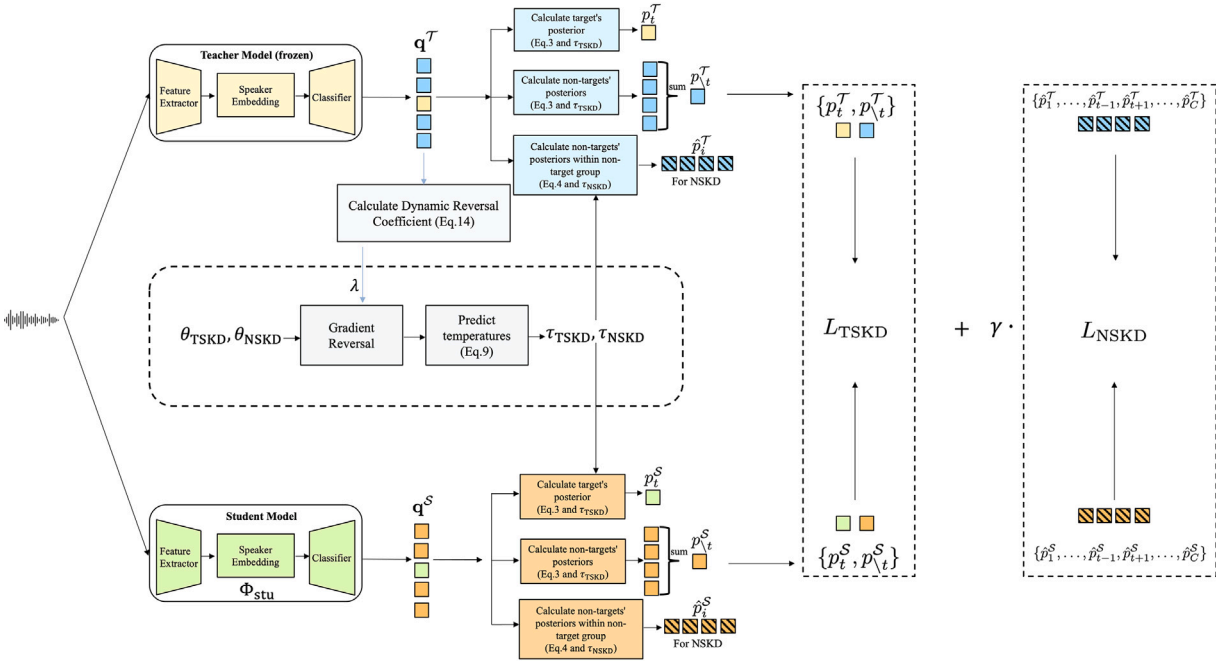


Fig. 1. Schematic of the proposed **A**dversarially **A**daptive **T**emperature for **DKD** (AAT-DKD). TSKD stands for Target Speaker Knowledge Distillation, and NSKD stands for Non-Target Speaker Knowledge Distillation. λ is set to 1.0 during the forward pass, and it follows Eq. (14) during backpropagation. For notational simplicity, we have omitted the classification loss.

3.2. Decoupled knowledge distillation

For decoupled knowledge distillation (DKD), we separate the predictions into a target and a non-target groups. We define $\mathbf{b} = \{p_t, p_{\setminus t}\} \in \mathbb{R}^2$ to represent the probabilities of the target and non-target groups:

$$p_t = \frac{e^{q_t/\tau}}{\sum_{j=1}^{C} e^{q_j/\tau}}, \quad p_{\setminus t} = \frac{\sum_{k=1,k\neq t}^{C} e^{q_k/\tau}}{\sum_{j=1}^{C} e^{q_j/\tau}}, \quad t \in \{1,\dots,C\}. \tag{3}$$

The probability of a non-target speaker $i \neq t$ over all non-target speakers is defined as:

$$\hat{p}_i = \frac{e^{q_i/\tau}}{\sum_{j=1,j\neq t}^{C} e^{q_j/\tau}}, \quad i \in \{1,\dots,t-1,t+1,\dots,C\}. \tag{4}$$

The original KD loss in Eq. (2) can be split into the target speaker and non-target speaker parts:

$$L_{\mathrm{KD}} = p_t^{\mathcal{T}} \log\left(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}\right) + \sum_{i=1,i\neq t}^{C} p_i^{\mathcal{T}} \log\left(\frac{p_i^{\mathcal{T}}}{p_i^{\mathcal{S}}}\right). \tag{5}$$

According to Eqs. (3) and (4), we have $p_i = \hat{p}_i \times p_{\setminus t}$. Thus, we can rewrite Eq. (5) as

$$\begin{aligned} L_{\mathrm{KD}} &= p_t^{\mathcal{T}} \log\left(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}\right) + \sum_{i=1,i\neq t}^{C} p_{\setminus t}^{\mathcal{T}} \hat{p}_i^{\mathcal{T}} \log\left(\frac{p_{\setminus t}^{\mathcal{T}} \hat{p}_i^{\mathcal{T}}}{p_{\setminus t}^{\mathcal{S}} \hat{p}_i^{\mathcal{S}}}\right) \\ &= p_t^{\mathcal{T}} \log\left(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}\right) + \sum_{i=1,i\neq t}^{C} p_{\setminus t}^{\mathcal{T}} \hat{p}_i^{\mathcal{T}} \log\left(\frac{\hat{p}_i^{\mathcal{T}}}{\hat{p}_i^{\mathcal{S}}}\right) \\ &\quad + \sum_{i=1,i\neq t}^{C} p_{\setminus t}^{\mathcal{T}} \hat{p}_i^{\mathcal{T}} \log\left(\frac{p_{\setminus t}^{\mathcal{T}}}{p_{\setminus t}^{\mathcal{S}}}\right). \end{aligned} \tag{6}$$

Because $p_{\setminus t}^{\mathcal{S}}$ and $p_{\setminus t}^{\mathcal{T}}$ are independent of i and $\sum_{i=1,i\neq t}^{C} \hat{p}_i^{\mathcal{T}} = 1$, we can simplify Eq. (6) to

$$L_{\mathrm{KD}} = \underbrace{p_t^{\mathcal{T}} \log\left(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}\right) + p_{\setminus t}^{\mathcal{T}} \log\left(\frac{p_{\setminus t}^{\mathcal{T}}}{p_{\setminus t}^{\mathcal{S}}}\right)}_{L_{\mathrm{TSKD}}} + \underbrace{p_{\setminus t}^{\mathcal{T}} \sum_{i=1,i\neq t}^{C} \hat{p}_i^{\mathcal{T}} \log\left(\frac{\hat{p}_i^{\mathcal{T}}}{\hat{p}_i^{\mathcal{S}}}\right)}_{(1-p_t^{\mathcal{T}})L_{\mathrm{NSKD}}}, \tag{7}$$

where $p_{\setminus t}^{\mathcal{T}} = 1 - p_t^{\mathcal{T}}$. The first two terms of Eq. (7) constitute the target speaker KD (TSKD) loss and the last term denotes the non-target speaker KD (NSKD) loss.

Since the teacher is a well-trained model, the value of $p_t^{\mathcal{T}}$ tends to be relatively large, typically approaching 1.0. This phenomenon could suppress the NSKD part in the total loss in Eq. (7). Therefore, we decoupled the knowledge distillation by replacing $(1 - p_t^{\mathcal{T}})$ with a hyperparameter γ as follows:

$$L_{\mathrm{DKD}} = \underbrace{p_t^{\mathcal{T}} \log\left(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}\right) + p_{\setminus t}^{\mathcal{T}} \log\left(\frac{p_{\setminus t}^{\mathcal{T}}}{p_{\setminus t}^{\mathcal{S}}}\right)}_{L_{\mathrm{TSKD}}} + \gamma \underbrace{\sum_{i=1,i\neq t}^{C} \hat{p}_i^{\mathcal{T}} \log\left(\frac{\hat{p}_i^{\mathcal{T}}}{\hat{p}_i^{\mathcal{S}}}\right)}_{L_{\mathrm{NSKD}}}. \tag{8}$$

3.3. Adversarially adaptive temperature for decoupled knowledge distillation

In previous works [20,25,34], τ was usually set to 1.0. However, a fixed τ is not suitable for all training stages, which can reduce the model's generalization ability. To address this limitation, we treat τ as a variable predicted by a function with learnable parameters. Additionally, we use different temperatures τ_{TSKD} and τ_{NSKD} for TSKD and NSKD, respectively.

The overall process is shown in Fig. 1. The temperatures τ_{TSKD}, and τ_{NSKD} are predicted from a function parameterized by learnable θ_{TSKD} and θ_{NSKD}, respectively:

$$\begin{aligned} \tau_{\mathrm{TSKD}} &= \alpha_1 + \alpha_2(\sigma(\theta_{\mathrm{TSKD}})) \\ \tau_{\mathrm{NSKD}} &= \alpha_1 + \alpha_2(\sigma(\theta_{\mathrm{NSKD}})), \end{aligned} \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function. We use two hyperparameters, α_1 and α_2, to ensure the non-negativity of τ_{TSKD} and τ_{NSKD} and set their range to $[\alpha_1, \alpha_1 + \alpha_2]$.

Inspired by GANs [23,24], we propose to adversarially learn the parameters θ_{TSKD} and θ_{NSKD}, which predict suitable temperatures τ_{TSKD} and τ_{NSKD} by using Eq. (9). We define $L_{\mathrm{AAT-DKD}}$ in the same form as Eq. (8). However, unlike L_{DKD}, it employs different temperatures for L_{TSKD} and L_{NSKD}. These temperatures are derived through Eq. (9), using learnable parameters θ_{TSKD} and θ_{NSKD}. The student module $\boldsymbol{\Phi}_{\mathrm{stu}}$ and the two learnable parameters θ_{TSKD} and θ_{NSKD} play a min–max

game with the following equation:

$$\begin{aligned}&\min_{\boldsymbol{\Phi}_{\text{stu}}}\max_{\theta_{\text{TSKD}},\theta_{\text{NSKD}}} L_{\text{AAT-DKD}}(\mathbf{x};\boldsymbol{\Phi}_{\text{stu}},\theta_{\text{TSKD}},\theta_{\text{NSKD}})\\&=\min_{\boldsymbol{\Phi}_{\text{stu}}}\max_{\theta_{\text{TSKD}},\theta_{\text{NSKD}}} L_{\text{TSKD}}(\mathbf{x};\boldsymbol{\Phi}_{\text{stu}},\theta_{\text{TSKD}})\\&\qquad+\gamma L_{\text{NSKD}}(\mathbf{x};\boldsymbol{\Phi}_{\text{stu}},\theta_{\text{NSKD}}).\end{aligned}\tag{10}$$

The total loss is the sum of the knowledge distillation loss and classification loss:

$$L = L_{\text{cls}} + \beta L_{\text{AAT-DKD}},\tag{11}$$

where L_{cls} is the classification loss such as the AAMSoftmax loss [35], calculated with the ground truth labels. β is a parameter that balances the classification loss and the knowledge distillation loss. The parameters of the student network $\boldsymbol{\Phi}_{\text{stu}}$ are updated as follows:

$$\boldsymbol{\Phi}_{\text{stu}} \leftarrow \boldsymbol{\Phi}_{\text{stu}} - \eta\frac{\partial\sum_{\mathbf{x}\in D} L(\mathbf{x})}{\partial\boldsymbol{\Phi}_{\text{stu}}},\tag{12}$$

where D comprises the sample in a mini-batch, and η is a constant learning rate.

3.4. Dynamic reversal coefficient in adversarial learning

To implement AAT-DKD, we learn the parameters θ_{TSKD} and θ_{NSKD} in Eq. (9) adversarially. Adversarial learning is achieved by performing gradient ascent on L_{TSKD} and L_{NSKD} in Eq. (10) with respect to θ_{TSKD} and θ_{NSKD}, respectively. Specifically, we have

$$\begin{aligned}\theta_{\text{TSKD}} &\leftarrow \theta_{\text{TSKD}} + \eta\beta\lambda\sum_{\mathbf{x}\in D}\frac{\partial(L_{\text{TSKD}}(\mathbf{x}))}{\partial\theta_{\text{TSKD}}},\\\theta_{\text{NSKD}} &\leftarrow \theta_{\text{NSKD}} + \eta\beta\lambda\gamma\sum_{\mathbf{x}\in D}\frac{\partial(L_{\text{NSKD}}(\mathbf{x}))}{\partial\theta_{\text{NSKD}}},\end{aligned}\tag{13}$$

where λ is a parameter controlling the strength of adversarial learning. In previous work [23], λ was fixed, meaning that the strength of adversarial learning depends on the total gradient in the mini-batch (because η, β, and γ, are fixed hyperparameters.). Due to the varying quality of samples in a mini-batch, the strength of adversarial learning should also be dynamically adjusted. To implement this, we define λ to be dependent on the overall quality of a mini-batch as follows:

$$\lambda = \frac{1}{B}\sum_{j=1}^{B}\bar{p}_{t,j}^{\mathcal{T}},\quad t\in\{1,\dots,C\},\tag{14}$$

where B is the batch size, t denotes the target class for sample j, and $\bar{p}_{t,j}^{\mathcal{T}} = \frac{e^{q_{t,j}^{\mathcal{T}}}}{\sum_{k=1}^{C} e^{q_{k,j}^{\mathcal{T}}}}$. It is worth noting that for the calculation of $\bar{p}_{t,j}^{\mathcal{T}}$, we use a constant temperature of 1.0. $\bar{p}_{t,j}^{\mathcal{T}}$ to some extent can reflect the quality of the j-th sample [22]. For example, a high target probability indicates that the corresponding sample is of high quality and its contribution to adversarial learning will be large. Conversely, a low target probability suggests a challenging sample. In this case, the contribution of this sample to adversarial learning will be small to prevent excessive adversarial training.

4. Experimental setup

4.1. Datasets

We utilized the VoxCeleb2 development [36] set comprising 5994 speakers as the training set. For evaluation, we employed the test sets from VoxCeleb1 [37], i.e., Vox-O, Vox-E, and Vox-H. Vox-O consists of 37,611 trials from 40 speakers. Vox-E, which utilizes the entire dataset, comprises 579,818 trials from 1251 speakers. Vox-H is a challenging evaluation set, containing 550,894 pairs sampled from 1190 speakers in VoxCeleb1. VoxCeleb1 and VoxCeleb2 are multilingual, but the majority of utterances were spoken in English. We also tested the performance of various systems on the VoxSRC 2021 dataset. Since the labels of the VoxSRC 2021 test set are not publicly available, we used the validation set (VoxSRC21-val), which contains 60,000 evaluation trials from VoxCeleb1. This dataset is commonly used to test the robustness of speaker verification systems.

To further demonstrate the effectiveness of our method, we conducted experiments on the CN-Celeb [38] dataset, a multi-genre Mandarin corpus collected from Chinese open media. We used the development set from CN-Celeb1 and CN-Celeb2 as the training set, which comprise 2793 speakers. The test set is the CN-Celeb evaluation set (CN-eval), which contains 18,000 utterances from 196 speakers.

We followed the data augmentation strategy in Kaldi's recipes [39]. Specifically, we added noise, music, and babble to the training data using MUSAN [40] and created reverberated speech using the RIR dataset [41].

To demonstrate the generalizability of our approach, we performed experiments on the image classification task using the CIFAR-100 dataset. This dataset comprises natural images with a resolution of 32×32 pixels. It includes 50,000 training images and 10,000 test images. For the image classification task, we did not apply any data augmentation techniques, such as cropping, flipping, or rotation.

4.2. Teacher and student models

For the experiments on VoxCeleb, to maintain consistent with [20], we used the same teacher model as in [20], combining WavLM-large [7] with ECAPA-TDNN [3].[1] For the experiments on CN-Celeb, to maintain the same number of parameters in the teacher model, we continued using the WavLM-large with ECAPA-TDNN configuration. We used the WavLM pre-trained model, as in [20], to extract features, and we trained ECAPA-TDNN on the development set from CN-Celeb1 and CN-Celeb2 sets. For the experiments on CIFAR-100, we used an ResNet50 as the teacher model.

We utilized two different architectures for the ASV student model: a standard x-vector [1] and an ECAPA-TDNN with 512 channels (a smaller version of ECAPA-TDNN). We used a cosine backend in all ASV experiments. For image classification, we used a ResNet18 as the student model.

4.3. Network training

For VoxCeleb, during training, each speech signal was randomly cropped into 2 s and augmented with a probability of 0.6. For CN-Celeb, each audio waveform was randomly cropped into 3 s and augmented with a probability of 0.8. After augmentation, 80-dimensional filter bank (Fbank) features were extracted using a frame length of 25 ms and a frameshift of 10 ms. These Fbank features were then fed into the student model, while the augmented raw waveform was fed into the teacher network.

For the classification loss, we utilized the AAMSoftmax [15] with a scale of 32 and a margin scheduler, where the margin was 0 for the first 20 epochs and exponentially increased to 0.2 over the next 20 epochs, after which it remained constant. For the knowledge distillation loss, we followed [20], and set the value of γ to 2.0. We adopted the same strategy as in [20], where β was linearly increased from 0.05 to 1 during the first 20 epochs and remained constant. We employed an SGD optimizer to optimize the student model. A linear learning rate warmup was employed during the first 6 epochs, increasing the learning rate from 5e−4 to 0.15. We set α_1 to 0.25 and α_2 to 5. The batch size was set to 512 for training the x-vector network and 256 for training the ECAPA-TDNN.

For CIFAR-100 experiments, we used a SGD optimizer with a learning rate of 0.1, momentum of 0.9, and weight decay of 5e−4. A cosine annealing learning rate scheduler was applied, with a maximum of 150 epochs and a minimum learning rate of 0.001.

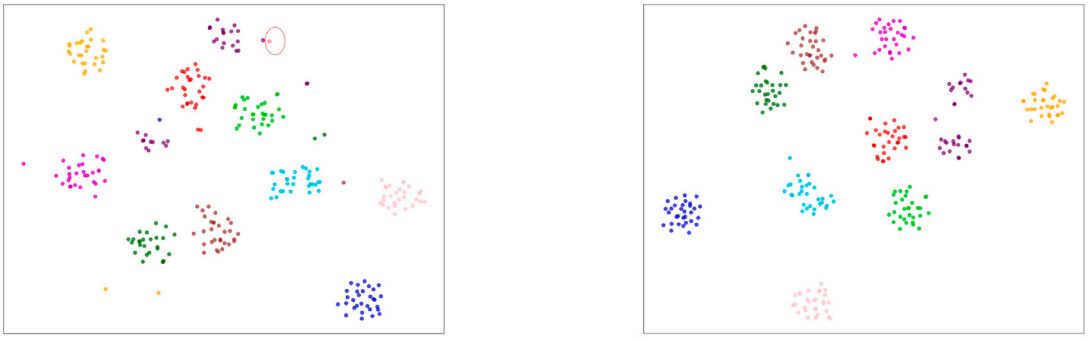
[1] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification

Table 1
Results on the VoxCeleb1 and VoxSRC21-val test sets. "–" denotes using the classification loss only.

System	# Params	Distillation	Vox1-O		Vox1-E		Vox1-H		VoxSRC21-val	
	(M)	method	EER (%)	minDCF	EER (%)	minDCF	EER (%)	minDCF	EER (%)	minDCF
Teacher model WavLM-TDNN [7]	316.62	–	0.431	–	0.538	–	1.154	–	–	–
Student model x-vector	4.39	–	2.06	0.190	1.99	0.213	3.33	0.317	5.60	0.419
	4.39	KD	1.71	**0.165**	1.72	0.187	2.98	0.284	4.93	0.414
	4.39	DKD	1.53	0.168	1.56	0.173	2.79	0.261	4.96	0.415
	4.39	AAT-DKD	**1.47**	0.174	**1.52**	**0.170**	**2.74**	**0.257**	**4.46**	**0.404**
Student model ECAPA-TDNN	7.00	–	1.02	0.106	1.15	0.129	2.22	0.221	4.26	0.384
	7.00	KD	0.90	**0.098**	0.99	0.117	1.96	0.195	4.01	0.351
	7.00	DKD	0.84	0.109	0.96	0.116	1.98	0.200	3.86	**0.334**
	7.00	AAT-DKD	**0.74**	0.108	**0.94**	**0.112**	**1.93**	**0.190**	**3.76**	0.344

Table 2
Results on CN-Celeb evaluation set. "–" denotes using the classification loss only.

System	# Params	Distillation	CN-eval	
	(M)	method	EER (%)	minDCF
Teacher model WavLM-TDNN [7]	316.62	–	7.33	0.421
Student model x-vector	4.39	–	9.40	0.502
	4.39	KD	9.29	0.501
	4.39	DKD	9.15	0.509
	4.39	AAT-DKD	**8.72**	**0.500**
Student model ECAPA-TDNN	7.00	–	8.87	0.462
	7.00	KD	8.21	0.451
	7.00	DKD	7.65	0.439
	7.00	AAT-DKD	**7.56**	**0.439**



(a) DKD (b) AAT-DKD

Fig. 2. t-SNE plots of speaker embeddings obtained from a student (ECAPA-TDNN) using (a) DKD and (b) AT-DKD.

4.4. Performance metrics

For ASV, the performance metrics include equal error rate (EER) and minimum detection cost function (minDCF) with $P_{\text{target}} = 0.01$. All ASV experiments were conducted based on the 3D-Speaker toolkit [42].[2]

5. Results and analyses

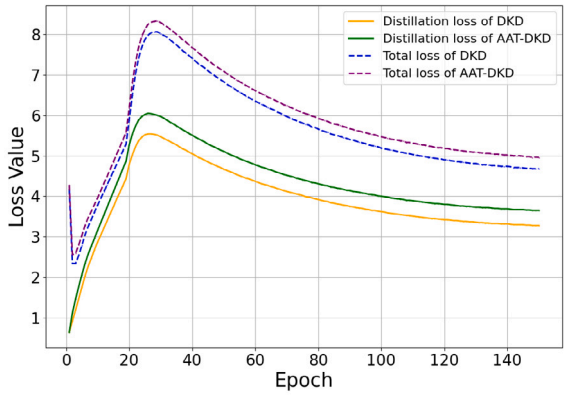
5.1. Main results

Table 1 presents our main results on the speaker embeddings extracted from the x-vector and the ECAPA-TDNN models. It is evident that KD outperforms methods relying solely on the ground-truth labels for classification. When employing the DKD distillation method, performance further improves, demonstrating the effectiveness of emphasizing non-target speaker information in knowledge distillation. For the x-vector, the proposed AAT-DKD outperforms both KD and DKD. AAT-DKD also reduces the EER on ECAPA-TDNN by 17.78% and 11.9% compared to KD and DKD, respectively. Our approach achieves an EER of 0.744% on VoxCeleb-O which is a notable achievement for the ECAPA-TDNN. These findings suggest that using adversarially adaptive temperatures can enhance the model's performance.

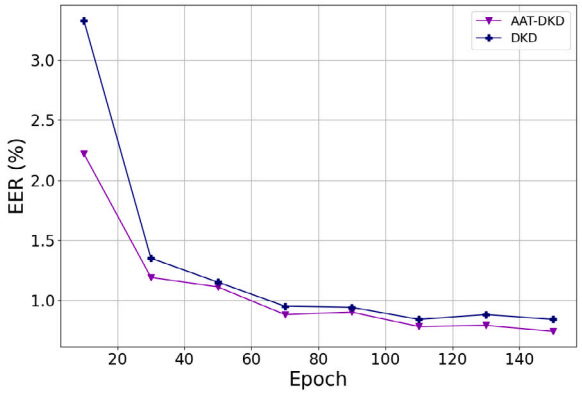
In addition to the VoxCeleb1 test set, we also tested on the challenging VoxSRC21-val set. According to Table 1, AAT-DKD performs the best for the x-vector and ECAPA-TDNN models. This indicates that knowledge distillation with adversarially adaptive temperatures can make the student model more robust.

To further demonstrate the effectiveness of AAT-DKD, we conducted experiments on a Mandarin corpus. Table 2 presents our results on the CN-eval dataset. We found that AAT-DKD outperforms KD and DKD on x-vector and ECAPA-TDNN. Specifically, it achieved an EER of 7.56% on ECAPA-TDNN, which is a highly competitive result. This result shows that the AAT-DKD method is effective not only in English but also in Chinese, demonstrating the robustness of the AAT-DKD approach.

[2] https://github.com/modelscope/3D-Speaker



(a)



(b)

Fig. 3. The curves of (a) distillation loss and total loss, and (b) EER during training.

To observe the representation power of AAT-DKD, we used the ECAPA-TDNNs trained with DKD and AAT-DKD to extract speaker embeddings and plot them using t-SNE in Fig. 2. We selected 10 speakers from the VoxCeleb1-dev set, each having 30 utterances. According to Fig. 2, AAT-DKD generally presents higher cluster compactness and fewer outliers compared with DKD. Especially, one instance of incorrect classification, highlighted with a red circle, was observed in DKD, whereas our method does not exhibit such an error. This verifies that the proposed method can achieve more discriminative speaker embeddings.

We also investigated the generalization of AAT-DKD. Fig. 3(a) shows the training loss curves of DKD and AAT-DKD on Vox1-O. During training, θ_{TSKD} and θ_{NSKD} were optimized to maximize the distillation loss, while Φ_{stu} was optimized to minimize the distillation loss. Since β follows a warm-up procedure in the first 20 epochs, the distillation loss increases in the early stages. After the warm-up period, as Φ_{stu} plays a leading role in this min–max game, the distillation loss shows an overall decreasing trend. Fig. 3(b) shows the EER curves of AAT-DKD and DKD during the training process. From Fig. 3, we observe that AAT-DKD has a higher loss than DKD, but its EER is lower than that of DKD. This observation suggests that AAT-DKD has better generalization capacity.

5.2. Impact of adversarial training

To better demonstrate the importance of the adversarial learning strategy in AAT-DKD, we trained the models using two strategies: direct learning of the temperature parameters ($\theta_{\text{TSKD}}, \theta_{\text{NSKD}}$) and adversarial learning of the temperature parameters. The results are shown in Table 3. Comparing the first (third) row and the second (fourth) row, the temperature obtained with adversarial learning demonstrates better performance. Moreover, the results of directly learning θ_{TSKD} and θ_{NSKD} without the adversarial strategy are worse than using DKD alone (compared with DKD in Table 1). This result underscores the importance of the adversarial learning strategy.

Table 3
Impact of learning strategies for θ_{TSKD} and θ_{NSKD} on Vox1-O. The "Normal" and "Adversarial" learning strategies were implemented by applying gradient descent and gradient ascent on the AAT-DKD loss (Eq. (10)), respectively. "None" means that a constant temperature was used.

Row	Learning strategy	Initial value		EER (%)
		τ_{TSKD}	τ_{NSKD}	
1	Normal	3.91	3.91	0.85
2	Adversarial	3.91	3.91	0.74
3	Normal	0.25	0.25	0.85
4	Adversarial	0.25	0.25	**0.71**
5	None	1.64	1.41	0.78

Table 4
Effect of using the same or different temperatures for the target (L_{TSKD}) and non-target (L_{NSKD}) distillation losses in AAT-DKD.

Distillation method	Temperature strategy	EER (%)
AT-DKD	Same τ	0.82
	Different τ	**0.74**

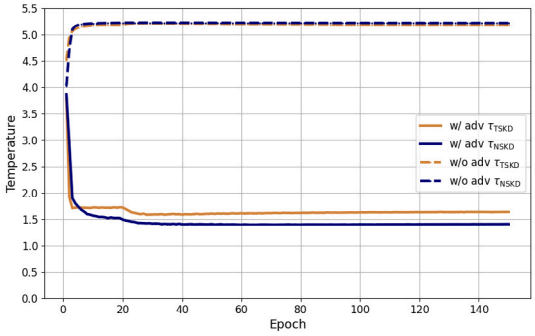
Additionally, we adjusted θ_{TSKD} and θ_{NSKD} to set the initial values of τ_{TSKD} and τ_{NSKD} to be greater than 1.0 and less than 1.0, respectively. This experiment aims to explore the impact of different initial values of τ_{TSKD} and τ_{NSKD} on system performance. We plotted the curves of τ_{TSKD} and τ_{NSKD} under different adaptive strategies and initial values, as shown in Fig. 4. From Fig. 4, when the initial value is larger than 1.0, τ_{TSKD} and τ_{NSKD} with the normal adaptive strategy quickly reaches the set maximum value and remains constant. When the initial value is less than 1.0, it stays at the lower boundary of the range and remains nearly unchanged. Neither τ_{TSKD} and τ_{NSKD} converges, indicating that this strategy does not work. When using adversarial strategy, regardless of the initial value, τ_{TSKD} eventually converges to around 1.6, and τ_{NSKD} converges to around 1.4. This indicates that the final convergence of τ_{TSKD} and τ_{NSKD} is independent of their initial values. Additionally, we set the temperature hyperparameters to constants and assigned them to the τ_{TSKD} and τ_{NSKD} in AAT-DKD. Specifically, τ_{TSKD} was set to 1.64 and τ_{NSKD} was set to 1.41. Comparing the result in Row 5 with those in Rows 2 and 4 in Table 3, we observe that AAT-DKD exhibits better performance. This confirms our hypothesis that different stages of knowledge distillation require different optimal temperatures. Therefore, using adaptive temperatures is more effective.

5.3. Impact of group-dependent temperatures

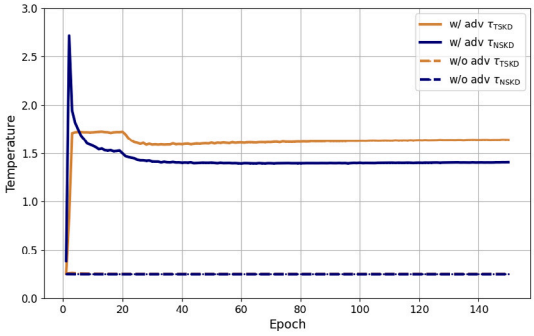
We explored the effectiveness of using different temperatures for different types of losses, i.e., having different temperatures for the target-speaker KD loss (L_{TSKD}) and the nontarget-speaker KD loss (L_{NSKD}). The studies, which use the ECAPA-TDNN models, were evaluated on Vox1-O. From Table 4, we observe that using different adaptive temperatures for L_{TSKD} and L_{NSKD} yields better performance compared to using the same temperature. This is reasonable because L_{TSKD} and L_{NSKD} distill different information from different distributions; hence, the temperatures controlling the smoothness of the distributions should also be different.

5.4. Impact of dynamic adversarial coefficient λ

We explored the impact of the dynamic adversarial coefficient λ in Eq. (14) on the system and compared AAT-DKD with that setting



(a) Initial values of τ_{TSKD} and τ_{NSKD} greater than 1



(b) Initial values of τ_{TSKD} and τ_{NSKD} smaller than 1

Fig. 4. Prediction curves of τ_{TSKD} and τ_{NSKD} with and without adversarial strategy.

Table 5
Effect of a dynamic adversarial coefficient λ in Eq. (14).

Distillation method	λ	EER (%)
AT-DKD	1.0 (fixed)	0.80
	Dynamic	**0.74**

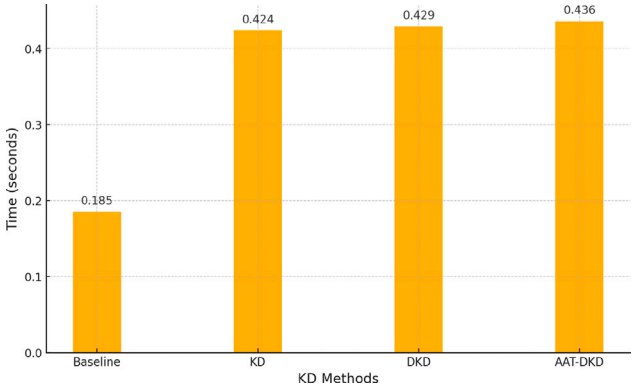


Fig. 5. Comparison of training time per batch across different KD methods.

λ to a fixed value of 1.0 for all mini-batches. Table 5 shows that a dynamic λ yields better performance, which further validates our argument that adopting varying degrees of adversarial learning for individual mini-batches is a reasonable approach.

5.5. Training efficiency

We compared the training time of AAT-DKD with the other methods under the same training conditions. The comparison results are shown in Fig. 5. Using two RTX 4090 GPUs and a batch size of 256, we conducted the experiments on the ECAPA-TDNN model and recorded the time required to process one batch (in seconds). As shown in Fig. 5, the additional computational overhead introduced by AAT-DKD is almost negligible compared to other KD methods.

5.6. Exploring effectiveness in the image classification

We also explored the effectiveness of AAT-DKD in the image classification domain, with the results shown in Table 6.[3] Table 6 shows that AAT-DKD demonstrates better performance, which proves its effectiveness and generalizability.

[3] We did not use state-of-the-art image classifiers or toolkits. Therefore, the performance of the baseline is not as competitive as those published in 2024.

Table 6
Performance (top-1 accuracy %) of the student model trained with various KD methods on the CIFAR-100 test set. "–" denotes using the classification loss only.

Distillation method	Acc (%)
[43]	60.58
–	62.55
KD	64.18
DKD	64.66
AAT-DKD	**65.10**

Table 7
Performance of different KD methods on short-duration scenarios (Vox1-O, 2s).

Distillation method	EER (%)
KD	2.96
DKD	3.05
AAT-DKD	**2.89**

Table 8
Performance of different KD methods on language mismatch scenarios (CN-Celeb evaluation set).

Distillation method	EER (%)
KD	**14.39**
DKD	14.40
AAT-DKD	14.43

5.7. Limitations and future works

The limitations of AAT-DKD are discussed here. We selected the ECAPA-TDNN models trained with the KD, DKD, and AAT-DKD methods, as shown in Table 1, and evaluated their performance in specific scenarios.

Challenges in Short-Utterance Scenarios: We evaluated our method in short-duration scenarios by randomly cropping the test audio from Vox1-O to 2 s to simulate short-duration conditions. As presented in Table 7, We observed that the performance of our method did not show significant improvement in short-duration audio in Vox1-O. We acknowledge that the proposed method has limitations under this condition, indicating that future work should focus on incorporating duration adaptation into the proposed knowledge distillation framework to enhance its robustness under the short-utterance scenario.

Challenges in Language Mismatch Scenarios: As shown in Table 8, We noticed that AAT-DKD did not achieve significant improvements when there was a language mismatch between the test and training audio. This can be attributed to the lack of language adaptation modules in our current framework. We recognize that this is a limitation and plan to address it in future work by incorporating a language adaptation module into our knowledge distillation framework to improve the method's robustness in such challenging scenarios.

6. Conclusions

We found that setting the temperature in the Decoupled Knowledge Distillation (DKD) loss to a fixed value is not an optimal solution. We proposed an adversarially adaptive temperatures for DKD (AAT-DKD) to address this issue. Furthermore, we used τ_{TSKD} and τ_{NSKD} for L_{TSKD} and L_{NSKD}, respectively. Additionally, the gradient reversal coefficient in the GRL was set to be different for individual mini-batches, adjusting the adversarial learning intensity based on the overall quality of the mini-batches. The experimental results on VoxCeleb1, VoxSRC21-val, CN-Celeb-eval, and CIFAR-100 demonstrate that our method is effective.

CRediT authorship contribution statement

Zezhong Jin: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Youzhi Tu:** Writing – review & editing, Supervision. **Chong-Xin Gan:** Investigation. **Man-Wai Mak:** Writing – review & editing, Supervision. **Kong-Aik Lee:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Research Grant Council of Hong Kong SAR, Grant Nos. 15210122 and 15228223.

Data availability

The data used in the paper is open-source, and links to download the data as well as the code for the experimental framework are provided.

References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust DNN embeddings for speaker recognition, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 5329–5333.

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[3] B. Desplanques, J. Thienpondt, K. Demuynck, Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification, in: Proc. Interspeech, 2020, pp. 3830–3834.

[4] H. Wang, S. Zheng, Y. Chen, L. Cheng, Q. Chen, Cam++: A fast and efficient network for speaker verification using context-aware masking, in: Proc. Interspeech, 2023, pp. 5301–5305.

[5] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, Wav2vec 2.0: A framework for self-supervised learning of speech representations, Adv. Neural Inf. Process. Syst. 33 (2020) 12449–12460.

[6] W.-N. Hsu, B. Bolte, Y.-H.H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, HuBERT: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ ACM Trans. Audio Speech Lang. Process. 29 (2021) 3451–3460.

[7] 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 J. Sel. Top. Signal Process. 16 (6) (2022) 1505–1518.

[8] Z. Fan, M. Li, S. Zhou, B. Xu, Exploring wav2vec 2.0 on speaker verification and language identification, 2020, arXiv preprint arXiv:2012.06185.

[9] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, M. Zeng, Large-scale self-supervised speech representation learning for automatic speaker verification, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2022, pp. 6147–6151.

[10] J. Peng, O. Plchot, T. Stafylakis, L. Mošner, L. Burget, J. Černockỳ, An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification, in: IEEE Spoken Language Technology Workshop, SLT, 2023, pp. 555–562.

[11] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, in: NIPS Deep Learning and Representation Learning Workshop, 2015.

[12] S. Wang, Y. Yang, Y. Qian, K. Yu, Revisiting the statistics pooling layer in deep speaker embedding learning, in: Proc. IEEE 12th International Symposium on Chinese Spoken Language Processing, 2021, pp. 1–5.

[13] D. Snyder, D. Garcia-Romero, D. Povey, S. Khudanpur, Deep neural network embeddings for text-independent speaker verification, in: Proc. Interspeech, 2017, pp. 999–1003.

[14] K. Okabe, T. Koshinaka, K. Shinoda, Attentive statistics pooling for deep speaker embedding, in: Proc. Interspeech, 2018, pp. 2252–2256.

[15] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.

[16] Z. Bai, J. Wang, X.-L. Zhang, J. Chen, End-to-end speaker verification via curriculum bipartite ranking weighted binary cross-entropy, in: IEEE/ ACM Trans. Audio Speech Lang. Process., 30 (2022) 1330–1344.

[17] J.S. Chung, J. Huh, S. Mun, M. Lee, H.S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, I. Han, In defence of metric learning for speaker recognition, in: Proc. Interspeech, 2020, pp. 2977–2981.

[18] S. Wang, Y. Yang, T. Wang, Y. Qian, K. Yu, Knowledge distillation for small foot-print deep speaker embedding, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 6021–6025.

[19] Z. Peng, X. He, K. Ding, T. Lee, G. 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] D.-T. Truong, R. Tao, J.Q. Yip, K.A. Lee, E.S. 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.

[21] L. Xu, J. Ren, Z. Huang, W. Zheng, Y. Chen, Improving knowledge distillation via head and tail categories, IEEE Trans. Circuits Syst. Video Technol. 34 (5) (2023) 3465–3480.

[22] B. Zhao, Q. Cui, R. Song, Y. Qiu, J. Liang, Decoupled knowledge distillation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 11953–11962.

[23] Z. Li, X. Li, L. Yang, B. Zhao, R. Song, L. Luo, J. Li, J. Yang, Curriculum temperature for knowledge distillation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 1504–1512.

[24] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: International Conference on Machine Learning, PMLR, 2015, pp. 1180–1189.

[25] B. Liu, H. Wang, Z. Chen, S. Wang, Y. Qian, Self-knowledge distillation via feature enhancement for speaker verification, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2022, pp. 7542–7546.

[26] Y. Jin, G. Hu, H. Chen, D. Miao, L. Hu, C. Zhao, Cross-modal distillation for speaker recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 12977–12985.

[27] V. Mingote, A. Miguel, D. Ribas, A. Ortega, E. Lleida, Knowledge distillation and random erasing data augmentation for text-dependent speaker verification, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2020, pp. 6824–6828.

[28] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 9650–9660.

[29] Z. Jin, Y. Tu, M.-W. Mak, W-GVKT: Within-global-view knowledge transfer for speaker verification, in: Proc. Interspeech, 2024, pp. 3779–3783.

[30] Z. Jin, Y. Tu, M.-W. Mak, Self-supervised learning with multi-head multi-mode knowledge distillation for speaker verification, in: Proc. Interspeech, 2024, pp. 4723–4727.

[31] Z. Zhao, Z. Li, X. Zhang, W. Wang, P. Zhang, Prototype division for self-supervised speaker verification, IEEE Signal Process. Lett. 31 (2024) 880–884.

[32] K. Chandrasegaran, N.-T. Tran, Y. Zhao, N.-M. Cheung, Revisiting label smoothing and knowledge distillation compatibility: What was missing? in: International Conference on Machine Learning, PMLR, 2022, pp. 2890–2916.

[33] J. Liu, B. Liu, H. Li, Y. Liu, Meta knowledge distillation, 2022, arXiv preprint arXiv:2202.07940.

[34] V. Mingote, A. Miguel, A. Ortega, E. Lleida, Class token and knowledge distillation for multi-head self-attention speaker verification systems, Digit. Signal Process. 133 (2023) 103859.

[35] X. Xiang, S. Wang, H. Huang, Y. Qian, K. Yu, Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition, in: Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 1652–1656.

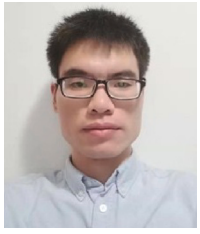
[36] J. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep speaker recognition, in: Proc. Interspeech, 2018, pp. 1086–1090.

[37] A. Nagrani, J. Chung, A. Zisserman, VoxCeleb: a large-scale speaker identification dataset, in: Proc. Interspeech, 2017, pp. 2616–2620.

[38] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T.F. Zheng, D. Wang, CN-celeb: multi-genre speaker recognition, Speech Commun. 137 (2022) 77–91.

[39] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The Kaldi speech recognition toolkit, in: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, 2011.
[40] D. Snyder, G. Chen, D. Povey, MUSAN: A music, speech, and noise corpus, 2015, arXiv preprint arXiv:1510.08484.
[41] M. Jeub, M. Schafer, P. Vary, A binaural room impulse response database for the evaluation of dereverberation algorithms, in: Proc. 16th International Conference on Digital Signal Processing, 2009.
[42] Y. Chen, S. Zheng, H. Wang, L. Cheng, T. Zhu, C. Song, R. Huang, Z. Ma, Q. Chen, S. Zhang, et al., 3D-speaker-toolkit: An open source toolkit for multi-modal speaker verification and diarization, 2024, arXiv preprint arXiv:2403.19971.
[43] T. Chen, Z. Zhang, S. Liu, S. Chang, Z. Wang, Robust overfitting may be mitigated by properly learned smoothening, in: International Conference on Learning Representations, 2020.

Zezhong Jin received his B.Eng. degree in Electronic information engineering from Hebei University in 2021 and M.Sc. degree in Electronic and Information Engineering from The Hong Kong Polytechnic University in 2023. He is currently pursuing a Ph.D. degree in the Department of Electrical and Electronic Engineering at The Hong Kong Polytechnic University. His research interests include speaker verification, self-supervised learning, and knowledge distillation.



Youzhi Tu received a B.Eng. degree and an M.Sc. degree from Harbin Engineering University in 2012 and 2015, respectively. He received a Ph.D. degree in electronic and information engineering at The Hong Kong Polytechnic University in 2022. He is now a postdoctoral fellow at The Hong Kong Polytechnic University. His research interests include speaker recognition and machine learning.

Chong-Xin Gan received his B.Eng. degree in Computer Science and Technology from Zhengzhou University in 2019 and M.Sc. degree in Electronic and Information Engineering from The Hong Kong Polytechnic University in 2021. He is currently pursuing a Ph.D. degree in the Department of Electrical and Electronic Engineering at The Hong Kong Polytechnic University. His research interests include speaker verification, sound event classification, and deep learning.

Man-Wai Mak (M'93–SM'15) received a Ph.D. in electronic engineering from the University of Northumbria in 1993. He joined the Department of Electronic and Information Engineering at The Hong Kong Polytechnic University in 1993 and is currently a professor in the same department. He has authored more than 200 technical articles in speaker recognition, machine learning, and bioinformatics. Dr. Mak also coauthored postgraduate textbooks Biometric Authentication: A Machine Learning Approach, Prentice-Hall, 2005 and Machine Learning for Speaker Recognition, Cambridge University Press, 2020. He served as a member of the IEEE Machine Learning for Signal Processing Technical Committee in 2005–2007. He has served as an associate editor of IEEE/ACM Transactions on Audio, Speech and Language Processing. He is currently an associate editor of Journal of Signal Processing Systems and IEEE Biometrics Compendium. He also served as Technical Committee Members of a number of international conferences, including ICASSP and Interspeech, and gave a tutorial on machine learning for speaker recognition in Interspeech'2016. Dr. Mak's research interests include speaker recognition, machine learning, and bioinformatics.

Kong Aik Lee (Senior Member, IEEE) received the Ph.D. degree from Nanyang Technological University, Singapore, in 2006. From 2006 to 2018, he was a Research Scientist and then a Strategic Planning Manager (concurrent appointment) with the Institute for Infocomm Research, Singapore. From 2018 to 2020, he was a Senior Principal Researcher with the Data Science Research Laboratories, NEC Corporation, Tokyo, Japan. He was an Associate Professor with the Singapore Institute of Technology, Singapore, while holding a concurrent appointment as a Principal Scientist and a Group Leader with the Agency for Science, Technology and Research (ASTAR), Singapore. He is currently an Associate Professor with the Hong Kong Polytechnic University, Hong Kong. His research interests include the automatic and paralinguistic analysis of speaker characteristics, ranging from speaker recognition, language and accent recognition, voice biometrics, spoofing, and countermeasures. From 2017 to 2021, he was an Associate Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. Since 2016, he has been an Editorial Board Member of Elsevier Computer Speech and Language. He is an elected Member of the IEEE Speech and Language Processing Technical Committee and was the General Chair of the Speaker Odyssey 2020 Workshop. He was the recipient of the Singapore IES Prestigious Engineering Achievement Award 2013 and the Outstanding Service Award by IEEE ICME 2020.