Grouped Knowledge Distillation with Adaptive Logit Softening for Speaker Recognition

Chong-Xin Gan, Youzhi Tu, Zezhong Jin, Man-Wai Mak, and Kong Aik Lee
Department of Electrical and Electronic Engineering,
The Hong Kong Polytechnic University,
Hong Kong SAR

***Abstract*—Recent works suggest that decoupling the information of non-target speakers from that of the target speaker in knowledge distillation (KD) and subsequently emphasizing the former can lead to significant performance improvement. However, a well-trained teacher model typically produces almost zero non-target speaker posteriors with limited contribution to knowledge transfer, resulting in a less effective KD. To address this problem, we advocate a dual-group knowledge distillation framework, wherein the primary group with top-*k* speaker posteriors captures most of the speaker discrimination knowledge in an utterance. The non-primary group contributes to the KD through a binary classification (distillation) between the primary and non-primary groups. In addition, adaptive logit softening is proposed to adjust the teacher's and student's logits in the binary distillation, further facilitating effective knowledge transfer. The proposed method trained with a simple x-vector pipeline obtains an impressive equal error rate of 1.46%, 1.47%, and 2.70% on three VoxCeleb1 test sets, outperforming the state-of-the-art methods with a noticeable margin.**

***Index Terms*—Speaker recognition; knowledge distillation; grouped knowledge transfer; adaptive logit softening**

I. Introduction

Speaker recognition [1]–[3] systems have achieved remarkable successes in recent years, largely attributed to the development of deep learning [4] and large-scale annotated datasets [5], [6]. Numerous studies used advanced architectures to capture speaker-dependent information from hand-crafted features, evolving from the early i-vector approach [7] to current deep neural network-based methods [8]–[10]. The networks are optimized via sophisticated margin-based [11]–[14] or metric-based [15]–[19] losses, thereby producing robust and discriminative speaker embeddings. During inference, the representations extracted from the speaker embedding networks are fed into various back-ends, including cosine similarity and probabilistic linear discriminant analysis [20] to facilitate decision-making.

Many powerful automatic speaker verification (ASV) systems have been proposed by combining the large-scale pre-trained speech models [21]–[23] followed by traditional speaker encoders, such as the ECAPA-TDNN [24] and the x-vector network [25]. However, the performance gains are at the expense of increased trainable parameters. Such cumbersome models are difficult to deploy on edge devices with limited computational resources. Knowledge distillation (KD) [26] provides an effective way to solve this problem. With KD, we can train a small student model with supervision from a large teacher model. This is achieved by minimizing the Kullback–Leibler (KL) divergence between the softened outputs of the teacher and the student networks at the logit level [27], [28] or reducing the discrepancies between their intermediate outputs at the feature level [28]–[30].

Recently, it was found that splitting the predictions of teacher and student networks into target and non-target groups is beneficial for knowledge transfer in computer vision [31]. This idea, known as decoupled knowledge distillation, has also achieved excellent performance in ASV [32], [33]. Furthermore, the authors of [34] introduced a strategy called grouped knowledge distillation (GKD), as they found that the majority of the non-target class posteriors are almost zero, thereby contributing only minor knowledge to the student network and making KD difficult. Inspired by GKD, we propose splitting the speaker posteriors into a primary and a non-primary group for each training utterance. Specifically, the teacher and student networks convert each input utterance into N speaker posteriors corresponding to N training speakers. These N speaker posteriors are split into two groups: a primary group and a non-primary group. The primary group comprises top-k speaker posteriors, while the non-primary group consists of the remaining $(N-k)$ speaker posteriors. From these posteriors, a binary group is constituted, which comprises accumulated speaker posteriors from the primary group and the accumulated speaker posteriors from the non-primary group. Based on this grouping, the traditional knowledge distillation loss can be reformulated into two terms: a primary loss and a binary loss. The primary loss is to transfer the knowledge in the primary groups from the teacher to the student. The binary loss is also proposed, enabling the student network to imitate the inter-group relation between the primary and non-primary groups that the teacher network possesses.

Despite GKD's effectiveness, utilizing a global temperature parameter to adjust the speaker posteriors across all training samples may lead to undesirable results because different samples have different posterior distributions and contribute differently to KD. On the other hand, it is suboptimal to use a constant temperature for both the primary and binary losses in GKD because the distributions of speaker posteriors vary

This work was supported by the RGC of Hong Kong SAR, Grant No. PolyU 15210122 and 15228223

significantly between the primary and non-primary groups. Moreover, given a well-trained teacher model, the posterior distribution of speakers in the primary group will be much sharper than that of the non-primary group, causing the binary cross-entropy loss to be dominated by a few posteriors only. In this work, instead of only using a static temperature, we advocate softening the logits dynamically so that a constant temperature can be applied. The logit softener aims to smooth out the sharp posterior distribution of speakers, and the degree of softening depends on individual utterances.

In summary, our contributions are as follows. Firstly, we introduce the strategy of GKD to ASV by splitting the conventional KD loss into a primary loss and a binary loss for efficient KD. Secondly, we propose applying adaptive logit softening in the binary distillation, which is part of GKD, to further smooth the speaker posteriors for better inter-group knowledge transfer.

II. Methodology

In this section, we briefly review the conventional and decoupled knowledge distillation. Our proposed method is then detailed.

A. Conventional and Decoupled Knowledge Distillation

Consider an utterance $\boldsymbol{u}$ and its corresponding speaker label y, where $y \in \{1, \ldots, C\}$ and C is the number of speakers in the training set. KD minimizes the Kullback–Leibler (KL) divergence between the classification probabilities of the student and teacher networks. Specifically, a short segment $\boldsymbol{x}$ is randomly truncated from $\boldsymbol{u}$ and then passed to both networks $f^{\text{tea}}(.)$ and $f^{\text{stu}}(.)$. The obtained logits $\boldsymbol{z}^{\text{tea}}$ and $\boldsymbol{z}^{\text{stu}}$ are transformed into respective speaker posterior probabilities $\boldsymbol{p}^{\text{tea}}$ and $\boldsymbol{p}^{\text{stu}}$ via a softmax function $p_i = \frac{\exp(z_i/\tau)}{\sum_{j=1}^{C}\exp(z_j/\tau)}$, where $i \in \{1, \ldots, C\}$ indexes the classes and τ is a temperature constant. Here z_i and p_i are the i-th element of vectors $\boldsymbol{z}$ and $\boldsymbol{p}$, respectively.

The KD loss can be formulated as

$$\mathcal{L}_{\text{KD}} = \text{KL}(\boldsymbol{p}^{\text{tea}} || \boldsymbol{p}^{\text{stu}}) = \sum_{i=1}^{C} p_i^{\text{tea}} \log\left(\frac{p_i^{\text{tea}}}{p_i^{\text{stu}}}\right). \quad (1)$$

In [33], it was shown that the ASV performance can be consistently improved if the number of non-target speakers increases during training. To highlight the role of non-target speakers, $\mathcal{L}_{\text{KD}}$ can be split into target and non-target terms:

$$\mathcal{L}_{\text{KD}} = \text{KL}(\boldsymbol{b}^{\text{tea}} || \boldsymbol{b}^{\text{stu}}) + (1 - p_t^{\text{tea}})\text{KL}(\hat{\boldsymbol{p}}^{\text{tea}} || \hat{\boldsymbol{p}}^{\text{stu}}), \quad (2)$$

where $\boldsymbol{b}^{\text{tea}} = \left[p_t^{\text{tea}}, p_{\setminus t}^{\text{tea}}\right]$, $\boldsymbol{b}^{\text{stu}} = \left[p_t^{\text{stu}}, p_{\setminus t}^{\text{stu}}\right]$, $\hat{\boldsymbol{p}}^{\text{tea}} = [\hat{p}_i^{\text{tea}}]_{i=1, i\neq t}^{C}$ and $\hat{\boldsymbol{p}}^{\text{stu}} = [\hat{p}_i^{\text{stu}}]_{i=1, i\neq t}^{C}$. For simplicity, we only state the teacher's probabilities:

$$p_t^{\text{tea}} = \frac{\exp(z_t^{\text{tea}}/\tau)}{\sum_{j=1}^{C}\exp(z_j^{\text{tea}}/\tau)}, \quad p_{\setminus t}^{\text{tea}} = \frac{\sum_{k=1, k\neq t}^{C}\exp(z_k^{\text{tea}}/\tau)}{\sum_{j=1}^{C}\exp(z_j^{\text{tea}}/\tau)}, \quad (3)$$

$$\hat{p}_i^{\text{tea}} = \frac{\exp(z_i^{\text{tea}}/\tau)}{\sum_{j=1, j\neq i}^{C}\exp(z_j^{\text{tea}}/\tau)}, \quad i \in \{1, \ldots, C\} \setminus t. \quad (4)$$

Because the target speaker's posterior p_t^{tea} dominates the speaker posterior distribution, the non-target speakers' information is potentially suppressed. Besides, p_t^{tea} induces dependence between the target and non-target terms, reducing the benefit of separating the KD into two loss terms. In [32], [33], a hyperparameter γ was introduced to replace $(1 - p_t^{\text{tea}})$ in Eq. 2, emphasizing the non-target information and decoupling it from the target speaker's information. This strategy leads to the decoupled knowledge distillation (DKD) loss, which is expressed as:

$$\mathcal{L}_{\text{DKD}} = \text{KL}(\boldsymbol{b}^{\text{tea}} || \boldsymbol{b}^{\text{stu}}) + \gamma\text{KL}(\hat{\boldsymbol{p}}^{\text{tea}} || \hat{\boldsymbol{p}}^{\text{stu}}). \quad (5)$$

B. Adaptive Logit Softening

Due to the dominance of the target speaker's probability exhibited in the well-trained teacher model, simply softening the logits with a global τ will fail to address the discrepancy in the speaker posterior distributions across different utterances. Thus, it is necessary to reduce the sharpness of the teachers' output probabilities on a sample-by-sample basis. To this end, we propose adaptively softening the logits from both the student and teacher networks by normalizing them with their respective standard deviation before passing them to a softmax function with constant temperature.

As shown in Fig. 1, the teacher and student logits ($\boldsymbol{z}^{\text{tea}}$, $\boldsymbol{z}^{\text{stu}}$) are normalized as follows:

$$\tilde{\boldsymbol{z}}^{\text{tea}} = \frac{\boldsymbol{z}^{\text{tea}}}{\sigma^{\text{tea}}} \quad \text{and} \quad \tilde{\boldsymbol{z}}^{\text{stu}} = \frac{\boldsymbol{z}^{\text{stu}}}{\sigma^{\text{stu}}}, \quad (6)$$

where $\sigma^{\text{tea}} = \sqrt{\frac{\sum_{i=1}^{C}(\boldsymbol{z}_i^{\text{tea}} - \bar{\boldsymbol{z}}^{\text{tea}})^2}{C}}$, $\sigma^{\text{stu}} = \sqrt{\frac{\sum_{i=1}^{C}(\boldsymbol{z}_i^{\text{stu}} - \bar{\boldsymbol{z}}^{\text{stu}})^2}{C}}$ represent the utterance-wise standard deviation of logits $\boldsymbol{z}^{\text{tea}}$ and $\boldsymbol{z}^{\text{stu}}$, respectively. $\bar{\boldsymbol{z}}^{\text{tea}}$ and $\bar{\boldsymbol{z}}^{\text{tea}}$ denote the means of teacher and student logits.

Unlike applying a fixed temperature for all inputs, adaptive logit softening enables KD to tune the probability distribution for each sample, resulting in higher flexibility.

C. Grouped Knowledge Distillation

Similar to DKD, we also partition the traditional KD loss into primary and binary losses:

$$\mathcal{L}_{\text{GKD}} \triangleq \alpha\mathcal{L}_{\text{primary}} + \beta\mathcal{L}_{\text{binary}}, \quad (7)$$

where α and β denote the contribution of the primary and binary losses, respectively. As presented in Fig. 1, the primary and binary losses are computed from the primary group and binary group, correspondingly. For the former, the top-k posteriors of the student network are contrasted with that of the teacher network. For the latter, the loss is derived by computing KL divergence between the corresponding groups: one corresponds to the sum of the top-k posteriors, and another one corresponds to the accumulation of the remaining posteriors.

Following the strategy in [34], we select the top-k speaker posteriors to form the primary speaker group for enhancing the discriminative capabilities of the student network. The

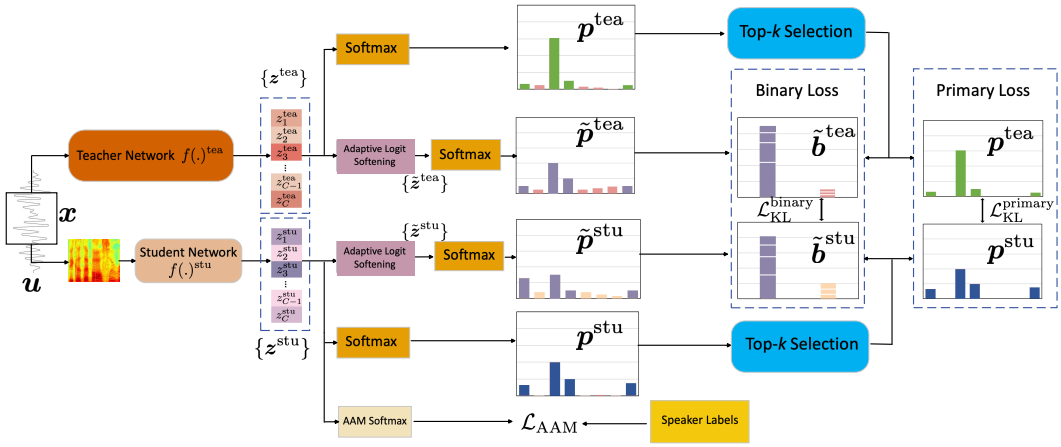

Fig. 1. Framework of the proposed distillation method. In the upper branch, a pre-trained teacher network maps an augmented waveform to logits $\boldsymbol{z}^{\text{tea}}$. Concurrently, mel-filter bank features are extracted and fed into a student network in the bottom branch to obtain student logits $\boldsymbol{z}^{\text{stu}}$, which are then transformed into speaker posteriors. In the binary group, adaptive logit softening is applied to the teacher and student logits for effective knowledge transfer. In the primary group, the probabilities of top-k speakers are selected, enabling the student network to learn discriminative information. The student network is optimized using both the classification loss and the proposed knowledge distillation loss.

student network's predictions are sorted in descending order to facilitate the selection of these top-k confusing speakers. The primary KD loss is formulated as

$$\mathcal{L}_{\text{primary}} = \text{KL}(\boldsymbol{p}^{\text{tea}}||\boldsymbol{p}^{\text{stu}}) = \sum_{i\in\Phi} p_i^{\text{tea}} \log\left(\frac{p_i^{\text{tea}}}{p_i^{\text{stu}}}\right), \quad (8)$$

where Φ is a set of indices corresponding to the top-k speaker posteriors of the student, and $p_i^{\text{tea}} = \frac{\exp(z_i^{\text{tea}}/\tau)}{\sum_{j=1}^{C}\exp(z_j^{\text{tea}}/\tau)}$ and $p_i^{\text{stu}} = \frac{\exp(z_i^{\text{stu}}/\tau)}{\sum_{j=1}^{C}\exp(z_j^{\text{stu}}/\tau)}$ are the speaker posteriors of the two networks with fixed logit softening.

Because the primary posterior group (defined by Φ) retains the majority of the knowledge necessary for effective distillation, it focuses on capturing the relationship between the target speaker and other confusing speakers. Therefore, $\mathcal{L}_{\text{primary}}$ improves the discriminative capabilities of the student by ignoring the non-primary speakers whose posterior probabilities are small. The student network is expected to learn inter-speaker knowledge in the primary group, enabling its predictions to closely align with those of the teacher.

The non-primary speaker/posterior group aggregates the posteriors of non-primary speakers, contrasting with the aggregation of the primary posteriors. As a result, we consider the primary and non-primary posteriors to be coming from two distinct classes and formulate a binary cross-entropy loss that distills global knowledge. The binary KD is formulated as

$$\mathcal{L}_{\text{binary}} = \text{KL}(\tilde{\boldsymbol{b}}^{\text{tea}}||\tilde{\boldsymbol{b}}^{\text{stu}}), \quad (9)$$

where $\tilde{\boldsymbol{b}}^{\text{tea}} = \left[\frac{\sum_{t\in\Phi}\exp(\tilde{z}_t^{tea}/\tau)}{\sum_{j=1}^{C}\exp(\tilde{z}_j^{tea}/\tau)}, \frac{\sum_{t\notin\Phi}\exp(\tilde{z}_t^{tea}/\tau)}{\sum_{j=1}^{C}\exp(\tilde{z}_j^{tea}/\tau)}\right]$ represents the speaker posteriors of the primary group and the non-primary group. $\tilde{\boldsymbol{b}}^{\text{stu}}$ can be expressed in a similar way. Instead of directly using two logits for computing KL divergence, adaptive logit softening is applied to smooth the teacher and student output distributions, potentially reducing the discrepancy between primary and non-primary groups.

The student network is optimized by minimizing $\mathcal{L}_{\text{GKD}}$ and an AAM softmax classification loss [12]:

$$\mathcal{L} = \mathcal{L}_{\text{AAM}} + \omega\mathcal{L}_{\text{GKD}}, \quad (10)$$

where ω is a hyperparameter for balancing two losses.

III. Experimental Settings

A. Training and Evaluation Strategy

The development set of VoxCeleb2 [5] was utilized to train the student network, while three test sets sampled from VoxCeleb1 [35], i.e., Vox1-O, Vox1-E, and Vox1-H, were used to evaluate the performance of the proposed method. The teacher model comprises a large WavLM combined with a powerful ECAPA-TDNN. The WavLM was pre-trained on LibriSpeech and subsequently fine-tuned on the VoxCeleb2 development set. A simple x-vector network was selected as the student.

During training, we fed the augmented waveform of a short segment containing 200 frames to the frozen teacher model. 80-dim Mel-filter bank features were extracted and then presented to the student network after applying augmentations with a probability of 0.6. MUSAN [36] and RIRs [37] were adopted in the augmentation process. For AAMSoftmax, the scale and margin are consistent with the settings in [33].

During inference, speaker embeddings were extracted from the student network. A cosine similarity score was computed for each trial. No score normalization techniques were applied. Both EER and minDCF were reported for comparison.

B. Hyperparameters Settings

The base temperature τ was set to 4. In the proposed GKD method, β was set to 1, whereas α was set to 4 for

TABLE I

PERFORMANCE ON VOX1-O, VOX1-E, AND VOX1-H WITH k = 200. THE FIRST AND SECOND ROWS DO NOT USE ANY KD, AND THEY USE THE TEACHER AND STUDENT DURING INFERENCING, RESPECTIVELY.

Knowledge Distillation Method	Teacher	Student	#Param(M) during inferencing	Vox1-O EER% / minDCF	Vox1-E EER% / minDCF	Vox1-H EER% / minDCF
None	WavLM + ECAPA-TDNN	–	316.62	0.43 / –	0.54 / –	1.15 / –
None	–	x-vector	4.61	1.99 / –	1.95 / –	3.41 / –
Conventional KD (Eq. 1)	WavLM + ECAPA-TDNN			1.74 / 0.162	1.69 / 0.185	2.96 / 0.283
DKD (Eq. 5)				1.55 / 0.166	1.53 / 0.173	2.77 / 0.266
GKD with adaptive logit softening (Eq. 8 and 9)				**1.46 / 0.157**	**1.47 / 0.168**	**2.70 / 0.256**

amplifying the contribution of the primary loss. The parameter ω was increased from 0.05 to 1 during the first 20 epochs and kept unchanged in the remaining epochs. For the top-k selection, the value of k was set to 50, 150, 200, and 300, respectively. The proposed method was primarily developed using the *WeSpeaker* toolkit [38].

IV. RESULTS AND DISCUSSIONS

A. Comparisons of KD methods

Table I reports the results on Vox1-O, Vox1-E, and Vox1-H, respectively, when k was set to 200. It is obvious that with the extra supervisory information provided by the teacher model, distillation-based approaches consistently yield better performance than the same student network without KD, i.e. using $\mathcal{L}_{\text{AAM}}$ in Eq. 10 only. Compared to an x-vector network trained in a conventional distillation manner, our proposed system achieves the largest improvement of 16.1% on Vox1-O in terms of EER. Although training an x-vector with a DKD loss can also enhance performance by decoupling the non-target speakers' information from the target speaker's one, our proposed method outperforms it by a large margin, validating that combining adaptive logit softening and top-k posterior selection for KD can promote discriminative knowledge transfer.

B. Effect of GKD and Adaptive Logit Softening

As shown in Table II, splitting the KD loss into primary and binary losses leads to a significant performance boost. This observation is evident by comparing Row 1 with Row 3. On the other hand, compared with the vanilla GKD in Row 3, the GKD with adaptive logit softening (Row 5) achieves better performance. The performance improvement suggests that adaptive logit softening is beneficial to knowledge transfer by smoothing out the speaker posteriors. However, it is observed that the improvement is not obvious when normalizing teacher and student logits in the conventional KD (Comparing Row 1 with Row 2). We conjecture that without focusing on the primary group as in the proposed method for major knowledge transfer, applying adaptive logit softening across all the classes may introduce noisy information for the distillation due to the small posteriors. To further validate the effect of adaptive logit softening, we built a small corpus comprising $29,970$ utterances sampled from 5,994 speakers in the VoxCeleb2 development set, with each speaker contributing 5 utterances. The corresponding teacher logits were obtained by feeding the utterances to the pre-trained teacher model. Subsequently, we computed the standard deviations of the teacher logits for each utterance and denoted it as σ_k. They were then averaged across those utterances, resulting in $\bar{\sigma} = 2.74$. We replaced σ^{tea} and σ^{stu} in Eq. 6 by $\bar{\sigma}$ to soften the speaker posteriors. Comparing Row 4 and Row 5, we observe that it is critical to soften the logits dynamically because the posterior distributions vary from utterance to utterance.

TABLE II

EFFECT OF ADAPTIVE LOGIT SOFTENING AND GKD.

Row	Loss Function	Vox1-O EER% / minDCF	Vox1-E EER% / minDCF	Vox1-H EER% / minDCF
1	KL	1.74 / 0.162	1.69 / 0.185	2.96 / 0.283
2	KL with adaptive logit softening	1.78 / 0.174	1.69 / 0.189	2.87 / 0.279
3	GKD with τ	1.51 / 0.158	1.50 / 0.173	2.75 / 0.270
4	GKD with $\bar{\sigma} * \tau$	1.47 / 0.164	1.50 / 0.175	2.76 / 0.266
5	GKD with adaptive logit softening	**1.46 / 0.157**	**1.47 / 0.168**	**2.70 / 0.256**

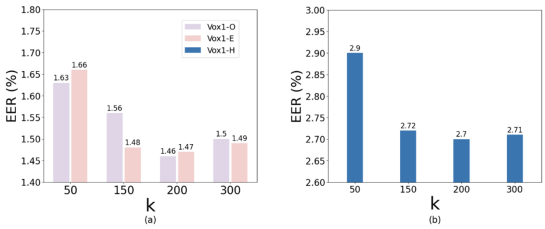

Fig. 2. EER on (a) Vox1-O, Vox1-E and (b) Vox1-H with respect to k.

C. Effect of k in Posteriors Selection

The hyperparameter k plays a critical role in GKD. We initially set k to 50. Fig. 2 shows the correlation between k and EER. Specifically, a notable improvement is observed when transitioning the value of k, followed by a subsequent drop in performance in terms of EER. It is observed that the best performance was obtained when k was set to 200.

V. CONCLUSIONS

This paper introduced a straightforward knowledge distillation method by selecting and emphasizing the information of top-k similar speakers for knowledge transfer, thereby enhancing the discriminative power of the student network. Adaptive logit softening incorporated in the binary distillation can further enhance the KD performance. The proposed method exhibited outstanding performance when compared to the method based on vanilla KD and also outperformed the current DKD-based approach with a noticeable margin.

REFERENCES

[1] Man-Wai Mak and Jen-Tzung Chien, *Machine Learning for Speaker Recognition*, Cambridge University Press, 2020.

[2] Zhongxin Bai and Xiao-Lei Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[3] Youzhi Tu, Weiwei Lin, and Man-Wai Mak, "A survey on text-dependent and text-independent speaker verification," *IEEE Access*, vol. 10, pp. 99038–99049, 2022.

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "VoxCeleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, pp. 101027, 2020.

[6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

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

[8] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.

[9] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[10] Jee-weon Jung, You Jin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung, "Pushing the limits of raw waveform speaker recognition," in *Proc. Interspeech*, 2022.

[11] Lantian Li, Ruiqian Nai, and Dong Wang, "Real additive margin softmax for speaker verification," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7527–7531.

[12] Zili Huang, Shuai Wang, and Kai Yu, "Angular softmax for short-duration text-independent speaker verification.," in *Proc. Interspeech*, 2018, pp. 3623–3627.

[13] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4879–4883.

[14] Dao Zhou, Longbiao Wang, Kong Aik Lee, Yibo Wu, Meng Liu, Jianwu Dang, and Jianguo Wei, "Dynamic margin softmax loss for speaker verification.," in *Proc. Interspeech*, 2020, pp. 3800–3804.

[15] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han, "In defence of metric learning for speaker recognition," in *Proc. Interspeech*, 2020.

[16] Yoohwan Kwon, Hee Soo Heo, Bong-Jin Lee, and Joon Son Chung, "The ins and outs of speaker recognition: lessons from VoxSRC 2020," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

[17] Chong-Xin Gan, Man-Wai Mak, Weiwei Lin, and Jen-Tzung Chien, "Asymmetric clean segments-guided self-supervised learning for robust speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2024, pp. 11081–11085.

[18] Zhe Li, Man-Wai Mak, and Helen Mei-Ling Meng, "Discriminative speaker representation via contrastive learning with class-aware attention in angular space," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

[19] Ruijie Tao, Kong Aik Lee, Rohan Kumar Das, Ville Hautamäki, and Haizhou Li, "Self-supervised speaker recognition with loss-gated learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 6142–6146.

[20] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. International Conference on Computer Vision*, 2007, pp. 1–8.

[21] 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*, vol. 33, pp. 12449–12460, 2020.

[22] 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*, vol. 29, pp. 3451–3460, 2021.

[23] 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*, vol. 16, no. 6, pp. 1505–1518, 2022.

[24] Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6147–6151.

[25] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, and et al, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

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

[27] Leying Zhang, Zhengyang Chen, and Yanmin Qian, "Knowledge distillation from multi-modality to single-modality for person verification," in *Proc. Interspeech*, 2021, pp. 1897–1901.

[28] Shuai Wang, Yexin Yang, Tianzhe Wang, Yanmin Qian, and Kai Yu, "Knowledge distillation for small foot-print deep speaker embedding," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 6021–6025.

[29] Xuechen Liu, Md Sahidullah, and Tomi Kinnunen, "Distilling multi-level x-vector knowledge for small-footprint speaker verification," *arXiv preprint arXiv:2303.01125*, 2023.

[30] Jungwoo Heo, Chan yeong Lim, Ju ho Kim, Hyun seo Shin, and Ha-Jin Yu, "One-Step Knowledge Distillation and Fine-Tuning in Using Large Pre-Trained Self-Supervised Learning Models for Speaker Verification," in *Proc. Interspeech*, 2023, pp. 5271–5275.

[31] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang, "Decoupled knowledge distillation," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11953–11962.

[32] Ting-Wei Chen, Chia-Ping Chen, Chung-Li Lu, Bo-Cheng Chan, Yu-Han Cheng, Hsiang-Feng Chuang, and Wei-Yu Chen, "A lightweight speaker verification model for edge device," in *Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2023, pp. 1372–1377.

[33] 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. International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10336–10340.

[34] Weisong Zhao, Xiangyu Zhu, Kaiwen Guo, Xiao-Yu Zhang, and Zhen Lei, "Grouped knowledge distillation for deep face recognition," in *Proc. AAAI Conference on Artificial Intelligence*, 2023, pp. 3615–3623.

[35] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Proc. Interspeech*, 2017.

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

[37] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5220–5224.

[38] Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian, "WeSpeaker: A research and production oriented speaker embedding learning toolkit," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.