Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

IDIR: Identifying and Distilling Informative Relations for Speaker Verification

Chong-Xin Gan, Zhe Li, Zezhong Jin, Zilong Huang, Man-Wai Mak, Kong Aik Lee

Dept. of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

chong-xin.gan@connect.polyu.hk

Abstract

Traditional feature-based knowledge distillation aligns the student's features with the teacher's features. However, these one-to-one alignments overlook the structural relations between speakers in a mini-batch. Also, the large capacity gap between the two networks causes significant discrepancies between their features. To address these limitations, we propose distilling the inter- and intra-speaker relations. Instead of mimicking all pairwise relations between the student's and teacher's feature vectors, we propose **I**dentifying and **D**istilling **I**nformative **R**elations (IDIR), enabling the student network to acquire speakers' relationships from the teacher. Moreover, a margin is added to the similarity scores of the informative pairs, further reducing intra-speaker variances and increasing inter-speaker separations. Evaluations with a simple x-vector student network demonstrate the method's superb performance across three test sets, showcasing its merits and effectiveness.

Index Terms: knowledge distillation, feature distillation, speaker recognition

1. Introduction

Deep neural network (DNN)-based approaches have greatly improved the performance of speaker verification systems [1, 2], particularly when trained on large-scale datasets [3] with sophisticated loss functions [4, 5]. Compared to the traditional *i*-vector framework [6], which relies on a Gaussian mixture model (GMM) and a universal background model (UBM), DNN-based speaker embedding networks have shown superior capabilities in feature extraction and noise robustness. During inference, the speaker representations extracted from the penultimate layer of the speaker embedding network are typically processed using a cosine similarity-based backend for decision making.

Recently, Transformer-based architectures [7], originally proposed for natural language processing (NLP), have been successfully repurposed for computer vision and speech tasks. In the context of speaker verification, numerous studies leveraged the strong capabilities of large speech models [8–10] that were pre-trained on extensive corpora, cascading them with traditional speaker encoders [11, 12]. This combination has demonstrated impressive results, largely boosting the performance of speaker verification [13, 14]. However, the performance improves with the increasing number of trainable parameters, resulting in powerful yet cumbersome models. To adapt the model to some specific tasks, parameter-efficient fine-tuning (PEFT) is developed to update only a subset of the model's parameters instead of training the entire model from scratch [15, 16]. While these techniques effectively reduce training costs, deploying these bulky models is still quite difficult, especially for resource-constrained edge devices. To address this limitation, techniques such as model pruning [17], quantization [18], and knowledge distillation [19] have been extensively explored. Among these, knowledge distillation (KD) stands out for its simplicity and effectiveness. It trains a lightweight student network by leveraging additional supervision of a large teacher network, by aligning predictions at the logit level, or by distilling feature-level knowledge.

Knowledge distillation is generally classified into three categories based on the distillation position: logit-level KD [19], feature-level KD [20], and a hybrid approach that combines the two [21]. Logit-level KD directly minimizes the Kullback–Leibler (KL) divergence between the predicted output distributions of the teacher and student networks. Previous studies, such as [22, 23], have demonstrated that speaker verification performance can be improved by leveraging the decoupled information of non-target speakers. Building on this concept, grouped knowledge distillation is introduced to further enhance KD by dividing the speaker posteriors into a primary group and a non-primary group [24]. Although these methods have shown promising results, their performance is highly sensitive to hyperparameter configurations, which increases the computational cost of identifying optimal settings. To address this issue, an adversarial temperature adjustment mechanism was proposed for KD loss, where the temperature varies dynamically across different training stages [25]. While this approach mitigates some of the challenges related to hyperparameter sensitivity, the requirement of training a discriminator to estimate optimal temperatures introduces additional complexity and effort, making it less practical in certain scenarios.

Unlike logit-based KD, feature-based KD is intuitively more suitable for speaker verification tasks, as the classification head is typically discarded after training. Moreover, as the number of training speakers continues to increase, it becomes more challenging for the student network to replicate the teacher's speaker posterior distributions since the posteriors of many non-target speakers are nearly zero. Feature-level KD can effectively bypass this issue by aligning the intermediate representations of the student and teacher networks with metrics such as mean squared error (MSE) [26, 27] or cosine similarity [28]. However, guiding the student network to produce identical speaker representations via MSE is unrealistic due to the significant capacity gap between the student and teacher networks. In contrast, maximizing cosine similarity offers greater flexibility and serves as a more effective training strategy. Nevertheless, traditional feature distillation operates at the sample level, failing

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

10.21437/Interspeech.2025-736

to capture the structural relationships among different speakers. To overcome this limitation, [28, 29] proposed transferring inter-class relations across all samples in a mini-batch. While effective, considering all pairwise relations is suboptimal, as not all relationships contribute equally to knowledge transfer. In [30], relational information among positive-anchor-negative pairs was repurposed to address domain mismatch. Different from them, we emphasize the informative inter-speaker relationships to enhance the distillation. Furthermore, the centers of class-dependent features derived from the teacher network are pre-computed to facilitate the transfer of intra-speaker relationships from the teacher network to the student network.

In summary, the contributions to this paper are as follows. Firstly, we designed two simple strategies to identify and transfer the most informative inter-speaker relations from the teacher to the student, enhancing the efficiency and effectiveness of the knowledge distillation process. Secondly, by utilizing a frozen teacher network to pre-compute the centers of speaker-dependent features, we enable the effective transfer of intra-speaker relations, promoting smaller intra-speaker variances. Thirdly, a fixed margin is applied to both inter- and intra-speaker relations, guiding the student to achieve better performance.

2. Methodology

2.1. Conventional Logit and Feature Distillations

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. The teacher $f^{\text{tea}}(.)$ and student $f^{\text{stu}}(.)$ networks map the short segment $\boldsymbol{x}$, which is randomly sampled from $\boldsymbol{u}$, to the corresponding embeddings $\boldsymbol{e}^{\text{tea}}$ and $\boldsymbol{e}^{\text{stu}}$, respectively. The embeddings are then transformed into respective speaker posterior probabilities $\boldsymbol{p}^{\text{tea}}$ and $\boldsymbol{p}^{\text{stu}}$ via a softmax function. Logit-KD minimizes the KL divergence between the classification probabilities outputted from student and teacher networks, and feature-KD reduces the discrepancies between the student's and teacher's intermediate features.

The logit-based KD and feature-based KD losses can be formulated as follows

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

$$\mathcal{L}_{\text{KD}}^{\text{feat}} = 1 - \frac{\tilde{\boldsymbol{e}}^{\text{stu}} \cdot \boldsymbol{e}^{\text{tea}}}{||\tilde{\boldsymbol{e}}^{\text{stu}}||\,||\boldsymbol{e}^{\text{tea}}||} \quad \text{or} \quad \sum_{d=1}^{D} (\boldsymbol{e}_d^{\text{tea}} - \tilde{\boldsymbol{e}}_d^{\text{stu}})^2, \quad \text{(1b)}$$

where $i \in \{1, \ldots, C\}$ indexes the classes and D denotes the dimension of speaker embeddings. Noted that a projector $g^{\text{stu}}(.)$ may be required to transform $\boldsymbol{e}^{\text{stu}}$ to $\tilde{\boldsymbol{e}}^{\text{stu}}$, enabling the dimension of the student embeddings to be consistent with that of the teacher embeddings.

2.2. Inter-Speaker Relations Distillation

Compared to the student network, the teacher network is more capable of establishing an embedding space where the embeddings of different speakers are well-separated and the embeddings of the same speaker remain closely clustered. However, simply forcing the student network's embeddings to align with those of the teacher may hinder the student network from developing its own embedding space, potentially resulting in suboptimal performance. To address this limitation, rather than simply limiting the knowledge transfer through the one-to-one correspondence between the teacher's and student's features, we propose forcing the student network to learn the *relations* between the embeddings of different speakers exhibited by the teacher network.

Fig. 1 shows the overview of the proposed framework. A mini-batch consisting of N segments from K speakers ($K \leq N$) is fed into the two networks. After obtaining the speaker embeddings of the N segments, a cosine similarity matrix $\boldsymbol{S}^{\text{tea}} \in \mathbb{R}^{N \times N}$ is constructed to capture the relations between the K speakers. A masking function is used to remove the similarity scores from the same speakers. The resulting matrix is defined as:

$$S_{k,j}^{\text{tea}} = \frac{\boldsymbol{e}_k^{\text{tea}} \cdot \boldsymbol{e}_j^{\text{tea}}}{||\boldsymbol{e}_k^{\text{tea}}||\,||\boldsymbol{e}_j^{\text{tea}}||}, \quad 1 \leq j, k \leq N \quad \text{and} \quad \text{ID}_k \neq \text{ID}_j, \quad (2)$$

where $\boldsymbol{e}_k^{\text{tea}}$ and $\boldsymbol{e}_j^{\text{tea}}$ represent the embeddings of the j-th and k-th utterances in the mini-batch, and ID_r is the speaker ID of utterance r. Therefore, after applying the mask, only the similarity scores of different speakers (the speaker labels of $\boldsymbol{e}_k^{\text{tea}}$ and $\boldsymbol{e}_j^{\text{tea}}$ are different) are retained. A similar operation is performed on the student's embeddings, resulting in a student's similarity matrix:

$$S_{k,j}^{\text{stu}} = \frac{\boldsymbol{e}_k^{\text{stu}} \cdot \boldsymbol{e}_j^{\text{stu}}}{||\boldsymbol{e}_k^{\text{stu}}||\,||\boldsymbol{e}_j^{\text{stu}}||}, \quad 1 \leq j, k \leq N \quad \text{and} \quad \text{ID}_k \neq \text{ID}_j. \quad (3)$$

Instead of transferring every inter-speaker relation, we devise two strategies to identify and focus on the most informative speaker pairs. The similarity scores in $\boldsymbol{S}^{\text{stu}}$ reflect the student network's current learning capability. The highest similarity scores in every row of $\boldsymbol{S}^{\text{stu}}$ are selected to form a vector $\boldsymbol{s}^{\text{stu}} = \left[s_1^{\text{stu}}, \ldots, s_N^{\text{stu}}\right]$. The corresponding similarity scores in $\boldsymbol{S}^{\text{tea}}$ are extracted to create a matching vector $\boldsymbol{s}^{\text{tea}} = \left[s_1^{\text{tea}}, \ldots, s_N^{\text{tea}}\right]$. To further promote inter-cluster separation, a fixed positive margin m_1 is added:

$$\mathcal{L}_{\text{KD}}^{\text{relation-max}} = \sum_{l=1}^{N} \left(\left(s_l^{\text{tea}} - m_1\right) - s_l^{\text{stu}}\right)^2. \quad (4)$$

However, the teacher model does not always outperform the student network. Thus, $\mathcal{L}_{\text{KD}}^{\text{relation-max}}$ is reformulated as:

$$\mathcal{L}_{\text{KD}}^{\text{relation-max}} = \sum_{l \in \mathcal{I}_1} \left[\min\left\{s_l^{\text{tea}} - m_1, s_l^{\text{stu}}\right\} - s_l^{\text{stu}}\right]^2, \quad (5)$$

where $\mathcal{I}_1 = \{l \mid s_l^{\text{tea}} - m_1 < s_l^{\text{stu}}, \forall\, 1 \leq l \leq N\}$ comprises a set of indices corresponding to the selected similarity scores such that the teacher is more capable than the student by the margin m_1.[1] In addition to those pairs with the maximum similarities, we also identify pairs with the largest discrepancies between the teacher and student networks. After filtering out irrelevant similarity scores based on the condition $S_{k,j}^{\text{tea}} < S_{k,j}^{\text{stu}}$, the gap between similarity scores is computed as:

$$S_{k,j}^{\text{gap}} = \left[\min\left\{S_{k,j}^{\text{tea}}, S_{k,j}^{\text{stu}}\right\} - S_{k,j}^{\text{stu}}\right]^2, \quad 1 \leq k, j \leq N. \quad (6)$$

The largest gaps from every row of $\boldsymbol{S}^{\text{gap}}$ form a vector $\boldsymbol{s}^{\text{gap}}$, indicating that the student and teacher have large discrepancies regarding inter-speaker relations. Since these discrepancies are inherently large, no margin is added to avoid complicating the

[1] For inter-speaker similarity, the lower the score, the higher the capability.

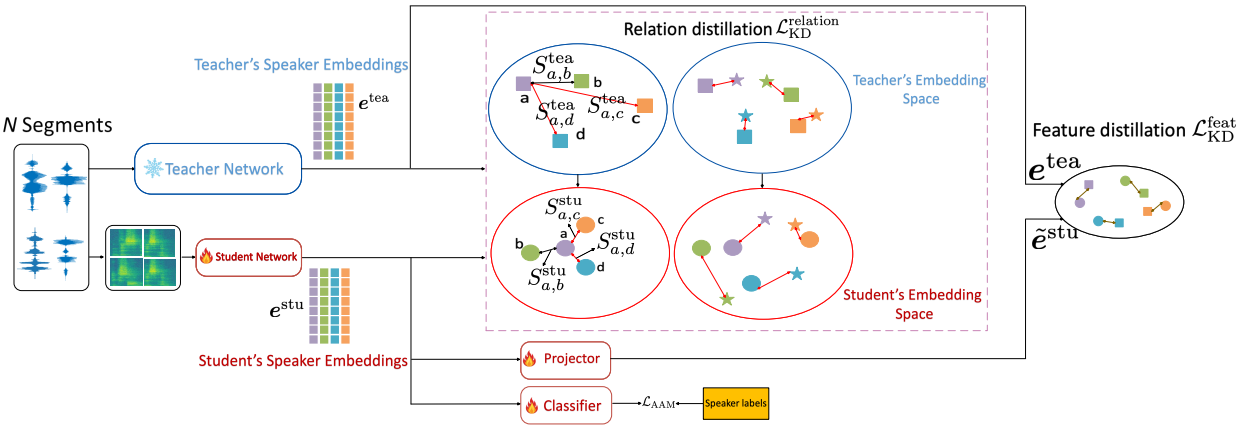

Figure 1: *Framework of the proposed relation distillation method. A mini-batch of N segments is fed into the teacher and student networks. The obtained speaker embeddings are used to compute the inter-speaker relations for the teacher and the student. Considering the purple marker "a" as an anchor embedding, as shown in Eq. 5, the student network learns to produce $S_{a,d}^{\text{stu}}$ that matches $S_{a,d}^{\text{tea}}$ as closely as possible. $S_{a,c}^{\text{stu}}$ is enforced to match $S_{a,c}^{\text{tea}}$ according to Eqs. 6 and 7. On the other hand, class-center embeddings (the star markers) are computed offline using the frozen teacher network so that the intra-speaker relations (teacher-to-center) can also be calculated, facilitating the intra-speaker relation knowledge transfer. The student network is optimized with three losses: relation and feature distillation losses and classification loss. Squares and circles represent teacher and student embeddings, respectively. Stars denote class-center embeddings. Different colors indicate different speakers. Red arrows denote the sources of relation distillation.*

student's training process. The gap-based inter-speaker relation loss is defined as:

$$\mathcal{L}_{\text{KD}}^{\text{relation-gap}} = \sum_{l \in \mathcal{I}_2} s_l^{\text{gap}}, \tag{7}$$

where $\mathcal{I}_2 = \{l \mid s_l^{\text{gap}} > 0, \forall\, 1 \le l \le N\}$. The combination of these two relation-based strategies helps the student network focus on the most informative inter-speaker relations, enhancing the robustness and generalization of the learned speaker embeddings. The overall inter-speaker relation distillation loss is formulated as:

$$\mathcal{L}_{\text{KD}}^{\text{inter}} = \mathcal{L}_{\text{KD}}^{\text{relation-max}} + \mathcal{L}_{\text{KD}}^{\text{relation-gap}} \tag{8}$$

2.3. Intra-Speaker Relations Distillation

In addition to capturing inter-speaker relations for knowledge distillation, it is equally critical to transfer the intra-speaker relations to the student network. However, with a small batch size and a large number of training speakers in a dataset, it is unlikely that a mini-batch contains multiple utterances from the same speaker, making the modeling of intra-speaker relations difficult. While it is possible to increase the number of same-speaker utterances by extracting multiple segments from the same utterances (similar to contrastive learning in speaker embedding networks), the intra-speaker similarities will be overestimated.

To address the above issue, we leverage the pre-trained teacher network to extract speaker representations from the entire training dataset. These representations are then averaged for each speaker, producing class-center embeddings $\boldsymbol{c}_i$ corresponding to speaker i. The intra-speaker similarities, $\boldsymbol{a}^{\text{tea}} = \left[a_1^{\text{tea}}, \ldots, a_K^{\text{tea}}\right]$, between the teacher's embeddings and their respective class-center embeddings, are defined as:

$$a_k^{\text{tea}} = \frac{\boldsymbol{e}_k^{\text{tea}} \cdot \boldsymbol{c}_k}{||\boldsymbol{e}_k^{\text{tea}}||\,||\boldsymbol{c}_k||}, \quad 1 \le k \le K. \tag{9}$$

Similarly, intra-speaker relations $\boldsymbol{a}^{\text{stu}} = \left[a_1^{\text{stu}}, \ldots, a_K^{\text{stu}}\right]$ for the student network are calculated using the same class-center embeddings:

$$a_k^{\text{stu}} = \frac{\boldsymbol{e}_k^{\text{stu}} \cdot \boldsymbol{c}_k}{||\boldsymbol{e}_k^{\text{stu}}||\,||\boldsymbol{c}_k||}, \quad 1 \le k \le K. \tag{10}$$

To consider the situation where the student network may outperform the teacher, we incorporate a fixed margin m_2 into the intra-speaker relation loss:

$$\mathcal{L}_{\text{KD}}^{\text{intra}} = \sum_{l \in \mathcal{I}_3} \left[\max\left\{a_l^{\text{tea}} + m_2, a_l^{\text{stu}}\right\} - a_l^{\text{stu}}\right]^2, \tag{11}$$

where $\mathcal{I}_3 = \left\{l \mid a_l^{\text{tea}} + m_2 > a_l^{\text{stu}}, \forall\, 1 \le l \le K\right\}$ performs a similar role to $\mathcal{I}_1$. The margin m_2 allows knowledge distillation to occur even if the student surpasses the teacher.[2]

The proposed IDIR framework effectively guides the student network in learning both intra-speaker compactness and inter-speaker separation. More importantly, this is achieved without introducing additional computational costs during either training or inference. The student network is optimized by jointly minimizing three KD losses and an AAM-softmax classification loss [31]. The final objective function is expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{AAM}} + \omega \left(\mathcal{L}_{\text{KD}}^{\text{feat}} + \underbrace{\mathcal{L}_{\text{KD}}^{\text{inter}} + \mathcal{L}_{\text{KD}}^{\text{intra}}}_{\mathcal{L}_{\text{KD}}^{\text{relation}}} \right), \tag{12}$$

where ω is a hyperparameter that balances the contribution of the classification loss and the distillation losses.

3. Experimental Settings

The development set of VoxCeleb2 [32] was utilized to train the student network, while three test sets sampled from VoxCeleb1 [33], i.e., Vox1-O, Vox1-E, and Vox1-H, were used to

[2]For intra-speaker similarity, the higher the score, the higher the capability.

Table 1: *Performance on Vox1-O, Vox1-E, and Vox1-H. The systems in Rows 1 and 2 do not use any KD during training, and they use the teacher and student networks during inference, respectively. Except for Row 3, all experiments were conducted using a 512-dim speaker embedding network .*

System	Row	Distillation Type	Feature Distillation Loss Metric	#Param(M) during inferencing	Vox1-O EER%/minDCF	Vox1-E EER%/minDCF	Vox1-H EER%/minDCF
Teacher model WavLM + ECAPA-TDNN	1	–	N/A	316.62	0.43/–	0.54/–	1.15/–
Student model x-vector	2	–	N/A	4.39	1.99/–	1.95/–	3.41/–
	3	Embedding (dim=256)	*Cosine*		1.93/0.199	1.86/0.203	3.18/0.289
	4	Embedding	*MSE*		1.73/0.173	1.74/0.194	3.02/0.287
	5	Embedding	*Cosine*		1.63/**0.163**	1.72/0.194	3.03/0.292
	6	Embedding + Relation (all)	*Cosine*		1.74/0.165	1.76/**0.193**	3.02/0.288
	7	Embedding + Relation (selected)	*Cosine*		**1.55**/0.171	**1.70**/0.195	**2.94/0.282**
	8	Logit + Embedding	*Cosine*		1.65/0.175	1.70/0.184	2.94/0.278
	9	Logit + Embedding + Relation (selected)	*Cosine*		**1.57/0.143**	**1.65/0.182**	**2.89/0.274**

evaluate the performance of the proposed method. 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. MUSAN [34] and RIR [35] were adopted in the augmentation process.

The teacher model comprises a large WavLM combined with a powerful ECAPA-TDNN pretrained on LibriSpeech, and subsequently fine-tuned on VoxCeleb2. A simple x-vector network was selected as the student. An additional projector comprising a linear layer followed by a normalization layer and an activation function is employed to match the dimension. It was removed during the inference stage.

In the proposed method, the parameter ω was increased from 0.05 to 1 during the first 20 epochs. The margins m_1 and m_2 were set to 0.3. MSE and Mean Absolute Error (MAE) were used interchangeably when computing relation distillation loss. In AAM-softmax, the margin was gradually increased from 0 to 0.2, and the scale was set to 32. During inference, speaker embeddings with a dimension of 512 were extracted from the student network. A cosine similarity score was computed for each trial. Both EER and minDCF were reported for comparison.

Table 2: *Comparions with the previous works on Vox1-O.*

Row	Method	Vox1-O EER% /minDCF
1	Vanilla logit-level KD	1.74/0.162
2	Vanilla feature-level KD	1.63/0.163
3	DKD [22]	1.56/0.166
4	IDIR with logit-level KD	1.57/**0.143**
5	IDIR	**1.55**/0.171

4. Results and Discussions

4.1. Main Results

Table 1 presents the performance of various feature-based KD methods across three test sets. Notably, the proposed method (Row 7) achieves superior performance compared to the baseline (Row 5), with a significant margin on Vox1-O and Vox1-H. This highlights the effectiveness of incorporating relational knowledge to enhance the KD process. Comparing Row 6 and Row 7 reveals that selecting all relations for knowledge transfer worsens the performance, as the student network is unable to replicate each pairwise relation, indicating the necessity of identifying informative relations. Furthermore, when compared to a simple x-vector network trained solely with the classification loss, our method achieves the most substantial relative improvements in EER, with reductions of 23%, 12%, and 14% on Vox1-O, Vox1-E, and Vox1-H, respectively. Using MSE to align the teacher and student embedding yields inferior results on Vox1-O compared to using the cosine criterion, as evidenced in Row 4 and Row 5. This suggests that the cosine-based alignment provides more flexibility for the student's learning. However, the performance improvements on Vox1-O and Vox1-E are marginal when combining logit-level KD with embedding-level KD, as reported in Row 5 and Row 8. We conjectured that this limited improvement could be caused by gradient conflicts between the objectives of logit-level and embedding-level KD, which could hinder the optimization process. Nevertheless, after integrating relational knowledge transfer, noticeable enhancements are observed across three test sets. This demonstrates that relational knowledge effectively complements embedding-level and logit-level KD by capturing structural relationships within a mini-batch. The integration not only alleviates potential gradient conflicts but also enriches the knowledge distilled to the student model, leading to consistent and significant performance gains across Vox1-O, Vox1-E, and Vox1-H.

4.2. Comparisons with the Existing Works

To further demonstrate the effectiveness of the proposed method, the comparisons on Vox1-O between our approach and other advanced methods are presented in Table. 2. The results clearly show that the proposed method could achieve the best performance. Even compared with DKD [22], which is considered one of the SOTA approaches, our method delivers competitive results, as evidenced by the comparison between Row 3 and Row 5. This underscores the merits of IDIR, which provides an effective solution without requiring additional computational overhead.

5. Conclusions

In this paper, we developed a feature-based distillation framework to capture the relations between different speakers within a mini-batch. Rather than transferring relations for all possible speaker pairs, we identified the informative inter-speaker relations and intra-speaker relations for knowledge transfer, enabling the student network to learn useful knowledge. The experimental results substantiate the effectiveness of the proposed method, offering a viable alternative for KD. Furthermore, when integrated with logit-level KD, our method yields promising results.

6. References

[1] Y. Tu, W. Lin, and M.-W. Mak, "A survey on text-dependent and text-independent speaker verification," *IEEE Access*, vol. 10, pp. 99 038–99 049, 2022.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[3] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018.

[4] J.-w. Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," in *Proc. Interspeech*, 2022.

[5] L. Li, R. Nai, and D. Wang, "Real additive margin softmax for speaker verification," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7527–7531.

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

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

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

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

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

[12] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[13] S. wen Yang, P.-H. Chi, Y.-S. Chuang, and et al, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

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

[15] J. Peng, T. Stafylakis, R. Gu, O. Plchot, L. Mošner, L. Burget, and J. Černockỳ, "Parameter-efficient transfer learning of pre-trained transformer models for speaker verification using adapters," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[16] Z. Li, M.-W. Mak, H.-y. Lee, and H. Meng, "Parameter-efficient fine-tuning of speaker-aware dynamic prompts for speaker verification," in *Proc. Interspeech*, 2024, pp. 2675–2679.

[17] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the value of network pruning," *arXiv preprint arXiv:1810.05270*, 2018.

[18] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, 2021.

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

[20] X. Liu, M. Sahidullah, and T. Kinnunen, "Distilling multi-level x-vector knowledge for small-footprint speaker verification," *arXiv preprint arXiv:2303.01125*, 2023.

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

[22] D.-T. Truong, R. Tao, J. Q. Yip, K. A. Lee, and E. S. 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. 10 336–10 340.

[23] T.-W. Chen, C.-P. Chen, C.-L. Lu, B.-C. Chan, Y.-H. Cheng, H.-F. Chuang, and W.-Y. 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.

[24] C.-X. Gan, Y. Tu, Z. Jin, M.-W. Mak, and K. A. Lee, "Grouped knowledge distillation with adaptive logit softening for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2025, pp. 1–5.

[25] Z. Jin, Y. Tu, C.-X. Gan, M.-W. Mak, and K.-A. Lee, "Adversarially adaptive temperatures for decoupled knowledge distillation with applications to speaker verification," *Neurocomputing*, p. 129481, 2025.

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

[27] Z. Jin, Y. Tu, Z. Li, Z. Huang, C.-X. Gan, and M.-W. Mak, "Denoising student features with diffusion models for knowledge distillation in speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2025, pp. 1–5.

[28] Y. Jin, G. Hu, H. Chen, D. Miao, L. Hu, and C. Zhao, "Cross-modal distillation for speaker recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 12 977–12 985.

[29] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[30] Q. Gu, Y. Song, N. Jiang, P. Cai, and I. McLoughlin, "PNP-RKD: A positive-negative pair based relational knowledge distillation method for cross-domain speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2025, pp. 1–5.

[31] Z. Huang, S. Wang, and K. Yu, "Angular softmax for short-duration text-independent speaker verification." in *Proc. Interspeech*, 2018, pp. 3623–3627.

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

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

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

[35] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5220–5224.