SELF-KNOWLEDGE DISTILLATION VIA FEATURE ENHANCEMENT FOR SPEAKER VERIFICATION

Bei Liu, Haoyu Wang, Zhengyang Chen, Shuai Wang, Yanmin Qian[†]

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

{beiliu, fayuge, zhengyang.chen, feixiang121976, yanminqian}@sjtu.edu.cn

ABSTRACT

As the most widely used technique, deep speaker embedding learning has become predominant in speaker verification task recently. Very large neural networks such as ECAPA-TDNN and ResNet can achieve the state-of-the-art performance. However, large models are computationally unfriendly in general, which require massive storage and computation resources. Model compression has been a hot research topic. Parameter quantization usually results in significant performance degradation. Knowledge distillation demands a pretrained complex teacher model. In this paper, we introduce a novel self-knowledge distillation method, namely **S**elf-**K**nowledge **D**istillation via **F**eature **E**nhancement (SKDFE). It utilizes an auxiliary self-teacher network to distill its own refined knowledge without the need of a pretrained teacher network. Additionally, we apply the self-knowledge distillation at two different levels: label level and feature level. Experiments on Voxceleb dataset show that our proposed self-knowledge distillation method can make small models have comparable or even better performance than large ones. Large models can also be further improved when applying our method.

***Index Terms*—** speaker verification, deep embedding learning, model compression, self-knowledge distillation

1. INTRODUCTION

Currently, deep neural networks (DNNs) have been widely applied in speaker verification task and presented remarkable results [1, 2, 3, 4, 5, 6, 7]. Impressive performance can be achieved by deep speaker embedding learning with very large architectures such as ECAPA-TDNN [7] and ResNet [8]. However, there is a trade-off between efficiency and effectiveness. Powerful models with millions of parameters generally require tremendous storage and computation resources, which is hard to be deployed onto resource-limited devices in real life. On the contrary, small models are much easier for distribution while the performance is unsatisfactory. How to make small models have comparable or even better performance than large ones is a demanding task for speaker verification.

Accordingly, model compression has attracted the attention of many researchers [9, 10, 11, 12, 13]. Compressing models directly can lead to significant performance degradation such as network quantization [9, 13] or model size reduction [10]. Although large compression ratio can be achieved, compressed models yield unsatisfying performance, which is the main problem for this type of compression strategy. To boost the results of small models, knowledge distillation [14] is another choice, which is characterized by good compression and comparable performances [10]. Whereas knowledge distillation enables student network to utilize the teacher's knowledge, preparing a pretrained teacher network in advance is still a computationally expensive task. In addition, there still exists the performance gap between the teacher and student network after distillation.

To deal with the limitations of the traditional knowledge distillation, this paper introduces a novel paradigm called self-knowledge distillation for speaker verification task. Combined with the proposed feature enhancement module, we present a new self-knowledge distillation method, namely **S**elf-**K**nowledge **D**istillation via **F**eature **E**nhancement (SKDFE). Self-knowledge distillation is designed to reduce the necessity of pretraining a teacher network beforehand, which utilizes an auxiliary self-teacher network to enhance feature representation and transfer the refined knowledge to self-student network. We employ multi-path feature pyramid network as self-teacher network to yield the enhanced and refined knowledge. Moreover, label level and feature level guidances are utilized to transfer knowledge better, which allows us to further narrow the performance gap between the self-teacher and self-student network.

2. RELATED WORKS

In this section, we briefly present previous works related to our proposed approach, including knowledge distillation and self-knowledge distillation.

2.1. Knowledge Distillation

Knowledge distillation is first introduced in [14], which has been widely used in various fields. For speaker verification [10, 11, 15], [10] introduces embedding level guidance that directly makes uses of teacher network's speaker embedding to boost student network, providing minimum square error (MSE) learning and cosine distance learning. [11] develops two alternatives of knowledge distillation and random erasing to improve the generalization and robustness of text-dependent speaker verification systems. [15] investigates the possibility of distilling knowledge from a multi-modality system to a single-modality system.

2.2. Self-Knowledge Distillation

Different from knowledge distillation, self-knowledge distillation transfers the enhanced and refined knowledge originating from stu-

[†]corresponding author

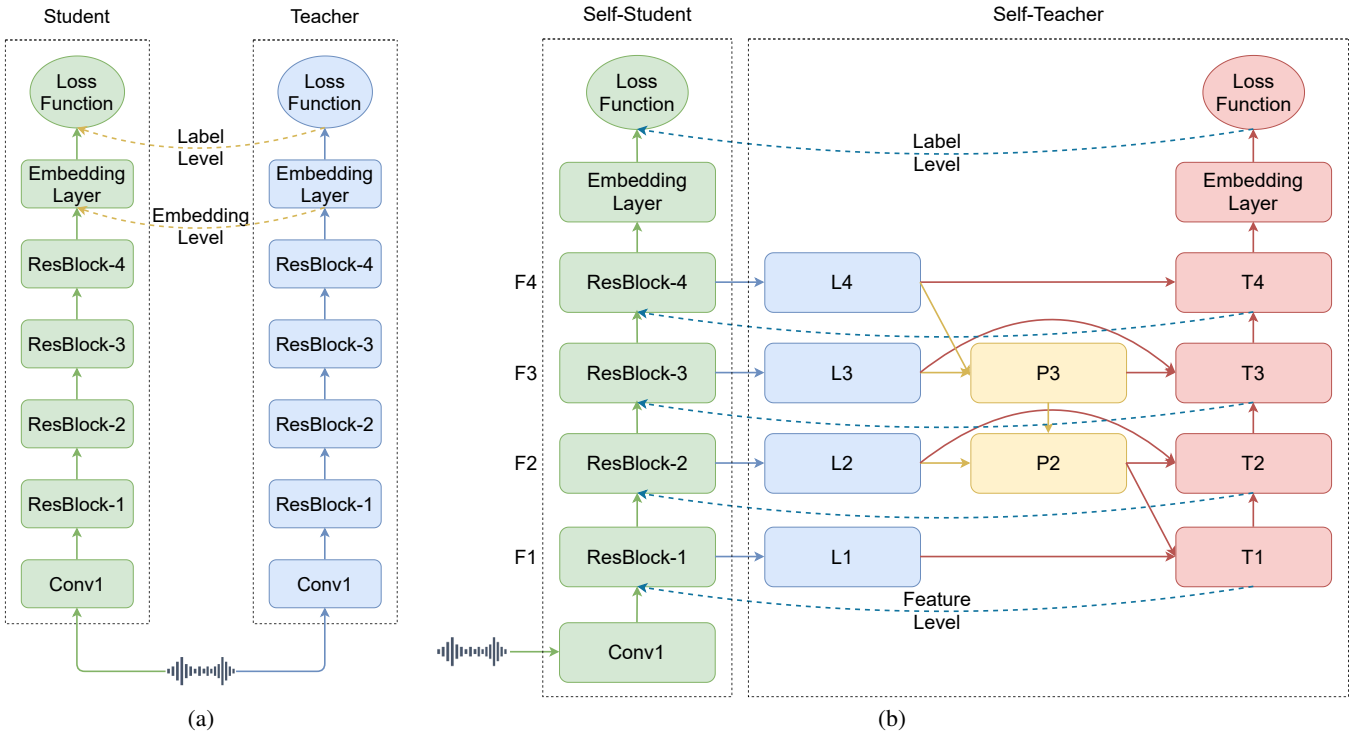

Fig. 1. Comparison of knowledge distillation and self-knowledge distillation. (a) Knowledge Distillation: A pretrained teacher network is employed to guide the training of student network. (b) Self-Knowledge Distillation via Feature Enhancement: An auxiliary self-teacher network is utilized to distill the refined knowledge to self-student network without a pretrained model.

dent network itself without a pretrained teacher network. Generally speaking, auxiliary network, which has the ability to capture global information and aggregate features from various layers, is adopted as self-teacher network to provide more powerful feature representation and guide the training of student network. In computer vision, several strategies have been explored to perform self-knowledge distillation. [16] introduces a set of auxiliary branches for the middle hidden layers within a network to enhance the performance of shallow layers. [17] proposes to train a single multi-branch network while establishing a strong teacher on-the-fly to enhance the learning of target network. In this paper, we introduce a more complex auxiliary feature enhancement network as self-teacher network to generate refined knowledge for speaker verification task.

3. PROPOSED METHODS

In this section, we introduce the self-knowledge distillation via feature enhancement (SKDFE) based on ResNet. Figure 1(a) shows the traditional knowledge distillation. Figure 1(b) is the overview of our proposed method.

3.1. Self-Student and Self-Teacher Network

For the self-student network, the left side in Figure 1(b), ResNet is used in this paper. As shown on the right side of Figure 1(b), the self-teacher network, as a main component of self-knowledge distillation method, is utilized to provide the self-student network with enhanced feature maps and soft labels for the purpose of distillation. We adopt BiFPN from [18] as the self-teacher network to produce refined knowledge. Architecture details of the self-student network and self-teacher network are listed in Table 1 and Table 2 respectively.

For the self-student network, we represent the i-th stage feature map as F_i, for $i = 1, ..., 4$. For the self-teacher network, lateral convolution is firstly calculated as follows:

$$L_i = Conv(F_i; d_i) \tag{1}$$

where $Conv$ is a depth-wise convolution [19] operation with a d_i number of output channels. The i-th lateral convolutional output is denoted as L_i.

Secondly, lateral outputs and previous top-down features are fed into top-down layers. A new intermediate feature map P_i is obtained:

$$P_i = Conv(I^P_{i,1} \cdot L_i + I^P_{i,2} \cdot Resize(P_{i+1}); d_i) \tag{2}$$

$$I^P_{i,j} = \frac{e^{w^P_{i,j}}}{\sum_k e^{w^P_{i,k}}}, j, k = 1, 2 \tag{3}$$

where P_i represents the i-th stage of the top-down path. $w^P_{i,j}$ is a learnable fusion weight parameter. $I^P_{i,j}$ is a normalized softmax value which means the importance of each input feature. The mismatch of dimensions between input features is dealt with by the $Resize$ operator where bilinear interpolation is adopted for up-sampling and max pooling is utilized for down-sampling.

Thirdly, bottom-up path is implemented to fuse features in bottom-up way, where the middle layer T_i accepts lateral, top-down and previous bottom-up features as inputs:

$$T_i = Conv(I^T_{i,1} \cdot L_i + I^T_{i,2} \cdot P_i + I^T_{i,3} \cdot Resize(T_{i-1}); d_i) \tag{4}$$

$$I^T_{i,j} = \frac{e^{w^T_{i,j}}}{\sum_k e^{w^T_{i,k}}}, j, k = 1, 2, 3 \tag{5}$$

Table 1. The ResNet34 self-student network. The dimensions of the output are $C \times F \times T$, i.e. the number of channels, filter-banks and frames.

Layer	Structure	Output	Stage
Conv1	[Conv2D-BN-ReLU]	$32 \times 40 \times T$	–
ResBlock-1	$\begin{bmatrix}\text{Conv2D-BN-ReLU}\\ \text{Conv2D-BN}\end{bmatrix} \times 3$	$32 \times 40 \times T$	1
ResBlock-2	$\begin{bmatrix}\text{Conv2D-BN-ReLU}\\ \text{Conv2D-BN}\end{bmatrix} \times 4$	$64 \times 20 \times T/2$	2
ResBlock-3	$\begin{bmatrix}\text{Conv2D-BN-ReLU}\\ \text{Conv2D-BN}\end{bmatrix} \times 6$	$128 \times 10 \times T/4$	3
ResBlock-4	$\begin{bmatrix}\text{Conv2D-BN-ReLU}\\ \text{Conv2D-BN}\end{bmatrix} \times 3$	$256 \times 5 \times T/8$	4

Table 2. The self-teacher network. The dimensions of the output are $C \times F \times T$, i.e. the number of channels, filter-banks and frames.

Layer	Structure	Output	Stage
L1		$256 \times 40 \times T$	1
L2	$\begin{bmatrix}\text{Conv2D-Conv2D-BN-ReLU}\\ \text{Conv2D-Conv2D-BN-ReLU}\end{bmatrix}$	$256 \times 20 \times T/2$	2
L3		$256 \times 10 \times T/4$	3
L4		$256 \times 5 \times T/8$	4
P2	$\begin{bmatrix}\text{Conv2D-Conv2D-BN-ReLU}\\ \text{Conv2D-Conv2D-BN-ReLU}\end{bmatrix}$	$256 \times 20 \times T/2$	2
P3		$256 \times 10 \times T/4$	3
T1		$256 \times 40 \times T$	1
T2	$\begin{bmatrix}\text{Conv2D-Conv2D-BN-ReLU}\\ \text{Conv2D-Conv2D-BN-ReLU}\end{bmatrix}$	$256 \times 20 \times T/2$	2
T3		$256 \times 10 \times T/4$	3
T4		$256 \times 5 \times T/8$	4

where T_i represents the i-th stage of the bottom-up path. Similarly, $w_{i,j}^T$ is a learnable fusion weight parameter. $I_{i,j}^T$ denotes a normalized softmax value, which ranges from 0 to 1. $Resize$ represents a resize operator.

3.2. Self-Knowledge Distillation for Speaker Verification

Traditionally, knowledge distillation utilizes a pretrained teacher network to guide the training of student network at label or embedding level as shown in Figure 1(a). By comparison, our proposed method SKDFE adopts a self-teacher network, which is trained jointly with the self-student network, to distill its own refined knowledge without a pretained model. Meanwhile, both label level and feature level distillations are performed to enhance the self-student network.

Firstly, the soft label $\tilde{y}$ of the self-teacher network is utilized to perform the label level distillation, which forces the self-student network to mimic the posteriors of the self-teacher network. The corresponding Kullback-Leibler divergence (KLD) loss $\mathcal{L}^{KLD}$ can be formulated as:

$$\mathcal{L}^{KLD} = -\sum_{i=1}^{N}\sum_{j=1}^{C} \tilde{y}_j^i \log y_j^i \tag{6}$$

where $\tilde{y}^i$ is the posteriors of the i-th sample predicted by the self-teacher network.

Secondly, the feature level distillation is employed to induce the self-student network to learn from the refined feature map T_i of the self-teacher network. Specifically, the attention transfer [20] is adopted for feature distillation in this paper. The feature level distillation loss $\mathcal{L}^F$ is defined as:

$$\mathcal{L}^F = \sum_{i=1}^{S} \left\| \phi(T_i) - \phi(F_i) \right\|_2 \tag{7}$$

where i indicates the the i-th stage feature map of the self-student network and self-teacher network. ϕ is a channel-wise pooling operator combined with L_2 normalization.

Thirdly, cross-entropy loss $\mathcal{L}^{CE}$ is applied to both the self-student network and self-teacher network, which makes them learn from the ground-truth labels.

For label level only distillation, the overall optimization objective is:

$$\mathcal{L}^{SKDFE} = \mathcal{L}_S^{CE} + \mathcal{L}_T^{CE} + \alpha\mathcal{L}^{KLD} \tag{8}$$

For feature level only distillation, the overall optimization objective is:

$$\mathcal{L}^{SKDFE} = \mathcal{L}_S^{CE} + \mathcal{L}_T^{CE} + \beta\mathcal{L}^{F} \tag{9}$$

For both label level and feature level distillations, the overall optimization objective is:

$$\mathcal{L}^{SKDFE} = \mathcal{L}_S^{CE} + \mathcal{L}_T^{CE} + \alpha\mathcal{L}^{KLD} + \beta\mathcal{L}^{F} \tag{10}$$

where α and β are hyperparameters, which are chosen from {1, 2, 3} and {100, 200} respectively.

4. EXPERIMENTAL SETUP

4.1. Datasets

Our experiments are conducted on the Voxceleb1&2 [21, 22] datasets. The development set of Voxceleb2 is adopted as training data. Voxceleb1 is used as testing data. Performance is measured on the three official trial lists. Specifically, no data augmentation is applied in the experiments for fair comparison with [10].

4.2. Implementation Details

40-dimensional Fbank with a frame length of 25 ms and a frame shift of 10ms are extracted as input features. We use a fixed frame number 300 to extract the Fbank features during training, which is randomly cropped from one utterance. Trial scores are evaluated using probabilistic linear discriminant analysis (PLDA) [23] since it provides better results than cosine distance. The equal error rate (EER) and the minimum detection cost function (MinDCF) with the settings of $P_{target} = 0.01$ and $C_{FA} = C_{Miss} = 1$ are adopted to measure performance. All the systems are implemented using PyTorch [24] framework. In order to compare with [10], normal softmax is employed to calculate the loss in the experiments.

5. RESULTS

5.1. Evaluation of the Proposed Self-Knowledge Distillation

The results of the baseline systems and our proposed self-knowledge distillation systems are listed in Table 3. We implement ResNet18, 34 and 50 as the baselines. It can be obviously observed that the performance of systems can become better with models being deeper and larger. Still, how to make small models have comparable or even better performance than large ones is important and difficult for speaker verification.

Our proposed self-knowledge distillation method can significantly improve the baselines without increasing the parameters. In our experiments, ResNet18 and ResNet34 are employed as the self-student networks individually. The channel number in the self-teacher network is set to 256 by default. For ResNet18-SKDFE, applying label level and feature level distillations simultaneously achieves the best performance, which results in relative improvements in EER by 21.5%, 17.8%, 18.9% and in MinDCF by 19.4%,

Table 3. Performance comparison of the baselines and the proposed self-knowledge distillation systems on the Voxceleb1 dataset.

Architecture	Distillation	# Params	Voxceleb-O		Voxceleb-E		Voxceleb-H	
			EER(%)	MinDCF	EER(%)	MinDCF	EER(%)	MinDCF
ResNet18	—	3.45M	2.00	0.2232	2.07	0.2453	3.65	0.3566
ResNet34	—	5.98M	1.78	0.2331	1.87	0.2195	3.22	0.3207
ResNet50	—	8.51M	1.45	0.2131	1.63	0.1903	2.93	0.2878
ResNet18-SKDFE	Label	3.45M	1.62	0.1952	1.76	0.1955	3.02	0.3052
	Feature		1.72	0.2019	1.87	0.2214	3.37	0.3311
	Label+Feature		1.57	0.1800	1.70	0.1873	2.96	0.2985
ResNet34-SKDFE	Label	5.98M	1.49	0.1738	1.65	0.1860	2.81	0.2852
	Feature		1.54	0.1989	1.70	0.1946	3.01	0.2934
	Label+Feature		1.44	0.1677	1.59	0.1789	2.76	0.2781

20.4%, 16.3% over the basic student network in the three official tasks. It is noteworthy that the best ResNet18-SKDFE can achieve better performance than ResNet34, and meanwhile obtain a substantial 45.0% reduction in the parameter size. Similarly, the best ResNet34-SKDFE decreases the EERs to 1.44%, 1.59% and 2.76% in the three official tasks, which are comparable or even better than the ResNet50 system, and on the other hand the proposed ResNet34-SKDFE even has 30.0% fewer parameters than the ResNet50.

From Table 3, we can see that both label level and feature level distillations can improve the baselines, which reveals that self-teacher network has the ability to yield more enhanced and powerful features via cross-path connections. Additionally, the best knowledge transfer can be achieved through distilling at label and feature level simultaneously, which is reasonable since different information exists in label and feature levels.

5.2. Comparison with Traditional Knowledge Distillation

We compare the traditional knowledge distillation [10] with our proposed self-knowledge distillation in this section. The pretrained ResNet34 is set as the teacher while an untrained ResNet18 is adopted as the student. Two knowledge distillation methods: label level and embedding level from [10] are implemented. Table 4 shows that our proposed SKDFE outperforms the three knowledge distillation variants from [10] significantly. Moreover, SKDFE reduces the necessity of pretraining a teacher network in advance.

We notice that only slight improvements are obtained when using ResNet34 as the teacher network. Since a stronger teacher network can boost student network further, we replace ResNet34 with ResNet50 as the teacher network. Results are listed in Table 5. It shows that SKDFE still outperforms the three knowledge distillation variants with a pretrained ResNet50 as the teacher network, which further demonstrates the superiority of our proposed self-knowledge distillation method. Joint training of the self-student network and self-teacher network can result in better knowledge transfer and a narrower performance gap between the teacher and student in self-knowledge distillation.

Table 4. The first line is the teacher network ResNet34. The middle part shows the student network ResNet18 and three traditional knowledge distillation variants. The last line is the proposed self-knowledge distillation with ResNet18 as the self-student network.

System	Distillation	Vox1-O	Vox1-E	Vox1-H
ResNet34	—	1.78	1.87	3.22
ResNet18	—	2.00	2.07	3.65
	Label	1.95	1.99	3.50
	$\text{Embedding}_{\text{MSE}}$	1.94	1.99	3.54
	$\text{Embedding}_{\text{COS}}$	1.86	1.97	3.53
ResNet18-SKDFE	Label+Feature	**1.57**	**1.70**	**2.96**

Table 5. The first line is the teacher network ResNet50. The middle part shows the student network ResNet18 and three traditional knowledge distillation variants. The last line is the proposed self-knowledge distillation with ResNet18 as the self-student network.

System	Distillation	Vox1-O	Vox1-E	Vox1-H
ResNet50	—	1.45	1.63	2.93
ResNet18	—	2.00	2.07	3.65
	Label	1.80	1.89	3.42
	$\text{Embedding}_{\text{MSE}}$	1.87	1.91	3.46
	$\text{Embedding}_{\text{COS}}$	1.78	1.88	3.45
ResNet18-SKDFE	Label+Feature	**1.57**	**1.70**	**2.96**

6. CONCLUSION

In this paper, we introduce a novel self-knowledge distillation paradigm to replace the conventional knowledge distillation for speaker verification. Two advantages can be obtained: 1) reduce the necessity of pretraining a teacher network beforehand. 2) make small models have comparable or even better performance than large ones. We propose an auxiliary feature enhancement network as self-teacher network to transfer the refined knowledge to self-student network through label level and feature level distillations. Experiments on Voxceleb dataset demonstrate the effectiveness of our proposed self-knowledge distillation via feature enhancement (SKDFE). Compared to the conventional knowledge distillation, the proposed SKDFE can achieve better knowledge transfer and improve the baseline systems significantly.

7. ACKNOWLEDGEMENT

This work was supported by the China NSFC projects (No. 62122050 and No. 62071288), and Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0102). Experiments have been carried out on the PI super-computer at Shanghai Jiao Tong University.

8. REFERENCES

[1] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4052–4056.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 999–1003.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: robust dnn embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[4] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5796–5800.

[5] D. Garcia-Romero, A. McCree, D. Snyder, and G. Sell, "Jhuhltcoe system for the voxsrc speaker recognition challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7559–7563.

[6] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[7] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 3830–3834.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[9] S. Koppula, J. Glass, and A. P. Chandrakasan, "Energy-efficient speaker identification with low-precision networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2246–2250.

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

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

[12] J. Balian, R. Tavarone, M. Poumeyrol, and A. Coucke, "Small footprint text-independent speaker verification for embedded systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6164–6168.

[13] T. Zhu, X. Qin, and M. Li, "Binary neural network for speaker verification," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021, pp. 86–90.

[14] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[15] L. Zhang, Z. Chen, and Y. Qian, "Knowledge distillation from multi-modality to single-modality for person verification," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021, pp. 1897–1901.

[16] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: improve the performance of convolutional neural networks via self distillation," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 3713–3722.

[17] X. Lan, X. Zhu, and S. Gong, "Knowledge distillation by on-the-fly native ensemble," in *Advances in Neural Information Processing Systems (NIPS)*, 2018, pp. 7517–3727.

[18] M. Ji, S. Shin, S. Hwang, G. Park, and I. Moon, "Refine myself by teaching myself: feature refinement via self-knowledge distillation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10664–10673.

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[20] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1605.07146*, 2016.

[21] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2616–2620.

[22] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: deep speaker recognition," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 1086–1090.

[23] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision (ECCV)*, 2006, pp. 531–542.

[24] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Advances in Neural Information Processing Systems (NIPS) Autodiff Workshop*, 2017.