KNOWLEDGE DISTILLATION FOR SMALL FOOT-PRINT DEEP SPEAKER EMBEDDING

Shuai Wang, Yexin Yang, Tianzhe Wang, Yanmin Qian, Kai Yu

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{feixiang121976, yangyexin, usedtobe, yanminqian, kai.yu}@sjtu.edu.cn

ABSTRACT

Deep speaker embedding learning is an effective method for speaker identity modelling. Very deep models such as ResNet can achieve remarkable results but are usually too computationally expensive for real applications with limited resources. On the other hand, simply reducing model size is likely to result in significant performance degradation. In this paper, label-level and embedding-level knowledge distillation are proposed to narrow down the performance gap between large and small models. Label-level distillation utilizes the posteriors obtained by a well-trained teacher model to guide the optimization of the student model, while embedding-level distillation directly constrains the similarity between embeddings learned by two models. Experiments were carried out on the VoxCeleb1 dataset. Results show that the proposed knowledge distillation methods can significantly boost the performance of highly compact student models.

***Index Terms*—** knowledge distillation, teacher-student learning, speaker verification, speaker embedding

1. INTRODUCTION

Recently, speaker embeddings learned by deep architectures have shown impressive performance for speaker recognition. Speaker embeddings denote fixed-dimensional vector-based representations for modelling speakers' identities. From Gaussian Mixture Model (GMM) based super-vector [1, 2], joint factor analysis (JFA) based eigen-voice vectors[3], factor analysis (FA) based *i*-vectors[4], to the recently arising deep speaker embeddings [5, 6, 7, 8, 9, 10], speaker embedding learning has been a mainstream for speaker modelling in speaker recognition now.

Speaker embeddings learned with very deep architectures such as ResNet[11] are proven to achieve a very good performance [12, 13, 14]. However, these models comprise of millions of parameters and demand tremendous memory and computation resources. For real applications which usually need to run the program on a resource-constrained embedded device, these advanced models cannot be deployed easily. On the other hand, small models need much less resources and are more suitable for deployment, but at the expense of performance degradation. Accordingly, we want to develop an effective mechanism to boost the system performance of small models. To reduce the performance gap compared to the large deep models, knowledge distillation will be a natural approach.

Knowledge distillation was proposed in [15] and has been successfully applied to many applications such as image recognition[16], speech recognition [17, 18, 19] and keyword spotting [20]. Knowledge distillation is often used for domain adaptation and model compression, the common method is to use the posteriors obtained via a well-trained teacher model to guide the optimization of the student model, this paradigm is often referred to as teacher-student learning. In this paper, we propose to introduce the teacher-student learning idea into the deep speaker embedding learning process. Two knowledge distillation methods are developed.

- Label level knowledge distillation: The teacher model provides the predicted posteriors as the reference label for the student model. The Kullback-Leibler divergence is used to supervise the model optimization.
- Embedding level knowledge distillation: Directly use the speaker embeddings learned by the teacher model to help the optimization of the student model. More specifically, similarity metrics such as minimum square error and cosine distance are used to constrain the similarity of embeddings learned from two models.

The remainder of this paper is organized as follows. Section 2 gives a brief introduction to deep speaker embedding learning. The proposed label-level and embedding-level knowledge distillation are introduced for speaker recognition in Section 3. Detailed experimental setups and result analysis will be given in Section 4. Section 5 concludes the whole paper.

2. DEEP SPEAKER EMBEDDING LEARNING

In the deep speaker embedding framework, a speaker discriminative DNN is first trained on the utterances from a large set of speakers. This training process can be performed at the frame level [21, 22] or utterance level [6, 7, 23], while the utterance-level training makes more sense and achieves a better performance. More powerful deep architectures such as ResNet and more advanced loss functions such as triplet loss [6, 9], angular softmax [10] and generalized end-to-end loss [24] have been developed, achieving impressive results on the standard datasets. In this work, we adopt the normal softmax combined with cross entropy loss as the training criterion, more complex frameworks will be left in future work.

†Yanmin Qian and Kai Yu are the corresponding authors

This work has been supported by the National Key Research and Development Program of China under Grant No.2017YFB1302402 and the China NSFC projects (No. 61603252 and No. U1736202) and the Shanghai sailing program (No.16YF1405300). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

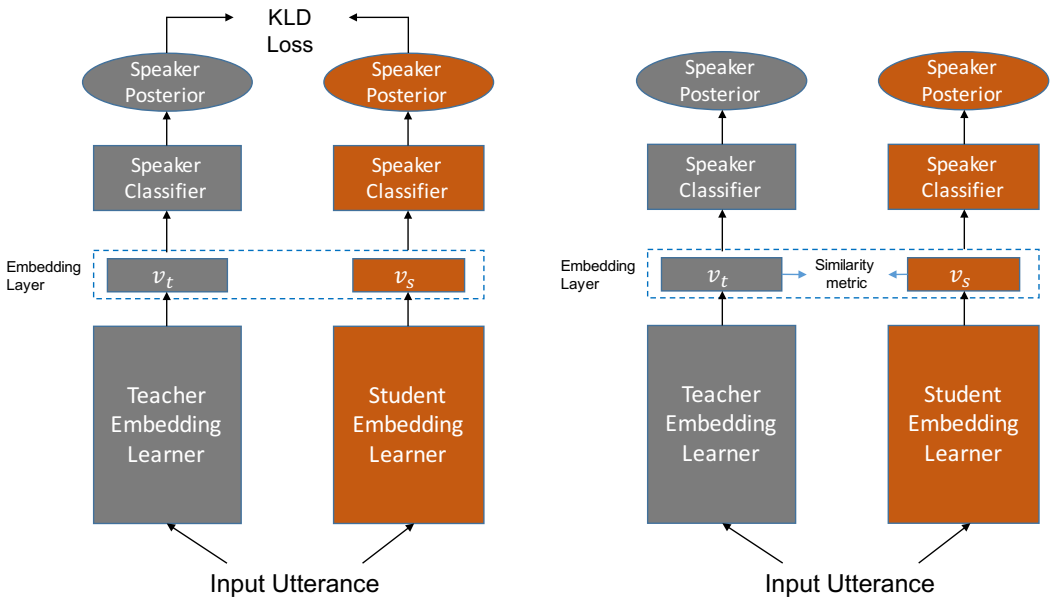

Fig. 1. Knowledge distillation for deep speaker embedding learning in speaker recognition. (1) Left: Label-level teacher-student learning architecture, the student optimization is guided by the posteriors predicted by a pretrained teacher model . (2) Right: Embedding-level teacher-student learning system, directly constraining the similarity of speaker embeddings learned from the teacher and student model.

3. THE TEACHER-STUDENT LEARNING FOR DEEP SPEAKER EMBEDDING

Teacher-student learning uses a well-performing teacher model to help the optimization of a student model. For instance, researchers in [18] use the ensembles of several acoustic models to help optimizing a single acoustic model for speech recognition. Similar to [17, 25] for speech recognition, we use teacher-student learning to reduce the performance gap between large deep models and small-footprint models for speaker recognition. In this paper, two frameworks are proposed for the knowledge distillation between deep speaker embeddings, including the label-level and embedding-level, which will be described in the following sections. The two different architectures are illustrated in Figure 1.

3.1. Cross-entropy training

The most common criterion for speaker embedding learning is the cross-entropy (CE), which is defined as following,

$$\mathcal{L}^{\text{CE}} = -\sum_{i=1}^{N}\sum_{j=1}^{C} \hat{y}_j^i \log y_j^i \tag{1}$$

where i is the sample index, N denotes the number of samples. $\hat{\boldsymbol{y}}^i$ represents the ground truth label which is a one-hot vector, $\boldsymbol{y}^i$ is the predicted output from the model. j denotes the j-th class, C denotes the number of classes.

3.2. Label-level knowledge distillation

In the speaker embedding learning task, the outputs of the teacher and student models are both posteriors over the same set of speakers and the student model is expected to mimic the teacher model if we force them to emit similar posteriors. This is usually achieved by minimizing the Kullback-Leibler divergence (KLD) between the student and teacher distributions [17, 25]. The corresponding KLD loss is defined in Equation 2

$$\mathcal{L}^{\text{KLD}} = -\sum_{i=1}^{N}\sum_{j=1}^{C} \tilde{y}_j^i \log y_j^i \tag{2}$$

where $\tilde{\boldsymbol{y}}^i$ is the posteriors of the i-th sample predicted by the teacher model, it's now a distribution (soft labels) rather than a simple one-hot vector (hard labels). Compared to the hard labels, soft labels contain more information (referred to as dark knowledge in [15]) of the underlying label distribution which may benefit the optimization of the student model. For simplicity, the "temperature term" in [15] is neglected in our experimental settings. In the optimization, both hard labels and soft labels are used, so the two losses can be combined for the student model training as

$$\mathcal{L} = \mathcal{L}^{\text{CE}} + \alpha\mathcal{L}^{\text{KLD}} \tag{3}$$

where α is a hyper-parameter to balance two losses.

3.3. Embedding-level knowledge distillation

Instead of performing the knowledge distillation at the label level, i.e. the distribution of model outputs, it's more intuitive to directly constrain the similarity of learned embeddings from two models in the deep embedding based speaker recognition framework. In this work, minimum square error (MSE) and cosine distance (COS) loss are developed as the optimization metric for embedding-level knowledge distillation.

$$\mathcal{L}^{\text{MSE}} = \sum_{i=1}^{N} \|\mathbf{v}_t^i - \mathbf{v}_s^i\|^2 \tag{4}$$

$$\mathcal{L}^{\text{COS}} = -\sum_{i=1}^{N} \frac{\mathbf{v}_t^i \cdot \mathbf{v}_s^i}{\|\mathbf{v}_t^i\| \|\mathbf{v}_s^i\|} \tag{5}$$

where $\mathbf{v}_t^i$ represents the embedding computed by the teacher model for the i-th sample, $\mathbf{v}_s^i$ denotes the embedding computed by the student model. The final loss function for model training is $\mathcal{L}^{\text{CE}} + \beta\mathcal{L}^{\text{MSE}}$ or $\mathcal{L}^{\text{CE}} + \gamma\mathcal{L}^{\text{COS}}$, respectively. β and γ are the corresponding weighting parameters.

4. EXPERIMENTS

4.1. Dataset

All the experiments were carried out on the VoxCeleb1[26] dataset which was recently released by Oxford. VoxCeleb is a large scale text-independent speaker recognition dataset comprised of two releases, VoxCeleb1 [26] and VoxCeleb2 [13]. Note that we only use VoxCeleb1 in this paper. Moreover, no data augmentation method was adopted in the experiments. Part 1 contains over 150000 utterances from 1251 different celebrities. For the speaker verification task, part 1 was split into the training part and the evaluation part. The training part contains 148642 utterances from 1211 celebrities, while the evaluation set contains about 4874 utterances from the rest 40 celebrities. The standard trial list for the verification contains 37720 pairs.

4.2. System setups and evaluation metric

The proposed knowledge distillation methods can be applied to standard speaker embedding learning models. In this work, we adopt a similar architecture as described in [14] for the teacher model, since a good performance was reported using this architecture on the VoxCeleb dataset. It's a 34-layer neural network comprising of 16 residual blocks ($\{3, 4, 6, 3\}$ [11]. The implementation of the ResNet architecture follows the standard one as depicted in [11]. The detailed network configuration of ResNet34 is shown in Table 1.

Table 1. The detailed configuration of the ResNet34 teacher model: all filter sizes are set to 3×3, N denotes the frame number of the input utterance.

Layers	Output Size	Channels	Blocks
Conv Layers	$64 \times N$	16	-
Res1	$64 \times N$	16	3
Res2	$32 \times N/2$	32	4
Res3	$16 \times N/4$	64	6
Res4	$8 \times N/8$	128	3
Reshape & Average	128×1	-	-
FC Layer (embedding)	128×1	-	-
Output	#speakers	-	-

For the student model, several different setups are investigated in the experiments. The most intuitive choice is to use a ResNet with less blocks. Two setups are adopted, namely ResNet16 and ResNet10, with block number of residual blocks set as $\{1, 2, 3, 1\}$ and $\{1, 1, 1, 1\}$, respectively. ResNet16 is around the half size of ResNet34, while ResNet10 is the smallest model we can obtain while keeping the same architecture with the ResNet34 teacher model. Moreover, a different architecture was also investigated, which is a simple 4-layer CNN with detailed configuration as shown in Table 2. The CNN model is designed to mimic the ResNet architecture, while each residual block is replaced with a simple convolutional layer. A comparison of different models in terms of parameter number and inference speed will be given in Section 4.4.

Table 2. The detailed configuration of the CNN student model, all filter sizes are set to 3×3, N denotes the frame number of the input utterance.

Layers	Output Size	Channels
Conv Layers	Output size	Channels
Conv1	$64 \times N$	16
Conv2	$32 \times N/2$	32
Conv3	$16 \times N/4$	64
Conv4	$8 \times N/8$	128
Reshape & Average	128×1	-
FC Layer (embedding)	128×1	-
Output	#speakers	-

For all neural network based systems, 64-dimensional Fbank features extracted with a frame-length of 25 ms are extracted at a 10 ms frame shift. Neural networks are trained with a mini-batch of 64 on a single GPU, stochastic gradient descent with momentum 0.9 and weight decay $1e-4$ is used in the optimizer. Although the original lengths of training utterances vary, we keep samples in one mini-batch sharing the same frame number, which is a random integer between 300 and 800. In the experiments, three hyper-parameters α, β and γ mentioned in Section 3 are set as 1.0, 0.4 and 0.4, respectively, which achieve the best results in the experiments.

Speaker embeddings are evaluated using both probabilistic linear discriminant analysis (PLDA) and cosine distance. All results are reported in terms of equal error rate (EER) and minimum of the normalized detection cost function, with the prior target probability P_{target} set as 0.01 ($\text{minDCF}_{0.01}$) and 0.001 ($\text{minDCF}_{0.001}$), and equal weights of 1.0 between misses C_{miss} and false alarms C_{fa}.

4.3. Results and analysis

Results of different systems are summarized in Table 3. ResNet34 is the teacher model, while ResNet16, ResNet10 and CNN with no knowledge distillation are three student model baselines. As shown in Table 3, a deeper architecture achieves a better performance. The ResNet34 teacher model achieves 4.852% and 6.045% EERs with PLDA and Cosine Distance Scoring, respectively, which is comparable to the results on the same dataset in the literature [14, 26, 27].

Different extents of performance degradation are observed with different student models, ResNet16, ResNet10 and CNN achieve EERs of 5.456%, 6.384% and 8.823% using PLDA backends, respectively. Label-level knowledge distillation reduces the EERs of three systems to 5.392%, 5.870% and 7.853%, while the embedding-level knowledge distillation further boosts the performance. From Table 3, it's observed that embedding-level knowledge distillation methods outperform the label-level one, which makes sense since the goal we are optimizing now is more relevant to the system performance. Cosine distance based distillation achieves a better performance than MSE, the reason could be that MSE constraint is too stringent, which harms the generalization ability.

It's noticeable that the performance of the ResNet16 with knowledge distillation using $\text{Embedding}_{\text{COS}}$ can almost obtain the same accuracy as the teacher model ResNet34, but with much less parameters. For the simplest CNN student model, the ability of the proposed knowledge distillation methods could be better reflected. A relative

Table 3. Performance comparison of different systems. The first line represents the teacher model ResNet34 and the following lines denote three student models, including ResNet16, ResNet10 and simple CNN as described in Section 4.2. Label, $\text{Embedding}_{\text{MSE}}$ and $\text{Embedding}_{\text{COS}}$ denote different knowledge distillation methods described in Section 3.

System	Distillation	PLDA Scoring			Cosine Scoring		
		EER (%)	$\text{minDCF}_{0.01}$	$\text{minDCF}_{0.001}$	EER (%)	$\text{minDCF}_{0.01}$	$\text{minDCF}_{0.001}$
ResNet34	-	4.852	0.5161	0.7268	6.045	0.5342	0.6422
ResNet16	-	5.456	0.5739	0.7364	6.591	0.5969	0.7325
	Label	5.392	0.5312	0.6613	**6.230**	0.5694	**0.6270**
	$\text{Embedding}_{\text{MSE}}$	5.154	0.5128	0.7080	6.479	**0.5256**	0.6745
	$\text{Embedding}_{\text{COS}}$	**4.857**	**0.5115**	**0.6700**	6.410	0.5705	0.7048
ResNet10	-	6.384	0.6354	0.7542	8.542	0.6971	0.7896
	Label	5.870	0.5603	0.7179	7.322	**0.5897**	0.7628
	$\text{Embedding}_{\text{MSE}}$	5.604	0.5696	0.7645	**7.200**	0.6159	**0.7105**
	$\text{Embedding}_{\text{COS}}$	**5.472**	**0.5309**	**0.7808**	7.312	0.6290	0.7639
CNN	-	8.823	0.6923	**0.7271**	23.26	0.8266	0.8883
	Label	7.853	**0.6262**	0.7628	14.59	0.7585	0.8979
	$\text{Embedding}_{\text{MSE}}$	7.794	0.6542	0.7369	11.30	**0.7141**	0.8372
	$\text{Embedding}_{\text{COS}}$	**6.914**	0.6706	0.7615	**9.464**	0.7169	**0.8080**

21.6% and 59.3% EER reduction is achieved using $\text{Embedding}_{\text{COS}}$ distillation in terms of PLDA and Cosine Scoring, respectively.

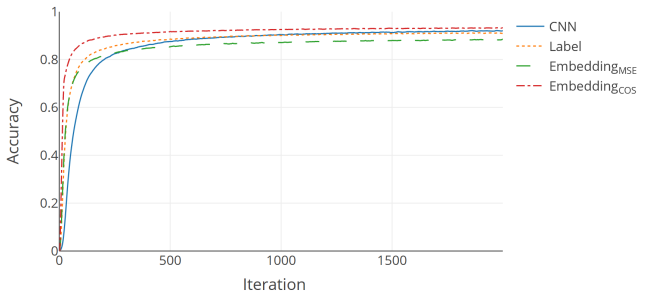

Fig. 2. Convergence comparison of student CNN model w/ or w/o knowledge distillation.

The convergence speeds of the student CNN model with/without knowledge distillation are depicted in Figure 2. It could be found that the convergence speed is improved to different extents with different knowledge distillation methods. One interesting observation is that the final accuracy achieved by the MSE loss distillation is even lower than the original CNN, but the former system outperforms the latter. Recall the limited performance gain obtained by label-level knowledge distillation, both observation exhibits that the softmax with cross entropy loss is not a perfect criterion for speaker embedding learning. More powerful criterion such as angular-softmax and end-to-end loss could be considered [6, 9, 14, 24], and knowledge distillation with these settings will be left as the future work.

4.4. Model size and inference speed

Excluding the last layers which will not be used in the system implementation, the mode size and inference speed are tested and compared, and the results are shown in Table 4.

Reducing the models size will increase the inference speed accordingly. Recall the performance reported in Table 3, ResNet16 obtains nearly the same performance with the teacher model ResNet34,

Table 4. Comparison on model sizes and inference speeds between the teacher and student models. Inference speeds are tested on both GPU (Tesla K40m) and CPU (Intel Xeon E5-2670)

Model	# Parameters	CPU time (ms)	GPU time (ms)
ResNet34	1.35M	365.7	12.77
ResNet16	0.49M	157.5	5.816
ResNet10	0.32M	98.81	4.850
CNN	0.11M	33.84	1.795

but with only half of the parameters and inference time. Another observation is that although the performance gap between the teacher and student model can be reduced with the proposed knowledge distillation methods, a larger model still gets better performance. For real applications, a trade-off is still considered between the model size and performance, while such a trade-off can be achieved more easily using the proposed knowledge distillation methods.

5. CONCLUSION

Speaker embeddings learned by very deep architectures have exhibited impressive performance on speaker recognition, however, these advanced deep models are not suitable for deployment. In this paper, we propose to use knowledge distillation with teacher-student learning framework to bridge the performance gap between speaker embeddings extracted by large and small models. Two knowledge distillation architectures are proposed: 1) Label-level knowledge distillation, in which the posterior outputs of the teacher model is used to guide the optimization of the student model. 2) Embedding-level knowledge distillation, in which the similarity between embeddings from teacher and student models is constrained. Experiments are carried out on the VoxCeleb1 dataset, a standard 34-layer ResNet is used as the teacher model, while three different models with different sizes are used as the student models. Results consistently show that the performance of the small student model can be boosted significantly by the proposed knowledge distillation methods.

6. REFERENCES

[1] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] William M Campbell, Douglas E Sturim, Douglas A Reynolds, and Alex Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2006.* IEEE, 2006, vol. 1, pp. I–I.

[3] Patrick Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.

[4] 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, 2011.

[5] Shuai Wang, Zili Huang, Yanmin Qian, and Kai Yu, "Deep discriminant analysis for i-vector based robust speaker recognition," *arXiv preprint arXiv:1805.01344*, 2018.

[6] Chunlei Zhang and Kazuhito Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," *Proc. Interspeech 2017*, pp. 1487–1491, 2017.

[7] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *ICASSP, Calgary*, 2018.

[8] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *Proc. Interspeech 2017*, pp. 999–1003, 2017.

[9] Zili Huang, Shuai Wang, and Yanmin Qian, "Joint i-vector with end-to-end system for short duration text-independent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018*. IEEE, 2018.

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

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] Na Li, Deyi Tuo, Dan Su, Zhifeng Li, and Dong Yu, "Deep discriminative embeddings for duration robust speaker verification," in *Proc. Interspeech 2018*, 2018, pp. 2262–2266.

[13] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[14] Weicheng Cai, Jinkun Chen, and Ming Li, "Analysis of length normalization in end-to-end speaker verification system," *arXiv preprint arXiv:1806.03209*, 2018.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[16] Bharat Bhusan Sau and Vineeth N Balasubramanian, "Deep model compression: Distilling knowledge from noisy teachers," *arXiv preprint arXiv:1610.09650*, 2016.

[17] Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong, "Learning small-size dnn with output-distribution-based criteria," in *Fifteenth annual conference of the international speech communication association*, 2014.

[18] Yevgen Chebotar and Austin Waters, "Distilling knowledge from ensembles of neural networks for speech recognition.," in *Interspeech*, 2016, pp. 3439–3443.

[19] Mingkun Huang, Yongbin You, Zhehuai Chen, Yanmin Qian, and Kai Yu, "Knowledge distillation for sequence model," in *Proc. Interspeech 2018*, 2018, pp. 3703–3707.

[20] Jinyu Li, Rui Zhao, Zhuo Chen, Changliang Liu, Xiong Xiao, Guoli Ye, and Yifan Gong, "Developing far-field speaker system via teacher-student learning," *arXiv preprint arXiv:1804.05166*, 2018.

[21] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*. IEEE, 2014, pp. 4052–4056.

[22] Nanxin Chen, Yanmin Qian, and Kai Yu, "Multi-task learning for text-dependent speaker verification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[23] Yao Tian, Meng Cai, Liang He, and Jia Liu, "Investigation of bottleneck features and multilingual deep neural networks for speaker verification.," in *Interspeech*, 2015, pp. 1151–1155.

[24] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," *arXiv preprint arXiv:1710.10467*, 2017.

[25] Liang Lu, Michelle Guo, and Steve Renals, "Knowledge distillation for small-footprint highway networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4820–4824.

[26] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[27] Suwon Shon, Hao Tang, and James Glass, "Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model," *arXiv preprint arXiv:1809.04437*, 2018.