*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Open-set Short Utterance Forensic Speaker Verification using Teacher-Student Network with Explicit Inductive Bias

*Mufan Sang, Wei Xia, John H.L. Hansen*

Center for Robust Speech Systems, University of Texas at Dallas, TX 75080
{mufan.sang, wei.xia, john.hansen}@utdallas.edu

## Abstract

In forensic applications, it is very common that only small naturalistic datasets consisting of short utterances in complex or unknown acoustic environments are available. In this study, we propose a pipeline solution to improve speaker verification on a small actual forensic field dataset. By leveraging large-scale out-of-domain datasets, a knowledge distillation based objective function is proposed for teacher-student learning, which is applied for short utterance forensic speaker verification. The objective function collectively considers speaker classification loss, Kullback-Leibler divergence, and similarity of embeddings. In order to advance the trained deep speaker embedding network to be robust for a small target dataset, we introduce a novel strategy to fine-tune the pre-trained student model towards a forensic target domain by utilizing the model as a fine-tuning start point and a reference in regularization. The proposed approaches are evaluated on the 1$^{st}$48-UTD forensic corpus, a newly established naturalistic dataset of actual homicide investigations consisting of short utterances recorded in uncontrolled conditions. We show that the proposed objective function can efficiently improve the performance of teacher-student learning on short utterances and that our fine-tuning strategy outperforms the commonly used weight decay method by providing an explicit inductive bias towards the pre-trained model.
**Index Terms**: Text-independent speaker verification, short utterance, teacher-student learning, transfer learning

## 1. Introduction

Speaker verification (SV) is defined as the process of identifying the true characteristics of the speaker and to accept or discard the identity claimed by the speaker. In recent years, speaker verification has drawn significantly improvement with the fast development of deep learning and the great success of various sophisticated neural networks applied to deep speaker embedding systems [1, 2, 3, 4, 5, 6]. Generally, speaker embedding systems consist of a front-end frame-level feature extractor, an utterance-level encoding layer, one or more fully-connected, and an output classifier. Like the I-vector system [7], speaker embedding networks can encode variable length of utterance into a fix-length speaker representation. Although studies [2, 4] have shown that deep speaker embedding systems outperformed the traditional I-vector especially on short utterance, short utterance speaker verification is still a challenging task because of insufficient phonetic information provided [8].

Recently, deep neural network (DNN) and convolutional neural network (CNN) based speaker embedding systems are applied to solve this problem and obtain great performance improvements [9, 10, 11, 12]. In [10], a raw waveform CNN-LSTM architecture is proposed to extract phonetic-level features which can help to compensate for missing phonetic information. In [9, 11], Some new aggregation methods such as NetVLAD layer and Time-Distributed Voting (TDV) are designed to improve the efficiency of the aggregation process. In order to keep a strong speaker discrimination, most of studies use large amounts of in-domain data to train the speaker embedding networks for short utterances and evaluate them in the same domain. However, it usually happens that only a very small size of dataset is available for the specific target domain, and directly training the model from scratch on it will greatly degrade the performance. As one of the application areas of speaker verification, forensic related problems are complex and challenging because it is hardly to get sufficient amount of data and length of utterances even for enrollment [13, 14]. Besides, inconsistencies usually exist in forensic relevant datasets, such as unknown recording locations, noise, reverberation, very limited duration of useful human speech, and so on [15, 16]. All of these variants can lead to significant performance degradation of speaker verification systems.

In this work, we collect a new challenging naturalistic dataset for forensic analysis of audio from various cities across the USA where detectives were investigating actual homicides. We focus on the problem of short utterance speaker verification in forensic domain with only a small naturalistic target dataset available. In order to alleviate this problem, firstly, we propose a knowledge distillation [17] based objective function which collectively considers Kullback-Leibler divergence (KLD) of posteriors, similarity of embeddings, and speaker classification loss to improve the performance of short utterance speaker verification with the teacher-student (T-S) learning applied. Secondly, to avoid losing initial speaker discriminative knowledge learned from a large number of speakers in out-of-domain datasets, we introduce a new fine-tuning strategy that can help to encode an explicit inductive bias towards the pre-trained model by using the pre-trained network as a fine-tuning start point as well as a reference penalized in the regularization. In this way, the fine-tuned model takes advantage of the strong speaker discriminative power on short utterances and learns new features from the small target dataset.

We evaluate the performance of our fine-tuning strategy based on the $L^2$ regularization (known as weight decay), which often results in a suboptimal solution for the target domain with an implicit inductive bias towards the pre-trained model. In this paper, we show that the proposed knowledge distillation loss function is able to transfer the knowledge from teacher to student more effectively, and the introduced fine-tuning strategy outperforms the $L^2$ regularization by adequately preserving the learned features from large-scale source datasets.

In Section 2, we describe the proposed pipeline, including the teacher-student learning framework used for short utterance speaker verification and the proposed fine-tuning strategy. Data description, experiment setting, and results analysis are reported in Section 3 and 4. Finally, we conclude this paper in Section 5.



http://dx.doi.org/10.21437/Interspeech.2020-2868

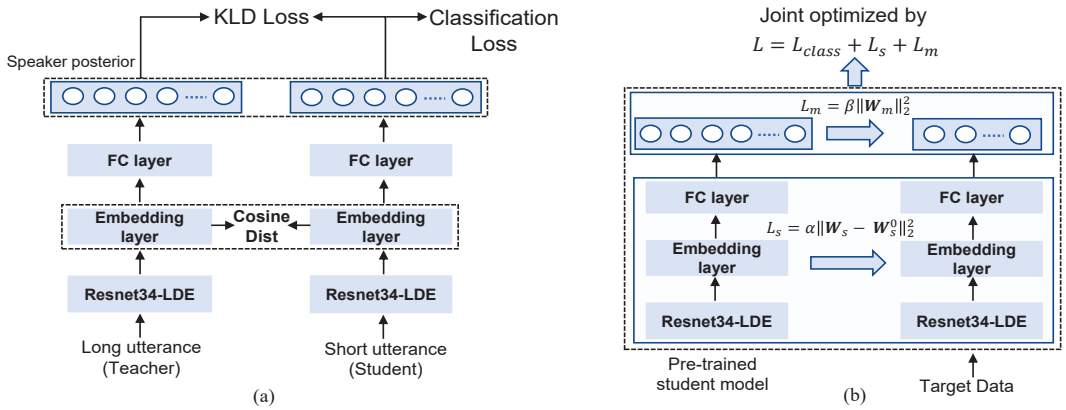


Figure 1: *(a) Flow-diagram of the teacher-student learning framework. The KL-divergence loss between posteriors, cosine distance between embeddings, and speaker classification loss are shown in the figure. (b) The fine-tuning strategy transfers the pre-trained student model to the small target dataset*

## 2. Methodology

In this section, we introduce our proposed approaches for short utterance speaker verification on the small forensic target dataset, where utterances in the target domain have much shorter length and several variation factors compared with the source domain. Firstly, the speaker embedding model should learn a strong speaker discriminative power on short utterances from large out-of-domain datasets. Then, it can be refined to perform well on short utterances in the target domain with only a much smaller in-domain dataset available. We utilize the T-S learning with the proposed objective function to improve system performance on short utterances. Then, we use our fine-tuning strategy to adapt the trained student model from the large source dataset to the small target dataset with an explicit inductive bias. The architecture of the whole pipeline is illustrated in Figure 1.

### 2.1. Teacher-Student Learning for short utterances

The teacher model is trained independently, and the student model is learned to mimic the probability distribution of teacher's output and account for speaker classification loss. Our teacher-student learning framework is described in the following sections.

#### *2.1.1. Speaker classification loss*

For both teacher and student models, softmax cross-entropy loss is used as the speaker classification loss. We also explore the angular softmax (A-softmax) [18] loss which has a stronger discriminative ability by maximizing the angular margin between speaker embeddings.

#### *2.1.2. Student model training loss*

In conventional T-S learning, the student is learned to mimic the teacher by minimizing the Kullback-Leibler divergence between the teacher and student output distributions given parallel data [19]. When adopting T-S learning for short utterance problem, the KLD function can be presented as,

$$L_{\text{KLD}} = -\sum_{i=1}^{I}\sum_{n=1}^{N} P\left(y_n|\mathbf{x}_{i,T}\right)\log\left(P\left(y_n|\mathbf{x}_{i,S}\right)\right) \quad (1)$$

Where $y_n$ refers to the speaker $n$ and $\mathbf{x}_{i,T}$ and $\mathbf{x}_{i,S}$ refer to the $i$-th input sample of the teacher and student models. In this work, the teacher model is fed with long utterance samples, and the student is fed with the shorter crop of same utterances. $P\left(y_n|\mathbf{x}_{i,T}\right)$ and $P\left(y_n|\mathbf{x}_{i,S}\right)$ are the posteriors of $i$-th sample predicted by the teacher and student model. In this way, the student model is enforced to generate similar posteriors as teacher model.

As the representation of speakers, embedding is the key to a speaker verification system. The problem will be tackled if a model can produce more similar or even identical embeddings for short utterances and long utterances. According to this idea, directly constricting the similarity of embeddings between short and long utterances would be more efficient [10, 20]. It can be achieved by minimizing the distance between the embeddings of short and long utterances. We use cosine distance (COS) as the distance metric, the corresponding loss function is formulated as,

$$L_{\text{EMD}} = -\sum_{i=1}^{1}\frac{\epsilon_T^i \cdot \epsilon_S^i}{\|\epsilon_T^i\|\,\|\epsilon_S^i\|} \quad (2)$$

Where $\epsilon_T^i$ and $\epsilon_S^i$ represent the embeddings predicted by the teacher and student models for the $i$-th sample.

In optimization, We use a multi-task objective function to combine the losses introduced above. $L_{\text{class}}$ denotes the speaker classification loss. Three different combinations of objective functions are described below:

(1)$L_{\text{class}} + L_{\text{KLD}}$: On the basis of speaker classification loss, adding soft labels enables the student model to enhance its discriminative power toward that of the teacher model trained on long utterances

(2)$L_{\text{class}} + L_{\text{EMD}}$: Replacing the KLD loss with the embedding-based loss enables the student model to generate more similar embeddings to that of teacher.

(3)$L_{\text{class}} + L_{\text{KLD}} + L_{\text{EMD}}$: Collectively combining them can guarantee both the speaker discriminative power and the similarity of embeddings between short and long utterances.

We compare the efficacy of trained student networks with the baseline models trained without T-S learning. We also compare the performances of student models optimized by different objective functions mentioned above.

### 2.2. Explicit Inductive bias for fine-tuning

Considering the limited source, naturalistic feature, and several inter-speaker and intra-speaker variations contained in the short utterance forensic corpus [21] , directly training the deep speaker embedding model from scratch cannot ensure the discriminative power of speakers, and will also cause severe overfitting problem. Fine-tuning with the commonly used $L^2$ reg-

ularization can help to alleviate overfitting to some extent, but it can only provide an implicit inductive bias towards the pre-trained model. Therefore, we introduce a novel fine-tuning strategy that is able to produce the explicit inductive bias towards the pre-trained model by setting the model as the start point of the fine-tuning as well as a reference for the regularization. With the restriction of the reference, the capacity of the fine-tuned model will not be adapted blindly. Accordingly, we investigate its efficiency on different regularizers during fine-tuning and add the suffix '*-SP*' to the regularizers. Assume $\boldsymbol{W} \in R^n$ represents the parameter matrix containing all adapted parameters of the fine-tuned model. Investigated regularizers are described as below:

$\boldsymbol{L^2}$*-norm*. It is the most common penalty term used for regularization, also called weight decay. The penalty term is shown as:

$$\Theta_2(\boldsymbol{W}) = \alpha \left\| \boldsymbol{W} \right\|_2^2, \tag{3}$$

where $\alpha$ is the weight of the penalty.

$\boldsymbol{L^2}$*-SP*. Using $\boldsymbol{W}^0$ denotes the parameter matrix of the pre-trained student model which is trained on the out-of-domain datasets. This pre-trained model is setting as the start point and the reference. Thus, we penalize the $L^2$ distance between adapted parameter matrix $\boldsymbol{W}$ and $\boldsymbol{W}^0$ to get:

$$\Theta_{\textit{2-SP}}(\boldsymbol{W}) = \alpha \left\| \boldsymbol{W} - \boldsymbol{W}^0 \right\|_2^2 \tag{4}$$

Considering the network architecture is usually changed when adapting the model from source datasets to a new target dataset, the penalty can be separated as two parts: one penalizes the part of unchanged architecture $\boldsymbol{W}_s$ between the pre-trained and fine-tuned networks, one penalizes the modified architecture $\boldsymbol{W}_m$. As shown in Figure 1-(b), they are represented as $L_s$ and $L_m$. Then we can get:

$$\hat{\Theta}_{\textit{2-SP}}(\boldsymbol{W}) = \alpha \left\| \boldsymbol{W}_s - \boldsymbol{W}_s^0 \right\|_2^2 + \beta \left\| \boldsymbol{W}_m \right\|_2^2 \tag{5}$$

$\boldsymbol{L^1}$*-SP*. Changing the $L^2$-norm to $L^1$-norm, we have:

$$\hat{\Theta}_{\textit{1-SP}}(\boldsymbol{W}) = \alpha \left\| \boldsymbol{W}_s - \boldsymbol{W}_s^0 \right\|_1 + \beta \left\| \boldsymbol{W}_m \right\|_2^2 \tag{6}$$

The $L^1$ penalty encourages some parts of the parameters to be equal to corresponding parts of the pre-trained model. Consequently, $L^1$*-SP* can be considered as a trade-off between $L^2$*-SP* and the scheme by freezing some parts of the pre-trained network.

In the experiments, We compare the investigated regularizers based on their performances on the small target dataset. Besides, we also explore the fine-tuning performance with different network layer selections.

# 3. Experiments

## 3.1. Dataset

### *3.1.1. In-domain target forensic dataset*

In this work, we collect a new challenging naturalistic forensic relevant dataset, called the 1$^{\text{st}}$48-UTD forensic corpus, and we will release this corpus with a license in the future. The corpus is intended for forensic analysis of audio from various cities across the USA where detectives were investigating actual homicides. The audio content was extracted from the USA TV program called "The First 48", and all audio contents are recorded from various real locations (e.g., interview rooms, cars, fields). We process the raw data with the following steps: (1) extract audio at 16kHz sample rate; (2) using human manual annotation, perform diarization on the audio stream (tag speaker identity); (3) tag audio segments based on neutral-vs-stress speaker state;(4) perform automatic segmentation following by manual check to verify segment boundaries.

This corpus has 49 episodes, with 300 speakers, and 5041 utterances turns consisting of 3.5 hours of actual situational crime audio. In this corpus, each episode contains disjoint speakers, and speakers in every episode are tagged as Detective, Witness, and Suspect according to their identifies. It is a small domain-specific dataset with short utterances. It involves utterances with an average length of 2.4 s and more than 50% of them are shorter than 2 seconds. Besides the length issue, context music, bleep used for covering bad words, modified speech sound, and some voice-over are also contained in the audio. In this study, we use the training portion of the 1$^{\text{st}}$48-UTD corpus to fine-tune the trained student model obtained from Section 2.1, and evaluate its performance on the test portion. After filtering the utterances consisting of non-speech content and the speakers with fewer than three utterances, the training set consists of 3755 utterances from 228 speakers, and the test set contains 882 utterances from 39 speakers.

### *3.1.2. Out-of-domain Voxceleb dataset*

For all the experiments in Section 2.1, we use Voxceleb1&2 datasets which are collected "in the wild" [22, 23]. The former contains 352 hours audio from 1251 speakers, and the latter consist of 2442 hours audio from 5994 speakers. We train the teacher and student speaker embedding networks on the whole Voxceleb2 dataset. A similar data augmentation method in [5] is adopted in the experiments.

## 3.2. Experiment settings

**T-S learning.** For both teacher and student models, 30-dimensional log-Mel filter-banks is extracted with a frame-length of 25 ms at a 10 ms shift. Mean-normalization is applied over a sliding window of up to 3 seconds, and Kaldi energy-based VAD is used to remove silence frames. We use the ResNet34 [24] as the encoder and the learnable dictionary encoding (LDE) layer [25] with 64 components to aggregate the frame-level features as utterance-level representations. Student is initialized as identical to the teacher model at the beginning. The weights of teacher model are fixed during the student model training. With a mini-batch of 64, we use the Adam optimizer [26] and the learning rate is decayed using the Noam method in [27]. For the teacher model, we apply the similar setting in [25] for network architecture and input samples. For the student, randomly cropped utterances at the fixed-length of 200 frames are utilized.

**Fine-tuning.** With the last layer replaced, the most common way is fine-tuning only the last two fully-connected (FC) layers. We also explore other two layer selections: (1) the last two FC layers plus the LDE layer and the last Residual block (Res4) of the ResNet34; (2) all layers of the embedding network. For all the fine-tuning experiments, we use Adam optimizer with a learning rate decreasing by 10 in every 15 epochs. Two learning rates are used for different layers: 1e-3 used for the replaced last layer and 1e-5 used for rest of the fine-tuned layers. For the best results achieved, the hyper-parameters $\alpha$ and $\beta$ for $L^2$*-SP* and $L^1$*-SP* are set as 0.1 and 0.01, and the weight decay is set as 0.001. The back-end part comprise of LDA with dimension 200, centering, whitening, length normalization, and PLDA.

## 4. Results and Analysis

We first investigate the performance of T-S learning on short utterances with different training objective functions. Table 1 presents the results of different models evaluated on the challenging 1st48-UTD dataset. ResNet34-LDE-S represents the baseline model which is directly trained on the short crop of utterances without T-S learning. ResNet34-LDE-L is the teacher model trained on the long utterances (3s-8s). As expected, the baselines produce better results than the teacher models by decreasing the EERs from 14.97% and 13.83% to 13.35% and 12.79% for using softmax and A-softmax respectively. The results show that training the speaker embedding network with short utterances can improve its performance on short evaluation utterances. But the baseline cannot obtain the improvement for a step further without using the knowledge transferred from the teacher model.

Table 1: *Performance of baseline systems, teacher and student models on the* 1st*48-UTD dataset. Column "Distillation" represents the objective functions used for student model training.*

| System | Distillation | EER(%) |
|---|---|---|
| (a) $L_{\text{class}}$ = Softmax | | |
| ResNet34-LDE-L | - | 14.97 |
| ResNet34-LDE-S | - | 13.35 |
| ResNet34-LDE | $L_{\text{class}}$+$L_{\text{KLD}}$ | 12.31 |
| | $L_{\text{class}}$+$L_{\text{EMD}}$ | 12.03 |
| | $L_{\text{class}}$+$L_{\text{KLD}}$+$L_{\text{EMD}}$ | **11.65** |
| (b) $L_{\text{class}}$ = A-Softmax | | |
| ResNet34-LDE-L | - | 13.83 |
| ResNet34-LDE-S | - | 12.79 |
| ResNet34-LDE | $L_{\text{class}}$+$L_{\text{KLD}}$ | 11.85 |
| | $L_{\text{class}}$+$L_{\text{EMD}}$ | 11.34 |
| | $L_{\text{class}}$+$L_{\text{KLD}}$+$L_{\text{EMD}}$ | **11.12** |

With the same embedding network utilized, the student model can still boost the performance with the advantage of the proposed objective function for T-S learning. As shown in Table 1, different student objective functions contribute to different extends of performance improvement, and all student models outperform their teacher and baseline models. In Table 1-(a), when using the softmax, three student models reduce the EER of the baseline system to 12.31%, 12.03%, and 11.65%, respectively. While using the A-softmax loss can still further boost the performance to 11.85%, 11.34%, and 11.12%. Based on the results, we can find that the models optimizted by $L_{\text{class}} + L_{\text{EMD}}$ consistently outperform that using $L_{\text{class}} + L_{\text{KLD}}$. It indicates that the speaker embedding quality is more relevant to the system performance. Compared with the baselines, it is consistent to observe that models optimized by $L_{\text{class}}+L_{\text{KLD}}+L_{\text{EMD}}$ can achieve the best performances for both softmax and A-softmax with the highest reductions of EER by 12.7% and 13.1% relatively. It makes sense that the student model can learn additional knowledge of embeddings from long utterances besides the speaker discriminative power, and it is more effective to compensate the short utterances.

Considering two student models optimized by the proposed objective function $L_{\text{class}} + L_{\text{KLD}} + L_{\text{EMD}}$ and $L_{\text{class}} + L_{\text{EMD}}$ with A-softmax used, they produce the best two performances on 2-second segments of Voxceleb1 evaluation set with EERs of 6.38% and 6.51%, respectively. Meanwhile, they also produce the best two performances on the target dataset in Table 1. Thus, we pick them as the start points for the fine-tuning step.

Table 2: *Performance comparison of different regularization strategies and layer selections on the best two student models. The first column represents proposed regularizers, the first row represents the different layer selections.*

| | Last 2 FC | Last 2 FC+LDE +Res4 | All Layers |
|---|---|---|---|
| (a) A-Softmax & $L_{\text{class}}$+$L_{\text{KLD}}$+$L_{\text{EMD}}$ | | | |
| w/o regularizer | 11.03 | 10.88 | 11.36 |
| $L^2$-*norm* | 10.85 | 10.52 | 10.97 |
| $L^1$-*SP* | 10.67 | 10.30 | 10.14 |
| $L^2$-*SP* | **10.48** | **10.13** | **9.85** |
| (b) A-Softmax & $L_{\text{class}}$+$L_{\text{EMD}}$ | | | |
| w/o regularizer | 11.28 | 11.20 | 11.52 |
| $L^2$-*norm* | 11.21 | 11.03 | 11.29 |
| $L^1$-*SP* | 10.92 | 10.53 | 10.25 |
| $L^2$-*SP* | **10.70** | **10.31** | **10.09** |

Table 2 shows their fine-tuning results on the 1st48-UTD dataset. We can find that adding regularizer can improve the fine-tuning performance. In Table 2-(a), the regularizers $L^2$-*norm*, $L^1$-*SP*, and $L^2$-*SP* achieve their best performances with 10.52%, 10.14%, and 9.85% in EER. In the lower Table (b), they achieve their best performances with EERs of 11.03%, 10.25%, and 10.09%, respectively. We can observe the trend that with the restriction of reference, $L^2$-*SP* and $L^1$-*SP* consistently outperform the $L^2$-*norm* for all the three layer selection methods. Thus, it is beneficial to fine-tune the pre-trained model with the explicit inductive bias especially when the target dataset is small. For $L^2$-*norm*, the problem of overfitting to a small amount of data will be more severe when lower layers are adapted, and it is shown that the performance of fine-tuning with it begins to degrade from adapting part of the network to adapting all layers. Comparing to the best results achieved by $L^1$-*SP* and $L^2$-*norm* in Table 2, $L^2$-*SP* outperforms them and reduces the EERs with relative 2.9% and 6.4%, respectively. Considering the small size, short duration and other variations mentioned in Section 3.1.1 about the 1st48-UTD dataset, this fine-tuning strategy not only helps to resist forgetting the features learned from large number of speakers in the source domain, but also adapts the model to the target domain efficiently. Our results interpret that the $L^2$-*SP* and $L^1$-*SP* are definitely more efficient than weight decay especially when lower layers are adapted, and $L^2$-*SP* can provide more strict regularization compared to $L^1$-*SP*.

## 5. Conclusion

In this paper, we focus on the approaches for short utterance speaker verification on a small challenging naturalistic forensic dataset. Our results indicate that speaker discriminative power and embedding similarity are two significant points for short utterance speaker verification. The proposed objective function for teacher-student learning could transfer the critical knowledge of these two points from long utterances to short utterances more effectively. Using the novel fine-tuning strategy, speaker embedding networks can be adapted to a new small target dataset with preserving the speaker discriminative power learned from large number of speakers. Experiment results show that our approaches can significantly improve the performance of speaker verification systems on small domain-specific short utterance datasets.

# 6. References


[1] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.

[2] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances." in *Interspeech*, 2017, pp. 1487–1491.

[3] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[4] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

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

[6] W. Xia, J. Huang, and J. H. Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5816–5820.

[7] 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.

[8] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "I-vector based speaker recognition on short utterances," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*. International Speech Communication Association (ISCA), 2011, pp. 2341–2344.

[9] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

[10] J.-w. Jung, H.-S. Heo, H.-j. Shim, and H.-J. Yu, "Short utterance compensation in speaker verification via cosine-based teacher-student learning of speaker embeddings," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 335–341.

[11] A. Hajavi and A. Etemad, "A deep neural network for short-segment speaker recognition," *arXiv preprint arXiv:1907.10420*, 2019.

[12] A. Gusev, V. Volokhov, T. Andzhukaev, S. Novoselov, G. Lavrentyeva, M. Volkova, A. Gazizullina, A. Shulipa, A. Gorlanov, A. Avdeeva *et al.*, "Deep speaker embeddings for far-field speaker recognition on short utterances," *arXiv preprint arXiv:2002.06033*, 2020.

[13] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "Evaluation of i-vector speaker recognition systems for forensic application," 2011.

[14] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 2, pp. 91–101, 2017.

[15] A. K. H. Al-Ali, D. Dean, B. Senadji, V. Chandran, and G. R. Naik, "Enhanced forensic speaker verification using a combination of dwt and mfcc feature warping in the presence of noise and reverberation conditions," *IEEE Access*, vol. 5, pp. 15 400–15 413, 2017.

[16] T. J. Machado, J. Vieira Filho, and M. A. de Oliveira, "Forensic speaker verification using ordinary least squares," *Sensors*, vol. 19, no. 20, p. 4385, 2019.

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

[18] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

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

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

[21] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, p. 101962, 2004.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[23] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," *arXiv preprint arXiv:1804.05160*, 2018.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.