The Speaker and Language Recognition Workshop (Odyssey 2020)
1-5 November 2020, Tokyo, Japan

SPEAKER CHARACTERIZATION USING TDNN, TDNN-LSTM, TDNN-LSTM-ATTENTION BASED SPEAKER EMBEDDINGS FOR NIST SRE 2019

Chien-Lin Huang

PAII Inc., Palo Alto, CA, USA

chiccoCL@gmail.com

ABSTRACT

In this paper, we explore speaker characterization using the time-delay neural network, long short-term memory neural network, and attention (TDNN-LSTM-Attention) based speaker embedding. The speaker embeddings of TDNN, TDNN-LSTM, TDNN-LSTM-Attention are investigated on a large scale of train and testing datasets. Different types of front-end feature extraction are investigated to find good features for speaker embedding. To increase the amount and diversity of the training data, 4 kinds of data augmentation are used to create 7 new copies of the original data. The proposed methods are evaluated with the National Institute of Standards and Technology (NIST) speaker recognition evaluation (SRE) tasks. Experimental results show that the proposed methods achieve the minimum decision cost function of 0.372 and 0.392 with the NIST SRE 2018 and SRE 2019 evaluation datasets, respectively.

Index Terms— speaker embedding, TDNN, TDNN-LSTM, TDNN-LSTM-Attention, NIST SRE 2019

1. INTRODUCTION

The embedding-based speaker recognition systems recently demonstrate very good performance and become the mainstream methods. The idea of speaker embedding is to find representation for speaker idiosyncrasy based on the specific hidden layer of neural networks, which can be extracted for both enrollment data and test utterance, and then used to make a decision regarding true speaker or imposter [1-6].

The NIST SRE is an ongoing speaker recognition evaluation conducted by the US National Institute of Standards and Technology since 1996. For real applications, different challenges are designed such as multilingual, multichannel, and noisy speech in every SRE evaluation. The SRE evaluations encourage the research community to explore the promising new ideas in speaker recognition. There are two tasks in NIST SRE 2019 including CTS and multimedia [7]. The multimedia SRE is a new task of speaker recognition evaluation conducted by NIST. So-called multimedia speaker recognition means we can use audio/visual information in speaker/face recognition. We focus on the conversational telephone speech (CTS) task in this study. The evaluation data of NIST SRE 2019 is spoken in Tunisian Arabic which is an extension of SRE 2018 [8].

In speaker embeddings, variable-length utterances are converted to fixed-dimensional embedding vectors. Different types of neural networks are used to build discriminate between speakers. A speaker embedding method based on deep neural network (DNN) has been proposed based on a feed-forward DNN architecture for text-independent speaker verification [9-12]. The key idea of long-term speaker characteristics can be captured in the network by a temporal pooling layer that aggregates over the input speech. To compute the speaker embedding as a weighted average of a speaker's frame-level hidden vectors, and their weights are automatically determined by an attention mechanism, a self-attention pooling layer was used to replace the temporal average pooling layer for text-independent speaker verification [13-16]. Besides of feed-forward neural networks, the convolutional neural networks (CNN) based speaker recognition was proposed in 2017 [17]. To better capture the temporal information in speech, the architecture of TDNN and long short-term memory (LSTM) recurrent neural networks (RNN) was proposed in 2018 [18, 19]. Such neural network based speaker embeddings can outperform the conventional i-vector when a large amount of training data is available. Different data augmentation methods are proposed to improve the performance of DNN embedding for speaker recognition.

In this study, we explore speaker-embedding models of TDNN, TDNN-LSTM, TDNN-LSTM-Attention for the CTS task of NIST SRE 2019 [20]. We investigate various front-end feature extraction methods to analyze speech from different signal aspects. In addition, 4 kinds of data augmentation methods are used to increase the amount and diversity of the available training data. The rest of this paper is organized as follows. In Section 2, we introduce the proposed methods of neural network based speaker embeddings, data augmentation, front-end feature analysis, back-end scoring, score fusion, and calibration. In Section 3, we present the experimental results on the CTS of NIST SRE 2019. In Section 4, we summarize this work and draw a conclusion.

2. THE PROPOSED METHODS

2.1. Neural network based speaker embeddings

We developed three neural network based speaker embedding systems including TDNN, TDNN-LSTM, and TDNN-LSTM-Attention. First, TDNN is a time-delayed neural network based speaker embedding as shown in Fig. 1(a). The statistics-pooling layer is used to estimate the mean and standard deviation from the variable-length inputs. The speaker-embeddings are extracted from the 512 dimensional affine components of the 6th and 7th layers, which mean the first and second segment-level layers. We found using both of them can have a complementary effect and obtain a better result. Second, TDNN-LSTM means the TDNN and long short-term memory recurrent neural network structure as shown in Fig 1(b). In frame-level layers, we apply 512 and 1,024

10.21437/Odyssey.2020-60

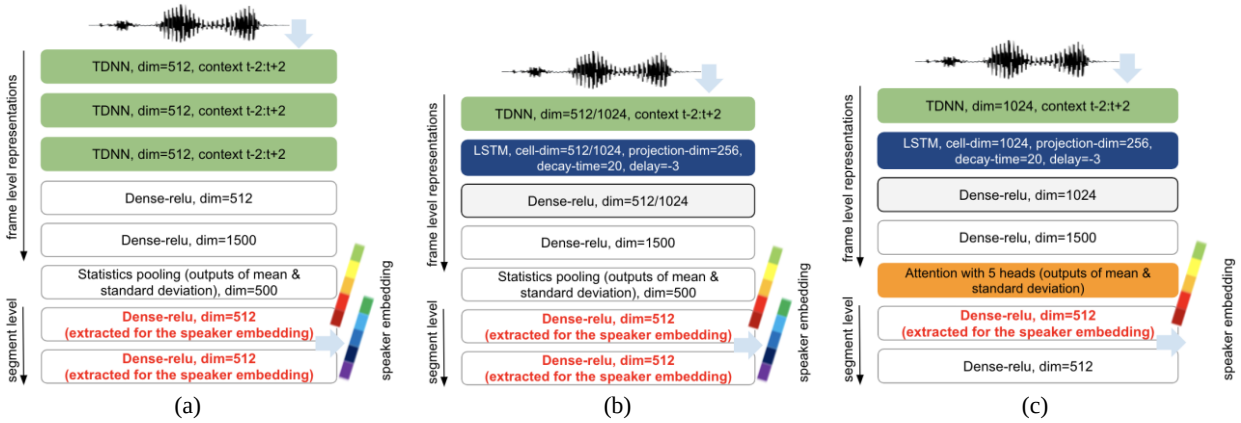

Figure 1. *Diagrams of (a)TDNN, (b)TDNN-LSTM, and (c)TDNN-LSTM-Attention speaker embeddings.*

dimensional neurons to build different systems. The wider neurons of frame-level hidden layers show better performance but are time-consuming. Due to the first and second segment-level layers are complementary, we use both of them in speaker embedding. The motivation of using both TDNN and LSTM is to better capture the temporal information in speech than using TDNN alone approach. Third, TDNN-LSTM-Attention is the TDNN-LSTM and Attention neural network structure in Fig. 1(c). In TDNN-LSTM-Attention, an attention layer instead of the statistical pooling layer is used to capture weighted mean and standard deviation vectors from outputs of all frames over each speech segment. Because the performance of the first segment-level layer is significantly better than the second segment-level layer, only the first segment-level layer is extracted for the speaker embedding of TDNN-LSTM-Attention.

2.2. Front-end feature analysis

Three acoustic feature sets are extracted from audio files, including the Mel-frequency cepstral coefficients (MFC), perceptual linear predictive (PLP) analysis of speech, and the linear mel-scale filter-bank energies with pitch (FBP). MFC features are computed using 24 Mel filter banks. The PLP analysis computes 18-order PLP-cepstra. FBP is estimated using 36 mel-scale filter-bank energies. The audio samples are coded with a 25-ms frame window, a 10-ms frame shift, and bandwidth is limited to the range of 100 Hz - 3,700 Hz. We apply three different front-end feature extraction of MFC, PLP, and FBP to train embedding models. The energy-based voice activity detection (VAD) is used to estimate frame-by-frame speech activity after doing feature extraction. The frames with silence or low signal-to-noise ratio in the audio samples are removed.

2.3. Dataset and data augmentation

The systems are trained on Fisher, Mixer6, NIST SRE, Switchboard (SWBD), and VoxCeleb [18, 21]. The VoxCeleb dataset is collected from Youtube with 16 kHz and in 16-bit format wideband speech. There are 1,245,525 utterances from 7,245 speakers in the VoxCeleb dataset (containing VoxCeleb1 and VoxCeleb2). Fisher dataset contains 23,392 utterances from 12,399 speakers. Mixer6 dataset contains telephone speech of 8,809 utterances from 591 speakers, and microphone speech of 3,423 utterances from 547 speakers, respectively. The NIST SRE dataset, consisting of SRE 2004, 2005, 2006, 2008, and 2010, contains 50,850 utterances from 4,263 speakers. There are 28,181 utterances from 2,594 speakers in the SWBD dataset (with phase 1, phase 2, phase 3, cellular part 1, and cellular part 2). In total, there are around 1,360,000 utterances from 26,600 speakers in the training data. Although the evaluation datasets of NIST SRE 2018 and SRE 2019 are Arabic, the most available train data are English. Because the CTS is 8 kHz telephone speech, all the audio samples are converted to 8 kHz and in 16-bit format for feature extraction.

Data augmentation is often used to increase the amount and diversity of the available training data [22-25]. Because the neural network based speaker embedding is a data greedy approach, the various data augmentation methods are used to create 7 copies of the original train data. First, the MUSAN dataset is applied to corrupt the original audio files with additive general noises, babble noises, and music [26]. Second, the simulated room impulse responses (RIRs) are used to corrupt the original audio by convolving with simulated RIRs including small and medium room size. Third, the speed perturbation is applied to modify the speed to 90% and 110% of the original rate. Fourth, the volume perturbation of original audio files is scaled with a random variable drawn from a uniform distribution over a range. The scaling factor for each audio is randomly chosen from a range (e.g. [0.5, 1.5]). By using the sampled scaling factors, the volume perturbation is used to create more available data for speaker recognition.

The number of speakers is the same after data augmentation but the number of utterances per speaker increases a lot in training neural networks. We throw away the speakers with fewer than 8 utterances and remove features that are too short after removing silence frames. We require at least 400 frames per utterance for training.

2.4. Back-end scoring

A classifier based on probabilistic linear discriminative analysis (PLDA) is used for our speaker embedding systems. All systems are centered and then projected to 150 dimensionalities using LDA. In addition, the length

Table 1. *24 individual system results on the NIST SRE 2018 evaluation dataset.*

		TDNN-LSTM, 5th layer		TDNN-LSTM, 6th layer		TDNN, 6th layer		TDNN, 7th layer	
	adapt	%EER	min_C	%EER	min_C	%EER	min_C	%EER	min_C
FBP	no	7.79	0.526	9.07	0.542	8.58	0.541	8.90	0.531
	yes	**7.77**	**0.455**	7.79	0.461	**8.06**	**0.477**	7.51	**0.445**
PLP	no	8.49	0.528	8.87	0.545	9.12	0.556	9.46	0.567
	yes	7.99	0.475	7.64	0.468	8.04	0.478	8.17	0.489
MFC	no	8.37	0.542	9.19	0.562	8.96	0.565	9.31	0.571
	yes	7.98	0.474	**7.55**	**0.454**	8.11	0.484	**7.45**	0.488

Table 2. *5 additional system results on the NIST SRE 2018 evaluation dataset.*

	adapt	%EER	min_C
AVF	no	8.84	0.587
	yes	**8.09**	**0.508**
LVF	no	8.55	0.564
	yes	**7.99**	**0.488**
VOX	yes	8.60	0.559

Table 3. *Final results on the NIST SRE 2018 and SRE 2019 evaluation datasets.*

dataset	%EER	min_C	act_C
SRE 2018 Evaluation	5.84	0.372	0.374
SRE 2019 Evaluation	6.02	0.392	0.395

normalization and PLDA are applied to speaker embeddings. The LDA and PLDA are trained using the SRE data with data augmentation. The evaluation dataset of the NIST SRE 2019 CTS task is Arabic. Because the training data is essentially all in English, the English (speaker) PLDA can be treated as an out-of-domain PLDA. The SRE 2018 unlabeled data can be used to adapt the out-of-domain PLDA to in-domain PLDA. The adapted PLDA can be treated as Arabic (speaker) PLDA because the SRE 2018 unlabeled data is Arabic. We also found that the PLDA and adapted PLDA are complementary. We can have a better result by the score fusion of two results of PLDA and adapted PLDA.

Instead of the small amount of the SRE 2018 unlabeled data, the SRE 2018 enrollment dataset can be used to train in-domain PLDA to achieve the best performance in the SRE 2019, However, we use the SRE 2018 unlabeled data for both NIST SRE 2018 and SRE 2019 evaluation for the comparison in this study.

3. EXPERIMENT AND ANALYSIS

The proposed systems are evaluated with NIST SRE 2018 and 2019 speaker detection tasks. In speaker detection, it is reasonable to assume the target ratio to be small. In verification, the target rate is often much higher. The performance metrics are the equal error rate in the percentage (%EER) and the minimum of the detection cost function (min_C). Experiments were conducted on the open-source Kaldi Speech Recognition Toolkit [27].

3.1. Effects of TDNN and TDNN-LSTM

Table 1 showed the results of 24 individual systems were evaluated on the NIST SRE 2018 evaluation dataset. The best result was bold face. The 24 results were created by using different combinations of TDNN, TDNN-LSTM, front-end feature analysis, PLDA and adapted PLDA. There were some findings. First, we found the first and second segment-level layers are complementary. By using the score fusion, a better result can be achieved in the SRE evaluation. In addition, the most results of TDNN-LSTM systems outperformed TDNN only systems. The in-domain PLDA adaptation indicated a stable gain compared with the out-of-domain PLDA. We also found that FBP was a good feature which might reflect the pitch feature was important to discriminate speakers.

3.2. Results of TDNN-LSTM-Attention

Since we found FBP was a good feature in the speaker characterization. The TDNN-LSTM-Attention system (AVF) and TDNN-LSTM with 1,024 neurons (LVF) instead of 512 in frame-level representations were trained based on the FBP features. With 1,024 neurons in frame-level representations, we found the first segment-level layer was significantly better than the second layer. Therefore, only the first segment-level layer of AVF and LVF was adopted as the speaker embedding. Both AVF and LVF were trained based on one-third of the total train dataset but the performance was already close to results in Table 1. In addition, the system of VOX denoted a wideband 16 kHz model which was trained on VoxCeleb1 and VoxCeleb2 datasets. The enrollment and testing speech was up-sampled from 8 kHz to 16 kHz including SRE 2018 evaluation and SRE 2019 evaluation datasets. The benefit of VOX was to provide a different type of input in the score fusion and improvement of the final score.

3.3. Score fusion

There was a total of 24+5 subsystems using different speaker embeddings, front-end feature analysis, and back-end scoring in our final submission. The calibration and fusion of 29 subsystems were done by using the BOSARIS toolkit [28]. The fusion weight and bias were learned according to the NIST SRE 2018 evaluation dataset. To prevent the system from overfitting, we applied the *K*-fold cross-validation

(K=10) to verify the results. Table 3 showed the score fusion results on NIST SRE 2018 and SRE 2019 evaluation datasets. Since the data source of SRE 2018 evaluation and SRE 2019 are similar, the performance gain between them is small. We showed an EER of 5.84%, a min_C of 0.372, the actual detection cost function (act_C) of 0.374 in the NIST SRE 2018 evaluation dataset with 2,063,007 trials. We achieved an EER of 6.02%, a min_C of 0.392, the act_C of 0.396 in the NIST SRE 2019 evaluation dataset with 2,688,376 trials. The experiments were tested on machines of NVIDIA DGX station equipped with Intel Xeon E5-2698 CPU 2.2 GHz, 256 GB RDIMM DDR4 and Tesla V100 GPUs. For training neural networks of speaker embeddings, it took about 2-6 weeks depending on methods of data augmentation and neural network parameters.

4. CONCLUSION

In this study, we introduce our systems submitted to the NIST SRE 2019 CTS task. We explore different neural network structures which are based on TDNN, TDNN-LSTM, and TDNN-LSTM-Attention and show the benefit of them. We investigate the MFCC, PLP, and FBP front-end feature analysis. We use training datasets (Fisher + Mixer6 + SRE + SWBD + VoxCeleb) and apply four data augmentation methods to increase the amount and diversity of available training data, including noise addition, convolution, volume perturbation, and speed perturbation. The results of in-domain adapted PLDA, out-of-domain PLDA, speaker embeddings of first and second segment-level layers were used in the final score fusion and calibration, because they are complementary. We evaluate the proposed methods with NIST SRE 2018 and SRE 2019 speaker detection tasks. The proposed methods achieve the act_C of 0.374 and 0.395 in the NIST SRE 2018 and SRE 2019 evaluation datasets, respectively.

5. REFERENCES

1. 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.
2. C.-L. Huang, H. Su, B. Ma, and H. Li, “Speaker characterization using long-term and temporal information,” in Proceedings of INTERSPEECH, 2010.
3. J. H. L. Hansen and T. Hasan, “Speaker recognition by machines and humans: A tutorial review,” IEEE Signal Processing Magazine, vol. 32, no. 6, pp. 74–99, 2015.
4. C.-L. Huang, C. Hori, H. Kashioka, and B. Ma, “Ensemble classifiers using unsupervised data selection for speaker recognition,” in Proceedings of INTERSPEECH, 2012.
5. Q.-B. Hong, C.-H. Wu, H.-M. Wang, C.-L. Huang, “Statistics pooling time delay neural network based on X-vector for speaker verification,” in Proceedings of ICASSP, 2020.
6. Q.-B. Hong, C.-H. Wu, H.-M. Wang, C.-L. Huang, “Combining deep embedding of acoustic and articulatory features for speaker identification,” in Proceedings of ICASSP, 2020.
7. NIST, “NIST 2019 Speaker Recognition Evaluation Plan,” 2019.
8. NIST, “NIST 2018 Speaker Recognition Evaluation Plan,” 2018.
9. 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,” IEEE Spoken Language Workshop (SLT), 2016.
10. D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in Proceedings of INTERSPEECH, 2017.
11. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in Proceedings of ICASSP, 2018.
12. V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in Proceedings of INTERSPEECH, 2015.
13. Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, “Self-attentive Speaker embeddings for text-independent speaker verification,” in Proceedings of INTERSPEECH, 2018.
14. T. Stafylakis, J. Rohdin, O. Plchot, and P. Mizera, L. Burget, “Self-supervised speaker embeddings,” in Proceedings of INTERSPEECH, 2019.
15. M. India, P. Safari, and J. Hernando, “Self multi-head attention for speaker recognition,” in Proceedings of INTERSPEECH, 2019.
16. I. Vinals, D. Ribas, V. Mingote, J. Llombart, P. Gimeno, A. Miguel, A. Ortega, and E. Lleida, “Phonetically-aware embeddings, wide residual networks with time-delay neural networks and self attention models for the 2018 NIST speaker recognition evaluation,” in Proceedings of INTERSPEECH, 2019.
17. A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A large-scale speaker identification dataset,” in Proceedings of INTERSPEECH, 2017.
18. C.-L. Huang, “Exploring effective data augmentation with TDNN-LSTM neural network embedding for speaker recognition,” in Proceedings of ASRU, 2019.
19. C.-P. Chen, S.-Y. Zhang, C.-T. Yeh, J.-C. Wang, T. Wang, and C.-L. Huang, “Speaker characterization using TDNN-LSTM based speaker embedding,” in Proceedings of ICASSP, 2019.
20. NIST, “NIST 2019 Speaker recognition evaluation: CTS challenge,” 2019.
21. J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep speaker recognition,” in Proceedings of INTERSPEECH, 2018.
22. T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in Proceedings of INTERSPEECH, 2015.
23. Z. Wu, S. Wang, Y. Qian, K. Yu, “Data augmentation using variational autoencoder for embedding based

speaker verification," in Proceedings of INTERSPEECH, 2019.

24. H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," in Proceedings of INTERSPEECH, 2019.
25. D. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in Proceedings of INTERSPEECH, 2019.
26. D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," arXiv preprint, 2015.
27. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in Proceedings of ASRU, 2011.
28. Niko Bru ̈mmer and Edward De Villiers, "The Bosaris toolkit: theory, algorithms and code for surviving the new DCF," arXiv preprint, 2013.