PingAn System Description for SdSV 2020 Challenge

Chien-Lin Huang

PAII Inc

chiccocl@gmail.com

Abstract

This paper shows the post-evaluation analysis of our efforts in the INTERSPEECH 2020 Short-duration Speaker Verification Challenge (SdSV). There are two tasks of text-dependent speaker verification (task1) and text-independent speaker verification (task2) from cross-lingual in the SdSV 2020. Our systems are based on x-vectors with different front-end feature extraction methods, data augmentations, and neural network topologies. The score fusion is used to combine different system results. We achieve the minimum decision cost function (minDCF) of 0.2709 and 0.1263 which are equal error rate (EER) of 5.77% and 2.93% for task1 and task2, respectively.

Index Terms: speaker recognition, SdSV 2020, speaker embedding, short-duration, x-vectors

1. Introduction

In the INTERSPEECH 2020 Short-duration Speaker Verification Challenge [1], two tasks are designed to evaluate new technologies for text-dependent and text-independent speaker verification in a short duration scenario with varying degree of phonetic overlap between the enrollment and test utterances (cross-lingual): Task1 of the SdSV Challenge is text-dependent speaker verification. In contrast to text-independent speaker verification, the lexical content of the utterance needs to be taken into consideration like a twofold verification task in which both the speaker and phrase are verified. Task2 of the SdSV Challenge is text-independent speaker verification. Each trial in this task contains a test segment of speech along with a model identifier which indicates one to several enrollment utterances. After applying an energy-based voice activity detection, the net enrollment speech for each model is uniformly distributed between 3 to 120 seconds. The duration of the test utterances varies between 1 to 8 seconds. We present the analysis of multiple speaker verification systems based on neural network based speaker embeddings for the SdSV 2020. We show the differences in performance of the state-of-the-art speaker embedding i-vector and x-vector systems [2]. We also analyze the impact of different front-end feature analysis, training data, data augmentation, and back-end scoring for short-duration data represented in the SdSV 2020 banchmarks. The main objective of this study is to provide a description and analysis of our submission to the SdSV 2020 challenge.

2. System Setup

In the SdSV 2020, a fixed training condition is required which means systems can only be trained using a designated set including VoxCeleb1 [3], VoxCeleb2 [4], LibriSpeech [5], and DeepMine [6]. In this study, we mainly use the VoxCeleb dataset for both task1 and task2 evaluation. The overall dataset involves two parts of VoxCeleb1 and Voxceleb2 which contains over 2,000 hours, over 1 million speech utterances for over 7,000 celebrities. The evaluation dataset used for the challenge is drawn from the recently released multi-purpose DeepMine dataset [7]. The in-domain training data is available for both task1 and task2 but cannot be used in the cross-task training and evaluation. There are 963 speakers in the task1 and some of which have only Persian phrases. The in-domain training data in the task2 contains text-independent Persian utterances from 588 speakers. This data can be used for any purpose such as probabilistic linear discriminant analysis (PLDA) and LDA models, score normalization, training data for neural network, reducing the effect of language for cross-lingual trials, etc.

2.1. Front-end feature analysis

Different types of front-end feature extraction are used to analyze speech from different signal aspects. Three speech feature sets are extracted from audio files, including the Mel-frequency cepstral coefficients (MFC), perceptual linear predictive (PLP) analysis of speech, and Mel-frequency cepstral coefficients with pitch (MFP). The bandwidth is limited between 20 Hz and 7600 Hz. Features are extracted from a 25 millisecond (ms) frame length and a 10ms frame-shift. We used the non-parameter approach of energy-based voice activity detection (VAD) to estimate frame-by-frame speech activity. Without modeling, such as Gaussian mixture model (GMM) classifiers, the frames with silence or low signal-to-noise ratio in the audio samples are removed.

2.2. TDNN-F-LSTM-Attention speaker embedding

The neural network based speaker embedding technologies demonstrate sound performance and become the mainstream methods in speaker recognition. Variable-length utterances are converted to fixed-dimensional embedding vectors. TDNN-F-LSTM-Attention neural network topology is proposed by considering long short-term memory (LSTM), factorized time-delay neural network (TDNN-F), and self-attention pooling. We compare different neural network topologies with the TDNN based x-vector including: TDNN (x-vector) [2], TDNN-LSTM, [8], TDNN-LSTM-Attention [9], and TDNN-F-LSTM-Attention. The long short-term memory recurrent neural networks (RNN) is applied to better capture the temporal information in speech than using TDNN alone as in x-vector [10]. The bigger hidden neurons (1,024 instead of 512) and factorized TDNN are considered in training speaker neural networks. The temporal average pooling layer in x-vector is replaced with an attention pooling layer is applied to automatically determine weights of the speaker's frame-level hidden vectors by an attention mechanism [11]. The self-attention pooling layer with 5 heads is used in this study. Mean

Table 1: Analysis of the systems on the SdSV 2020 challenge.

Task#	System name / Configuration	Progress		Evaluation	
		%EER	minDCF	%EER	minDCF
1	Baseline x-vector	9.05	0.5290	9.05	0.5287
1	Baseline i-vector	3.47	0.1472	3.49	0.1464
1	**Fusion submission of task1 (primary system)**	**5.75**	**0.2720**	**5.77**	**0.2709**
1	TDNN-LSTM-Attention (MFC, LDA-Cosine, dim=500)	6.00	0.2885	6.01	0.2888
1	TDNN-LSTM-Attention (PLP, LDA-Cosine, dim=500)	6.49	0.2901	-	-
2	Baseline x-vector	10.67	0.4319	10.67	0.4324
2	**Fusion submission of task2 (primary system)**	**2.95**	**0.1262**	**2.93**	**0.1263**
2	TDNN-F-LSTM-Attention (MFC, LDA-PLDA, dim=200)	3.22	0.1430	3.21	0.1429
2	TDNN-LSTM-Attention (MFP, LDA-PLDA, dim=200)	3.62	0.1608	-	-
2	TDNN-LSTM-Attention (MFC, LDA-PLDA, dim=200)	3.80	0.1699	-	-
2	TDNN-LSTM (MFC, LDA-PLDA, dim=200)	4.12	0.1818	-	-
2	TDNN-LSTM (PLP, LDA-PLDA, dim=200)	4.61	0.2104	-	-

and standard deviation from the variable-length inputs are estimated in the pooling layer. After the pooling layer, the speaker embedding representation is extracted from the first segment-level layers.

3. Results and Analysis

We evaluated the proposed methods and submitted results to the SdSV 2020 Challenge of task1 and task2. The whole test set of trials is divided into two subsets including a progress set (30%) and an evaluation subset (70%). There are 8,306,700 and 13,198,024 trials (pairs) in the progress subset of task1 and task2, respectively. All results are in Table 1. The best result is bold face. According to results of two baseline systems of the task1, we found the conventional i-vector system is much better than the state-of-the-art x-vector system in text-dependent speaker verification. We evaluate two TDNN-LSTM-Attention neural network speaker embedding systems using MFC and PLP feature extraction for the task1 evaluation. The feature MFC outperforms PLP. Two TDNN-LSTM-Attention systems show theory benefits and demonstrate better performance than the x-vector neural network topology. In our findings, the LDA-Cosine back-end scoring is better than LDA-PLDA in the task1. The LDA dimension is 500 in this study. In the task2, we combine and compare various neural network topologies and front-end feature extraction methods. There are 5 systems as shown in Table 1. We have some findings: First, the front-end feature MFP is better than MFC and PLP. In addition, TDNN-F-LSTM-Attention outperforms TDNN-LSTM-Attention and TDNN-LSTM. Instead of the LDA-Cosine back-end scoring in the task1, the LDA-PLDA scoring with 200 LDA dimensions is more suitable in the task2. Experiments were implemented using the open-source Kaldi Speech Recognition Toolkit [12]. The experiments are 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 takes about 4-8 weeks depending on data and neural network topologies.

4. Conclusions

In this paper, we proposed TDNN-F-LSTM-Attention based speaker embedding for the INTERSPEECH 2020 Short-duration Speaker Verification Challenge. The proposed methods were trained on the VoxCeleb dataset including more than 2,000 hours of speech and 7,000 speakers, and evaluated on the DeepMine dataset for the SdSV 2020 text-dependent and text-independent tasks.

5. References

[1] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Short-duration Speaker Verification (SdSV) Challenge 2020: the Challenge Evaluation Plan," 2020.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, "X-vectors: Robust DNN Embeddings for Speaker Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[3] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," *in Proceedings of INTERSPEECH*, 2017.

[4] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," *in Proceedings of INTERSPEECH*, 2018.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus based on Public Domain Audio Books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[6] H. Zeinali, H. Sameti, T. Stafylakis, "DeepMine Speech Processing Database: Text-Dependent and Independent Speaker Verification and Speech Recognition in Persian and English*,"* in *Odyssey 2018: The Speaker and Language Recognition Workshop*, pp. 386–392, 2018.

[7] H. Zeinali, L. Burget, J. Cernocky, "A Multi Purpose and Large Scale Speech Corpus in Persian and English for Speaker and Speech Recognition: the DeepMine database," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[8] C.-L. Huang, "Exploring Effective Data Augmentation with TDNN-LSTM Neural Network Embedding for Speaker Recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[9] C.-L. Huang, "Speaker Characterization using TDNN, TDNN-LSTM, TDNN-LSTM-Attention based Speaker Embeddings for NIST SRE 2019," in *Odyssey 2020: The Speaker and Language Recognition Workshop*, 2020.

[10] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[11] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *Proceedings of INTERSPEECH*, 2015.

[12] 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 *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.