SPEAKER CHARACTERIZATION USING TDNN-LSTM BASED SPEAKER EMBEDDING

Chia-Ping Chen[1], Su-Yu Zhang[1], Chih-Ting Yeh[1], Jia-Ching Wang[2], Tenghui Wang[2], Chien-Lin Huang[3]

{[1]National Sun Yat-sen University, Kaohsiung | [2]National Central University, Taoyuan }, Taiwan
[3]PAII Inc, Palo Alto CA, USA

國立中山大學 National Sun Yet-sen University
國立中央大學 National Central University

ABSTRACT

We propose speaker characterization using time delay neural networks and long short-term memory neural networks (TDNN-LSTM) speaker embedding. Three types of front-end feature extraction are investigated to find good features for speaker embedding. Three kinds of data augmentation are used to increase the amount and diversity of the training data. Experimental results were evaluated with the proposed methods on the SRE 2016 and SRE 2018.

INTRODUCTION

We propose a speaker-embedding model called L-vectors based on TDNN and LSTM. The motivation of using both TDNN and LSTM in L-vectors is to better capture the temporal information in speech than using TDNN alone as in X-vectors. We investigate three types of front-end feature extraction to analyze speech from different signal aspects. Three kinds of data augmentation are used to increase the amount and diversity of the available training data.

DATASET

A. Training data

The proposed systems are trained on Switchboard, NIST-SRE, Fisher, Mixer 6 and VoxCeleb datasets.

corpora	type	# utts	# spks
Mixer6	microphone	3,423	547
VoxCeleb		1,245,525	7,245
Mixer6	telephone	8,809	591
Switchboard		28,181	2,594
NIST-SRE		50,850	4,236
Fisher		23,392	12,399

Table 1. Training data

B. Data augmentation

1. Babble, noise, and music
2. Room impulse responses
3. Speed perturbation

SYSTEM ARCHITECTURE

A. Feature extraction

- **Feature extraction**:
  1. Mel-frequency cepstral coefficients (MFCCs)
  2. Linear mel-filterbank energies with pitch (FBP)
  3. Perceptual linear predictive (PLP)
- **Voice activity detection:** remove silence or low signal-to-noise ratio frames in the audio samples.

B. Model architecture

We modify the original TDNN-based x-vector by replacing two TDNN layers with an LSTM layer, we refer to this representation as **TDNN-LSTM (L-vector).**

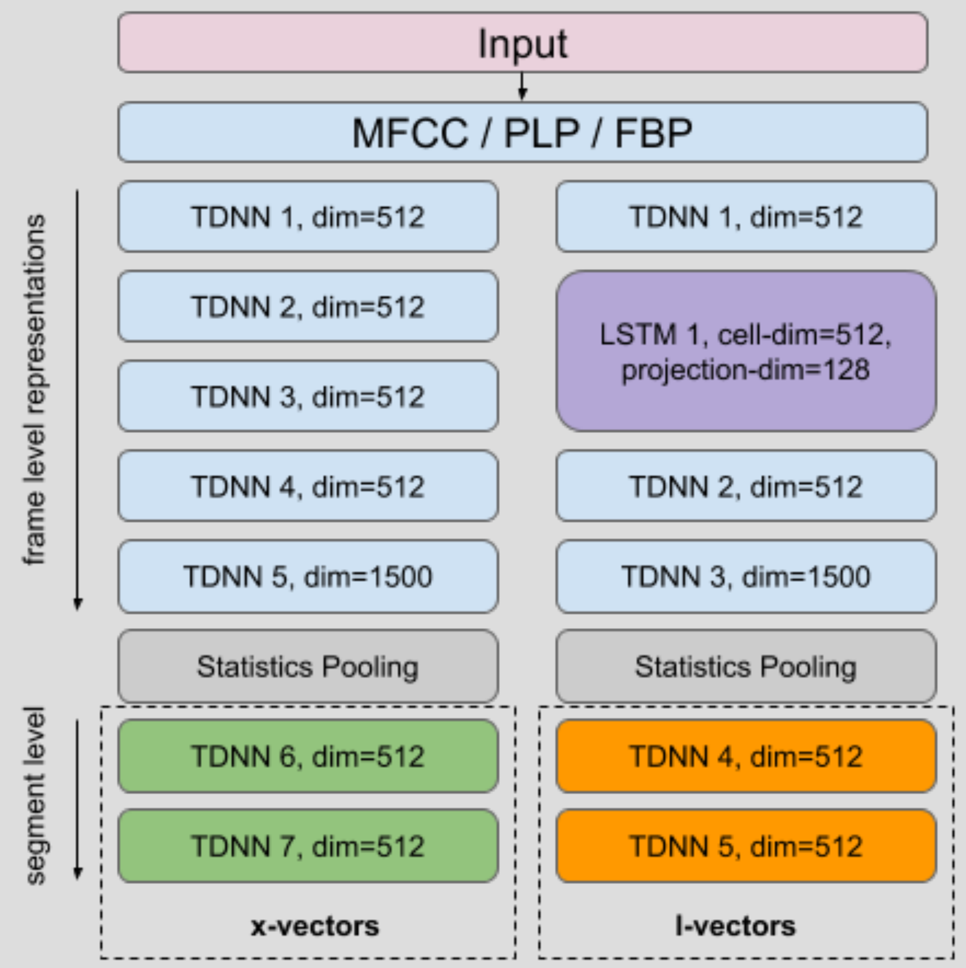

Figure 1. The network architecture of TDNN for X-vectors and TDNN-LSTM for L-vectors.

C. Classifier and score fusion

•Classifier:

1. Out-of-domain PLDA (English)
2. In-domain adapted PLDA : SRE unlabeled data (Arabic)

•Score fusion:

1. **BOSARIS toolkit**: fusion weight and bias are learned by SRE 2018 development data.
2. **Average**: simple average of the output scores of 12 systems.

EXPERIMENTS AND RESULTS

- **Toolkit:** Kaldi and NIST SRE toolkit

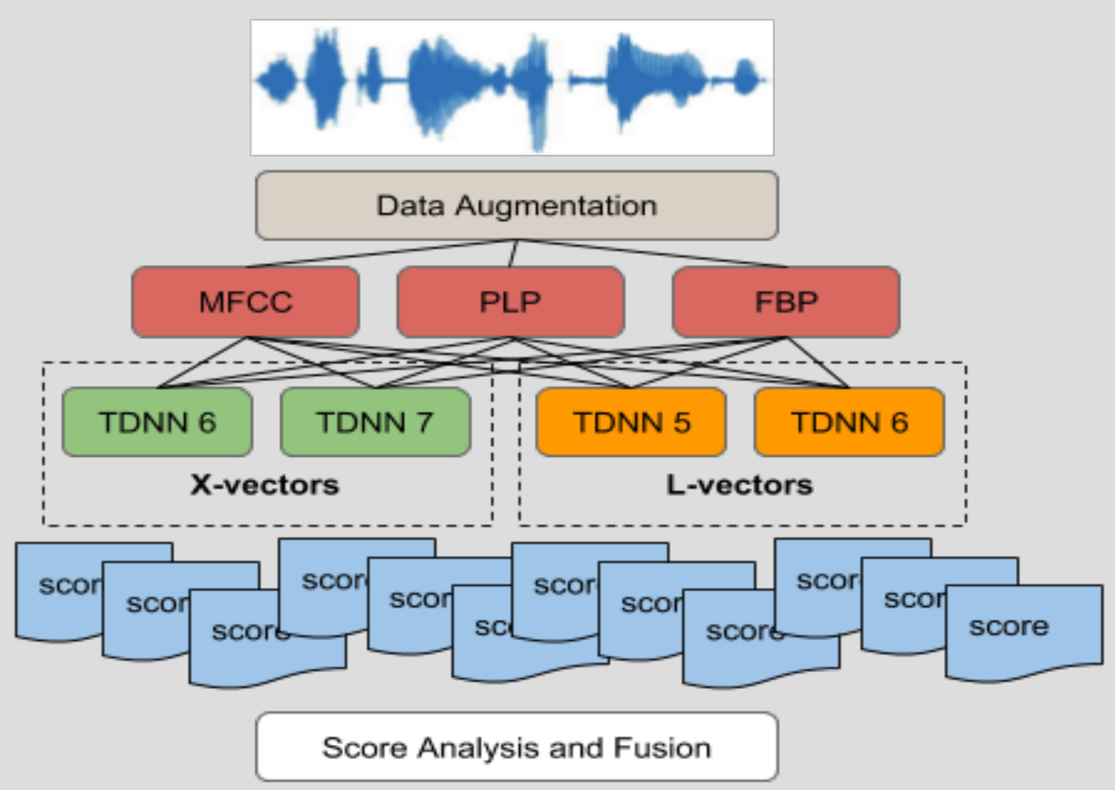

Figure 2. The step-by-step process of the proposed speaker embedding methods.

A. NIST SRE 2018 Results

- **Development set results of single systems**

		L-vector TDNN5			X-vector TDNN6		
		$EER^{\%}$	min_DCF	act_DCF	$EER^{\%}$	min_DCF	act_DCF
MFCC	CMN2	6.91	0.441	0.446	7.58	0.434	0.453
	VAST	3.70	0.416	0.490	5.35	0.333	0.519
	Pooled	--	--	0.468	--	--	0.486
PLP	CMN2	6.98	0.424	0.435	7.31	0.430	0.437
	VAST	3.70	0.267	0.407	7.41	0.412	0.481
	Pooled	--	--	0.421			0.459
FBP	CMN2	6.77	0.412	0.429	7.06	0.402	0.409
	VAST	3.70	0.296	0.370	7.41	0.379	0.416
	Pooled	--	--	**0.400**			0.412

Table 2. EER and DCF results of the SRE 2018 development set.

- **Evaluation set results of score fusion**

		$EER^{\%}$	min_DCF	act_DCF
BOSARIS	CMN2	6.15	0.392	**0.393**
	VAST	16.03	0.668	1.004
	Pooled	--	--	0.699
Average	VAST	11.93	0.489	0.608
	Pooled	--	--	0.501

Table 3. EER and DCF results of the SRE 2018 evaluation set.

B. NIST SRE 2016 Results

	L-vector TDNN5		L-vector TDNN6		X-vector TDNN6		X-vector TDNN7	
	$EER^{\%}$	min_DCF	$EER^{\%}$	min_DCF	$EER^{\%}$	min_DCF	$EER^{\%}$	min_DCF
MFCC	7.03	0.519	7.93	0.519	7.46	0.537	7.71	0.541
PLP	7.42	0.532	8.19	0.532	7.45	0.544	8.13	0.534
FBP	**6.99**	0.520	7.95	0.511	7.14	0.519	7.65	**0.505**

Table 4. EER and DCF results of the SRE 2016 evaluation set.

- **Average of all systems:** Achieve an EER of ***5.56%*** and a minimum DCF of ***0.423*** in NIST SRE 2016 evaluation set.

CONCLUSION

- **FBP** shows the best performance among 3 features.
- **L-vector** with FBP features is the best single system in SRE 2018 development set.
- **Average fusion** of 12 systems achieve better DCF than single system in NIST SRE 2016 evaluation set.