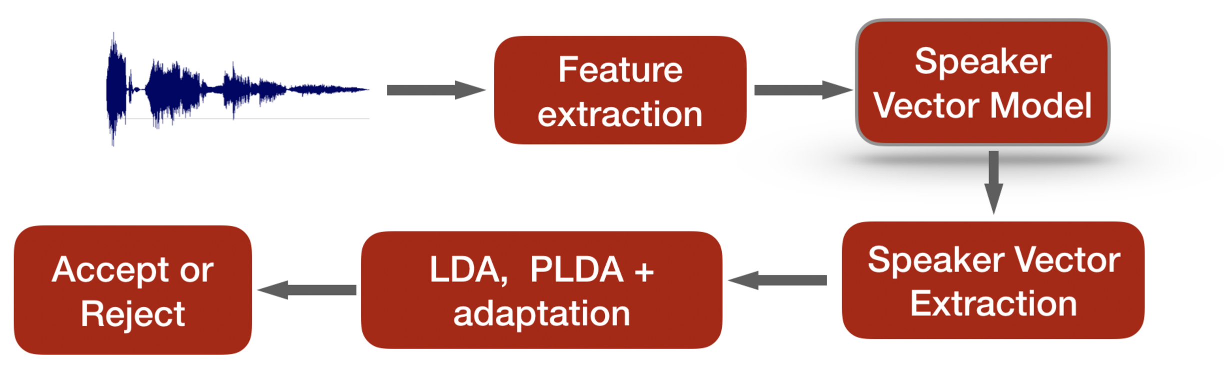
Deep Speaker Embedding Learning with Multi-Level Pooling for Text-Independent Speaker Verification

{ Yun Tang, Guohong Ding, Jing Huang, Xiaodong He, Bowen Zhou } JD AI Research

Speaker Verification



What's New

- A hybrid neural network structure using both TDNN and LSTM
- A multi-level pooling strategy to collect speaker information from both TDNN and LSTM layers
- A regularization scheme on the speaker embedding extraction layer to make the extracted embeddings suitable for the following fusion step

Data

Test Sets

- NIST SRE16 eval set
- NIST SRE18 dev set (CMN2)

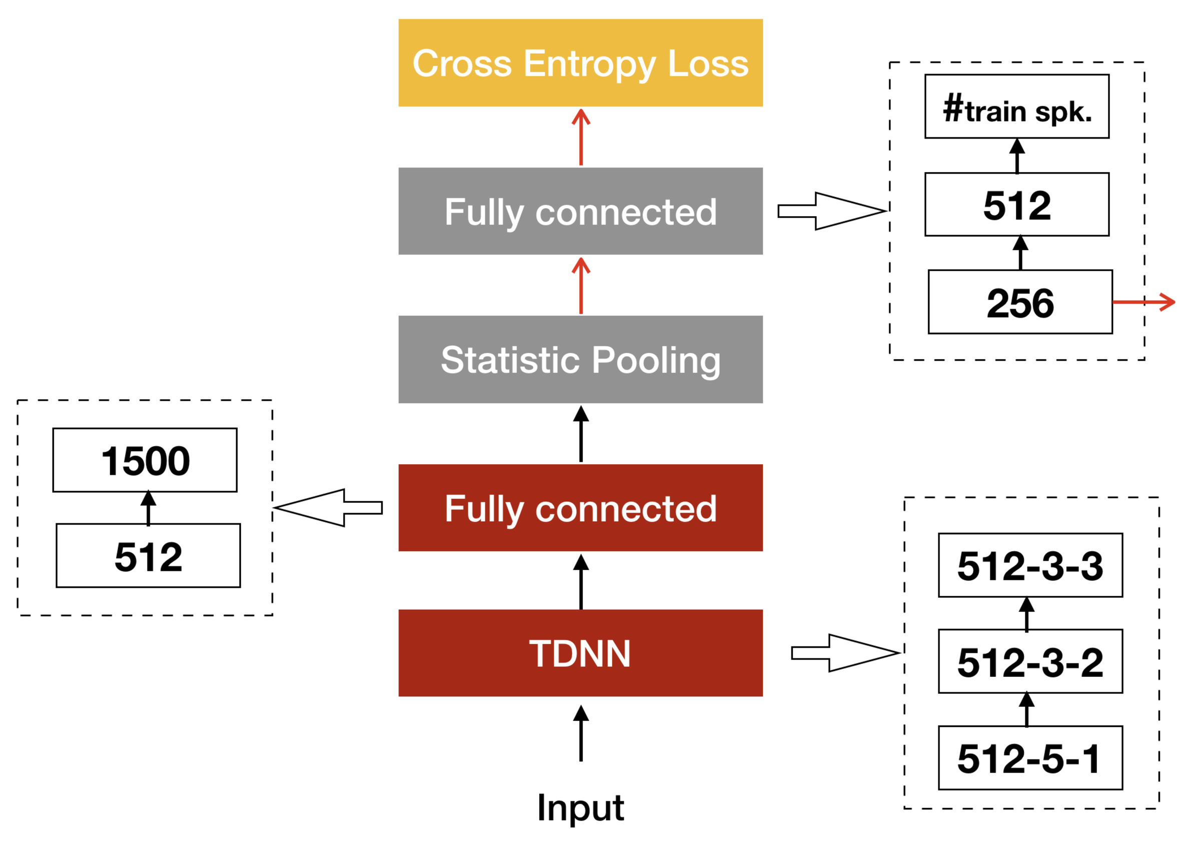
Training data sets:

- SRE data (2004-2006, 2008, and 2010), Switchboard, all Fisher data (1 & 2), all Voxceleb data
- 13,564 hours data from 20,803 speakers
- Data augmentation to deal with different noise conditions

LDA/PLDA adaptation:

- SRE16 unlabelled data is used for SRE16 LDA/PLDA adaptation;
- SRE18 unlabelled data is employed for SRE18 LDA/PLDA adaptation.

Kaldi x-vector Models



X-vector Baseline

- Frame level: 3 TDNN layers
- Speaker Level: Statistic Pooling + 3 fully connected layers

Experimental Setup

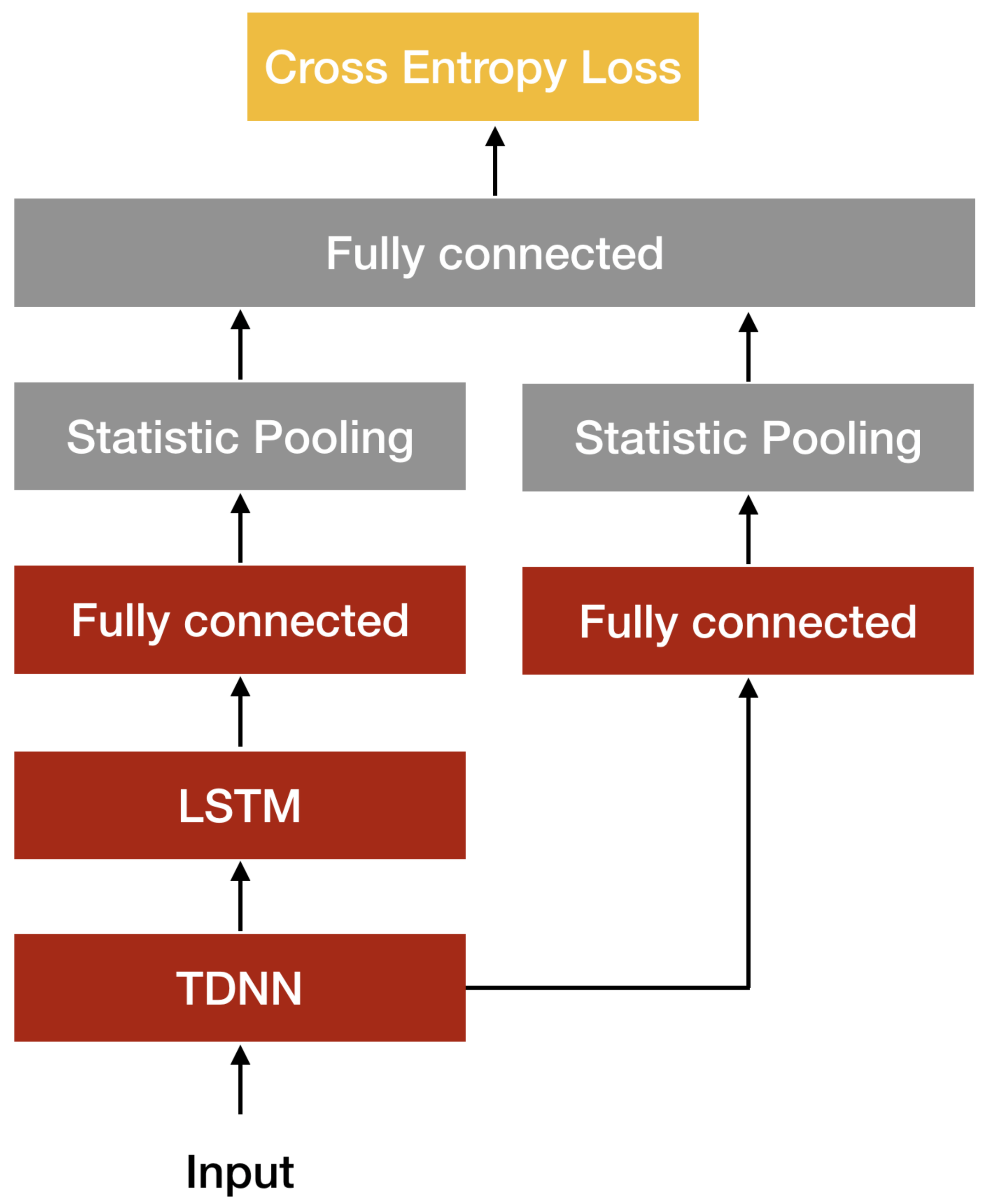
From X-vector model to multiple-level pooing model(MP)

Model	Model Configurations
x-vector	TDNN1-TDNN2-TDNN3-P
A	TDNN1-P-TDNN2-P-TDNN3-P
B	TDNN1-TDNN2-TDNN3-LSTM-P
MP	TDNN1-TDNN2-TDNN3-P-LSTM-P

Table 1: Experimental model configures

- 8kHz data, 40 dimensional filterbank feature + 3 pitch features
- Enrollment data varied from 10 to 60 seconds.

Multi-Level Pooling



$$\mathcal{L} = -\sum_{i=1}^{M} log \frac{\exp^{w_{c_i}^T x_i + b_{c_i}}}{\sum_j^N \exp^{w_j^T x_i + b_j}} + \lambda ||z_i||_2$$

- TDNN focuses on the local feature representation
- LSTM focuses on sequential and global feature representation.
- Multi-level pooling collects different level representations to model the target speaker
- regularization on the embedding extraction layer helps to extract robust representation for the backend process.

SRE16 Results

Model	$\lambda = 0$			$\lambda = 0.001$		
	Pooled	Tag.	Can.	Pooled	Tag.	Can.
x-vector	7.61	10.98	3.95	6.90	10.20	3.50
A	8.17	11.78	4.45	7.51	10.93	4.06
B	**6.64**	9.80	**3.40**	6.84	9.83	3.82
MP	6.68	**9.74**	3.51	**6.13**	**9.18**	**3.13**

Table 2: Results on SRE16 eval test.

model	$\lambda = 0$		$\lambda = 0.001$	
	$p = 0.01$	$p = 0.005$	$p = 0.01$	$p = 0.005$
x-vector	0.593	0.651	0.594	0.673
B	0.581	0.656	0.525	0.586
MP	0.567	0.632	**0.506**	**0.571**

Table 3: DCF scores for SRE16 (pooled) test set

SRE18 Results

λ	model	EER	$DCF(0.01)$	$DCF(0.005)$
0.0	x-vector	7.29	0.593	0.651
	B	7.90	0.581	0.656
	MP	7.16	0.567	0.632
0.001	x-vector	7.46	0.594	0.673
	B	7.77	0.525	0.586
	MP	**6.61**	**0.506**	**0.571**

Table 4: Evaluation results on the SRE18 (CMN2) dev set

- TDNN + LSTM helps to reduce EER by 12% in SRE16
- Regularization improves the verification performance on the backend
- Multiple-pooling from different sources gives the best results on both SRE16 and SRE18 test.