Improved Deep Speaker Feature Learning for Text-Dependent Speaker Recognition

Dong Wang

CSLT / RIIT, Tsinghua University

wangdong99@mails.tsinghua.edu.cn

Co-work with ***Lantian Li, Yiye Lin and Zhiyong Zhang***

IEEE APSIPA 2015, *Hong Kong*, *Dec.* 16-19, 2015

Outline

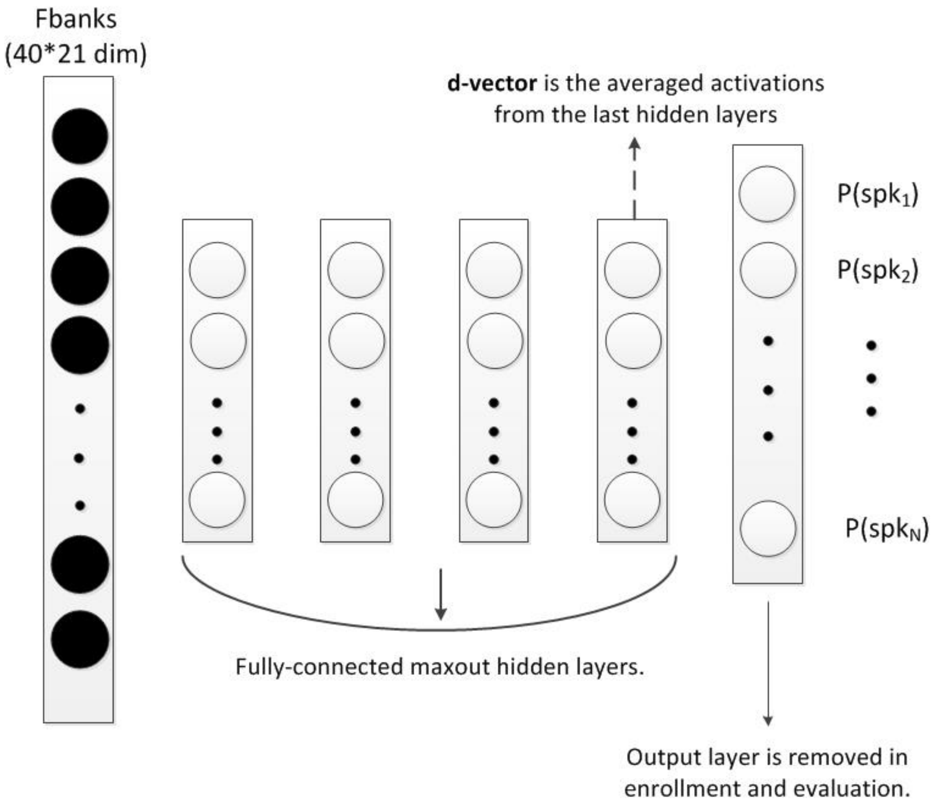
- Introduction
- Improved Deep Feature Learning
- Experiments
- Conclusions

Introduction

- Speaker recognition systems
  - ✧ Human-crafted acoustic features (e.g. *MFCC*)
  - ✧ Statistical models (e.g. *GMM-UBM* (Reynolds 2000), *JFA/i-vector* (Kenny 2007))
- Discriminative models
  - ✧ SVM for GMM-UBM (Campbell 2006)
  - ✧ PLDA for i-vector (Ioffe 2006)

Introduction

- Deep feature learning (Ehsan 2014)



✧ **Drawbacks** of d-vector on ***text-Dep.***

- ✓ Simple input ***feature***
  - No phone content information
- ✓ Simple average ***scoring***
  - Ignoring the temporal constraint

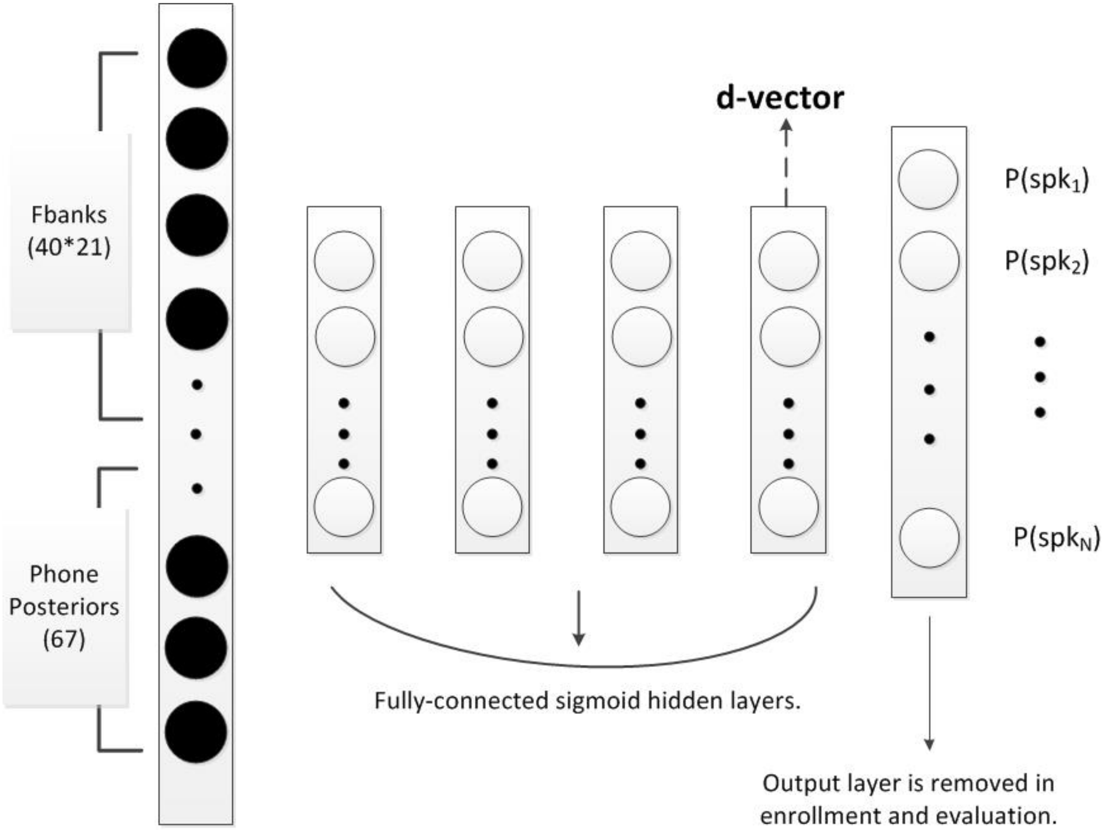
Outline

- Introduction
- **Improved Deep Feature Learning**
- Experiments
- Conclusions

Improved Deep Feature Learning

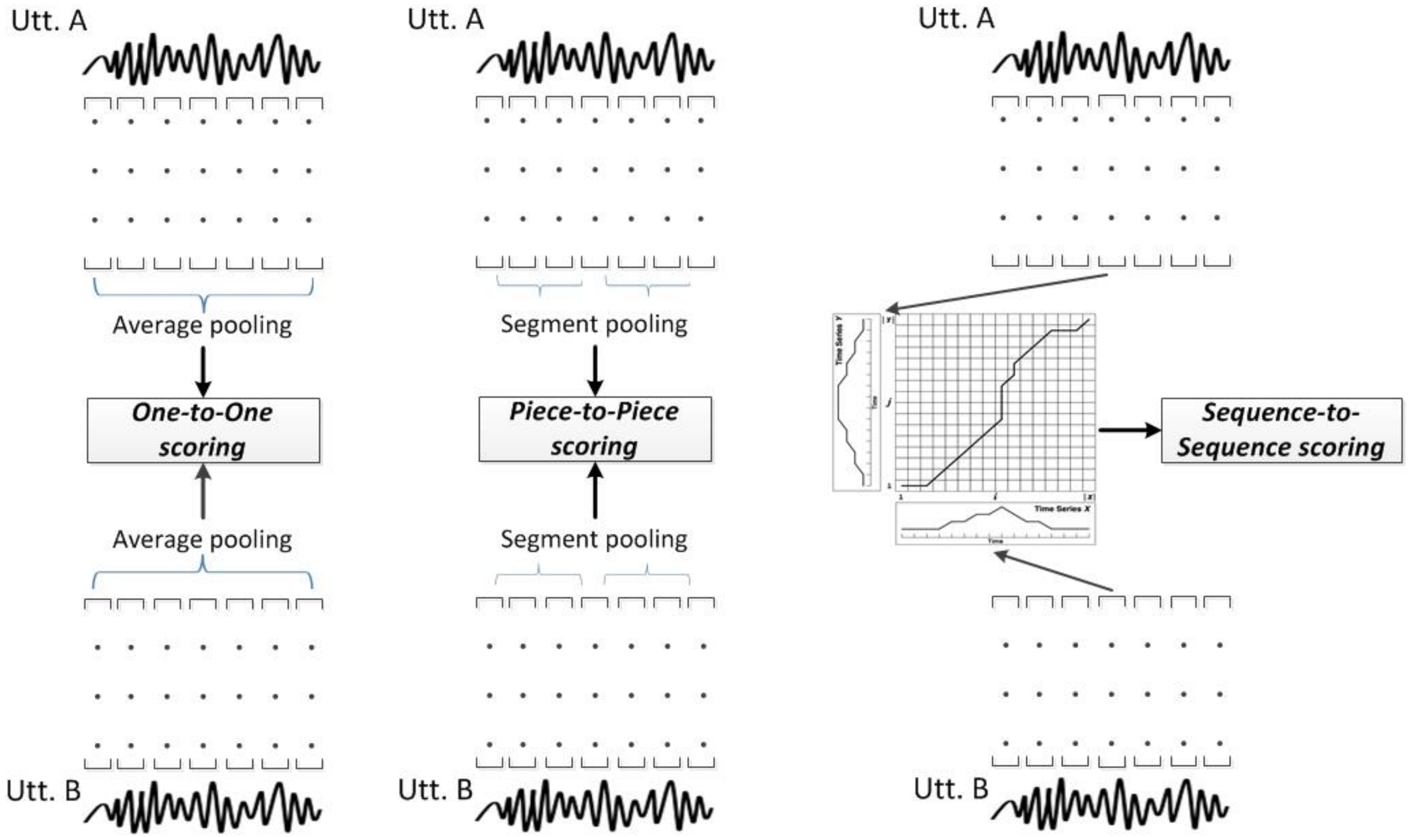
- Phone-depedent training

Speech data
↓
An ASR system
↓
Phone posteriors
↓
Phone-dependent DNN structure



Improved Deep Feature Learning

- Segment pooling and dynamic time warping (DTW) (Berndt 1994)



Outline

- Introduction
- Improved Deep Feature Learning
- **Experiments**
- Conclusions

Experiments

- Database
  - ✧ 100 speakers, 10 short phrases. Each phrase has 150 utterances per speaker.
    - ✓ Dev. Set: 80 speakers and 12000 utterances. → training DNN model / UBM / T matrix / LDA / PLDA.
    - ✓ Eva. Set: 20 speakers, 2100 target trials and 42750 non-target trials for each phrase.
- Experimental Setup
  - ✧ i-vector system
    - ✓ 39-dims MFCCs, 128-components UBM, 200-dims i-vector.
  - ✧ d-vector system
    - ✓ 40-dims Fbanks, 10 left and right frames, 200-dims of each hidden layer.

Experiments

- Baseline

TABLE I
PERFORMANCE OF BASELINE SYSTEMS

		EER%		
	Phrase	cosine	LDA	PLDA
i-vector	P1	2.86	1.81	1.71
	P2	1.52	2.29	1.57
	P3	3.43	3.05	3.05
	P4	3.19	2.86	2.71
	P5	3.57	3.00	2.67
d-vector	P1	10.29	9.81	12.67
	P2	10.52	10.57	12.29
	P3	10.10	9.33	10.48
	P4	10.38	9.95	11.10
	P5	9.14	9.29	11.10

✧ Observations

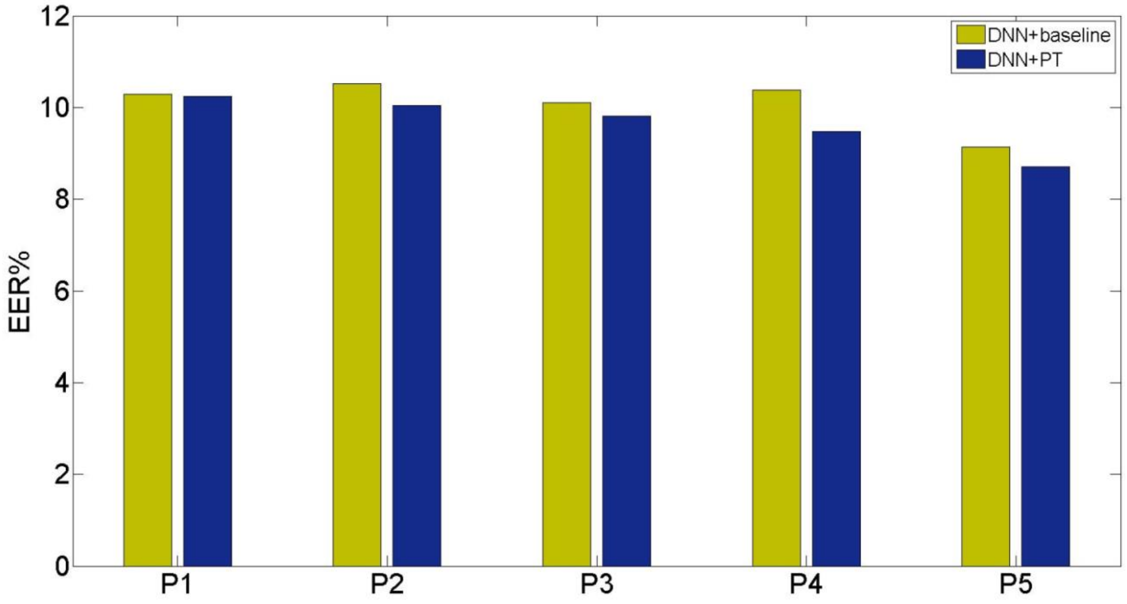
✓ The i-vector ***outperform***s the d-vector.

✓ ***LDA/PLDA*** is suitable for i-vector, while has no effect on d-vector.

✓ The d-vector is a '***discriminative***' vector.

Experiments

- Phone-dependent learning



✧ Descriptions

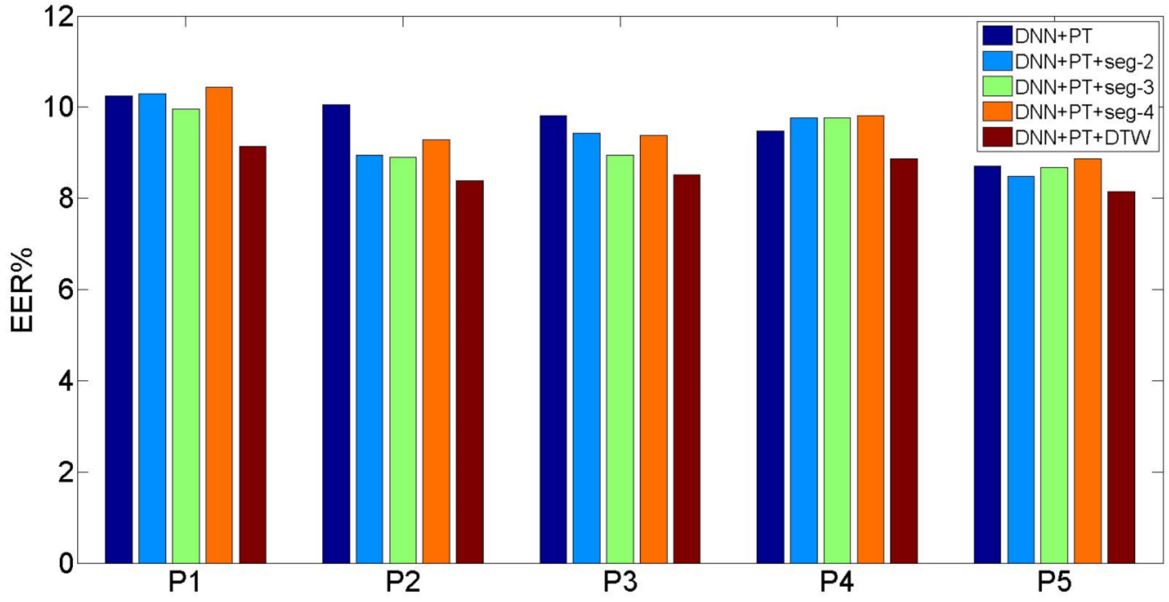
✓A ***DNN*** model was trained for ***ASR*** with a Chinese database consisting of 6000h.

✓The phone set consists of ***66*** initial and finals in Chinese.

✓The '***DNN+PT***' leads to marginal but consistent performance improvement.

Experiments

- Segment pooling and *DTW*



✧ Illustrations

✓The segment pooling(***DNN+PT+seg-n***) generally outperforms the 'DNN+PT'.

✓The '***DNN+PT+DTW***' offers clear performance improvement than the segment pooling.

Experiments

- System combination

TABLE II
PERFORMANCE OF SYSTEM COMBINATION

	EER%				
	P1	P2	P3	P4	P5
PLDA	1.71	1.57	3.05	2.71	2.67
DNN+PT+DTW	9.14	8.38	8.52	8.86	8.14
Combination	1.52	1.38	2.33	2.33	2.38

✧ Descriptions

✓Combine the best i-vector(***PLDA***) and the best d-vector (***DNN+PT+DTW***) from the ***score-level***.

$$s = \alpha s_{iv} + (1-\alpha) s_{dv}$$

where α is the interpolation factor.

✓The combination leads to the best performance.

Outline

- Introduction
- Improved Deep Feature Learning
- Experiments
- **Conclusions**

Conclusions

- A phone-dependent DNN structure.
- Two scoring strategies
  - ✧ Segment pooling
  - ✧ Dynamic time warping
- System combination

References

- D. Reynolds, T. Quatieri, and R. Dunn (**Reynolds 2000**), "Speaker verification using adapted gaussian mixture models," Digital Signal Processing, vol. 10, no. 1, pp. 19–41, 2000.
- P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel (**Kenny 2007**), "Joint factor analysis versus eigenchannels in speaker recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp. 1435–1447, 2007.
- ——, "Speaker and session variability in gmm-based speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp. 1448–1460, 2007.
- P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam (**Kenny 2014**), "Deep neural networks for extracting baum-welch statistics for speaker recognition," Odyssey, 2014.
- W. Campbell, D. Sturim, and D. Reynolds (**Campbell 2006**), "Support vector machines using gmm supervectors for speaker verification," Signal Processing Letters, IEEE, vol. 13, no. 5, pp. 308–311, 2006.
- S. Ioffe (**Ioffe 2006**), "Probabilistic linear discriminant analysis," Computer Vision ECCV 2006, Springer Berlin Heidelberg, pp. 531–542, 2006.
- T. Kinnunen and H. Li (**Kinnunen 2010**), "An overview of text-independent speaker recognition: From features to supervectors," Speech communication, vol. 52, no. 1, pp. 12–40, 2010.
- V. Ehsan, L. Xin, M. Erik, L. M. Ignacio, and G.-D. Javier (**Ehsan 2014**), "Deep neural networks for small footprint text-dependent speaker verification," IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), vol. 28, no. 4, pp. 357–366, 2014.
- D. Berndt and J. Clifford (**Berndt 1994**), "Using dynamic time warping to find patterns in time series," KDD workshop, vol. 10, no. 16, pp. 359–370, 1994.

Thank you

APSIPA ASC, Dec. 16-19, 2015