

# Discriminative Scoring for Speaker Recognition Based on I-vectors

Jun Wang

7/07/2014

# outline

- ◆ Introduction
- ◆ Background theory
- ◆ Motivation of NN approach
- ◆ NN-based discriminative model
- ◆ Experiments
- ◆ Conclusions

## ◆ Introduction

- Background

- i-vector: the most popular approach to speaker verification. [N. Dehak, 2011]
- PLDA: Probabilistic linear discriminant analysis, achieve state-of-the-art performance. [S. Ioffe, 2006][C. S. Greenberg, 2013]

- Motivation

- limitations of PLDA:
  - ✓ assumptions on data distributions.
  - ✓ not directly optimized with respect to speaker verification task.
- the difference between discriminative model and generative model.

- Our approach

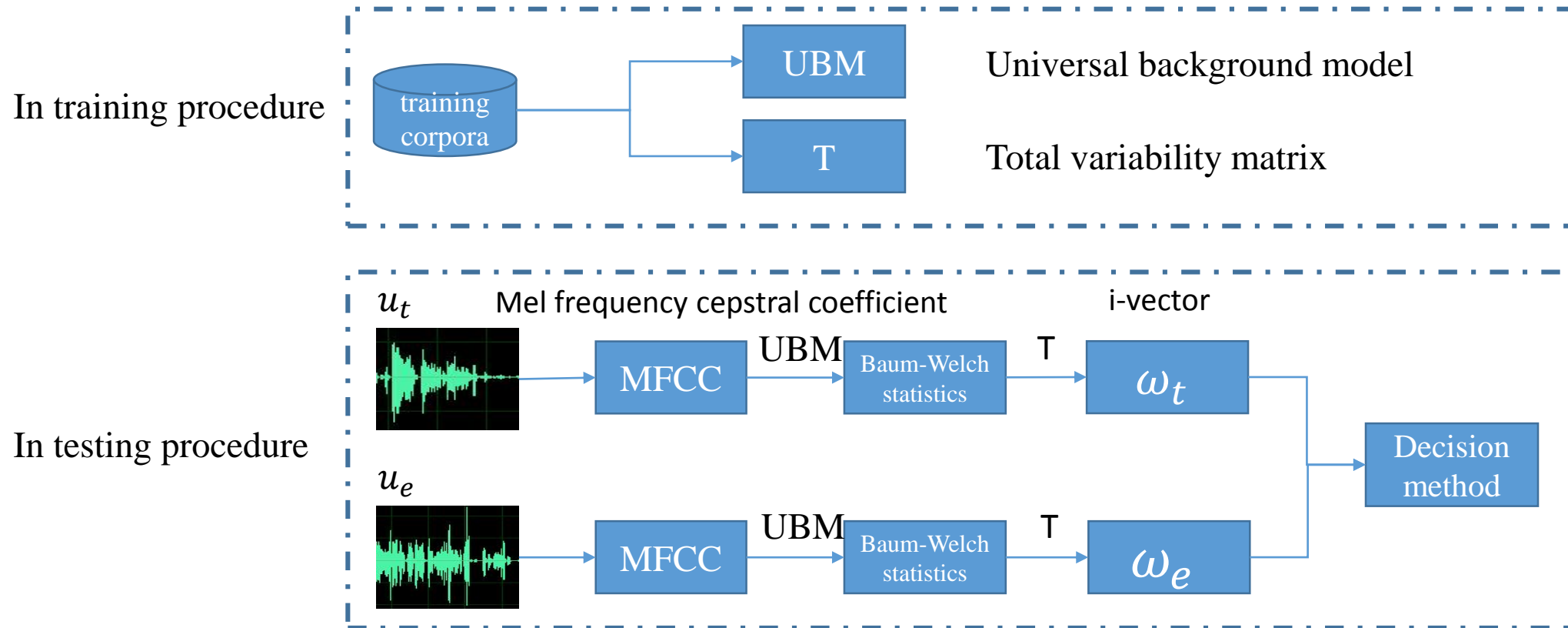
- NN-based ( neural-network-based ) discriminative scoring approach.

## ◆ Theory background

- i-vector[N. Dehak, 2011]

Given a test utterance  $u_t$  and an enrollment utterance  $u_e$ .

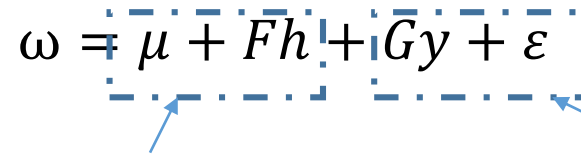
- Speaker verification task is to verify whether  $u_t$  and  $u_e$  are spoken from the same speaker or different speakers.
- i-vector training and testing.



- PLDA [S. J. D. Prince, 2007]

- Probabilistic linear discriminant analysis (PLDA) has been applied successfully to specify a generative model of the i-vector representation, achieves the state-of-the-art performance.

- Technically, assuming a factor analysis (FA) model of the i-vectors of the form:

$$\omega = \mu + Fh + Gy + \varepsilon$$


Speaker dependent part      Session dependent part

- $\omega$  is the i-vector,  $\mu$  is the mean of training i-vectors, and  $h \sim \mathcal{N}(0, I)$  is a vector of latent factors. The full covariance residual noise term  $\varepsilon$  explains the variability not captured through the latent variables.

- Comparison between generative model and discriminative model

	generative model	discriminative model
1	modeling observations drawn from a probability density function	do not need to model the distribution of the observed variables
2	can simulate values of any variable in the model	allows only sampling of the target variables conditional on the observed quantities
3	can generally express more complex relationships between the observed and target variables.	provides a model only for the target variable conditional on the observed variables
4	PLDA	NN

➤ The generative model and discriminative model are seen as complementary.

## ◆ Motivation of NN-based discriminative model

### • limitations of PLDA

- A disadvantage of PLDA lies in its Gaussian assumption of the prior or conditional distributions on the speaker and session variables, which is not necessarily true in reality.

$$\omega = \mu + Fh + \varepsilon$$

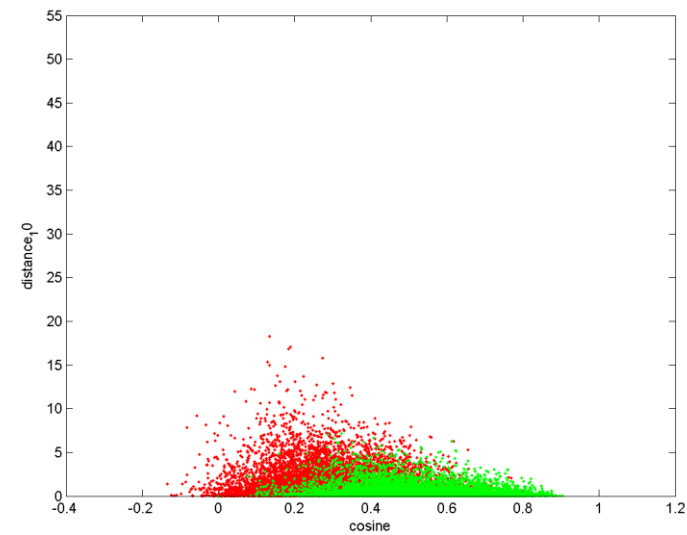
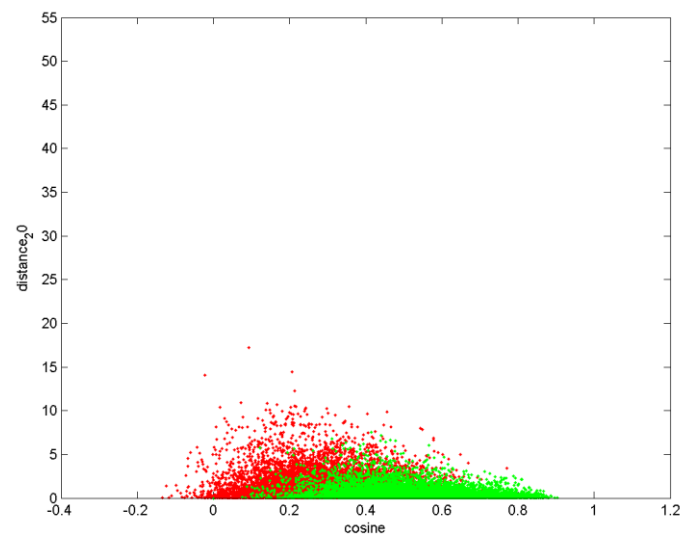
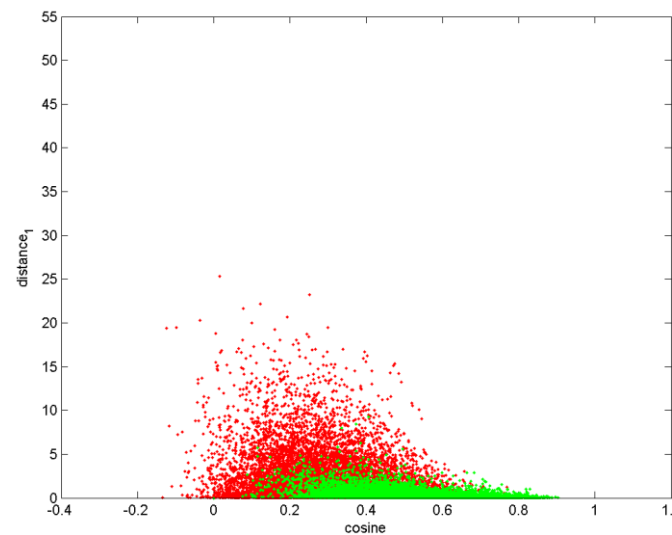
$$h \sim \mathcal{N}(0, I)$$

$$\bar{\omega} = \mu + \varepsilon$$

$$\text{prior: } \bar{\omega} \sim \mathcal{N}(\mu, \varepsilon^T \varepsilon)$$

- not directly optimized with respect to speaker verification task.
- We present a NN-based discriminative approach, which does not rely on any artificial assumptions on data distributions.
- The posterior probability that an i-vector pair belongs to the same person are read off from the NN output directly as the trial score.

- the amplitudes of i-vector also contain speaker information





## ◆ NN-based discriminative model

- We presents a discriminative approach which models i-vector pairs using a neural-network ( NN ).
- Suppose  $\omega_t$  and  $\omega_e$  are two total variability factor vectors extracted from test utterance and enrollment utterance respectively.
- Suppose A is projection matrix obtained by LDA (linear discriminant analysis )
- The cosine kernel [A. Hatch, 2006] between  $\omega_t$  and  $\omega_e$  can be written as:

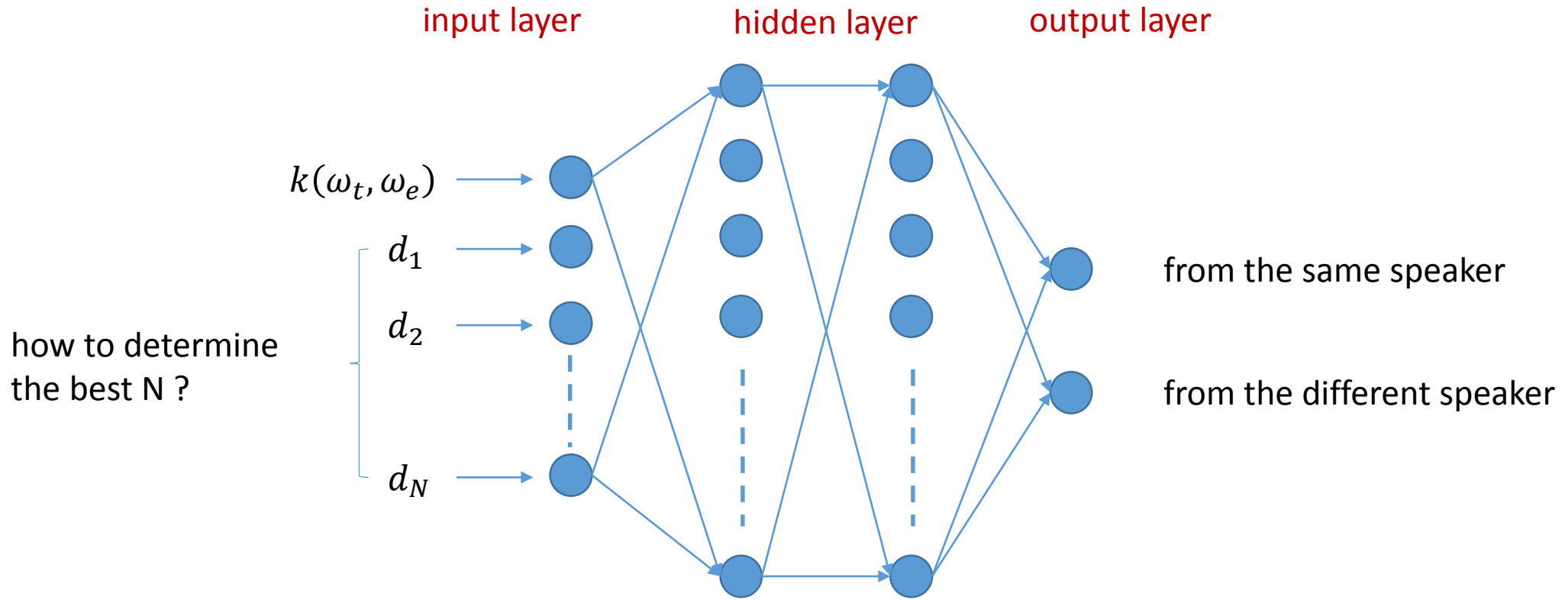
$$k(\omega_t, \omega_e) = \frac{(A' \omega_t)' (A' \omega_e)}{\sqrt{(A' \omega_t)' (A' \omega_t)} \sqrt{(A' \omega_e)' (A' \omega_e)}}$$

$v_t = A' \omega_t$  ,  $v_t^i$  corresponding to the i-th dimension of  $v_t$

$v_e = A' \omega_e$  ,  $v_e^i$  corresponding to the i-th dimension of  $v_e$

$$d_i = (v_t^i - v_e^i)^2$$

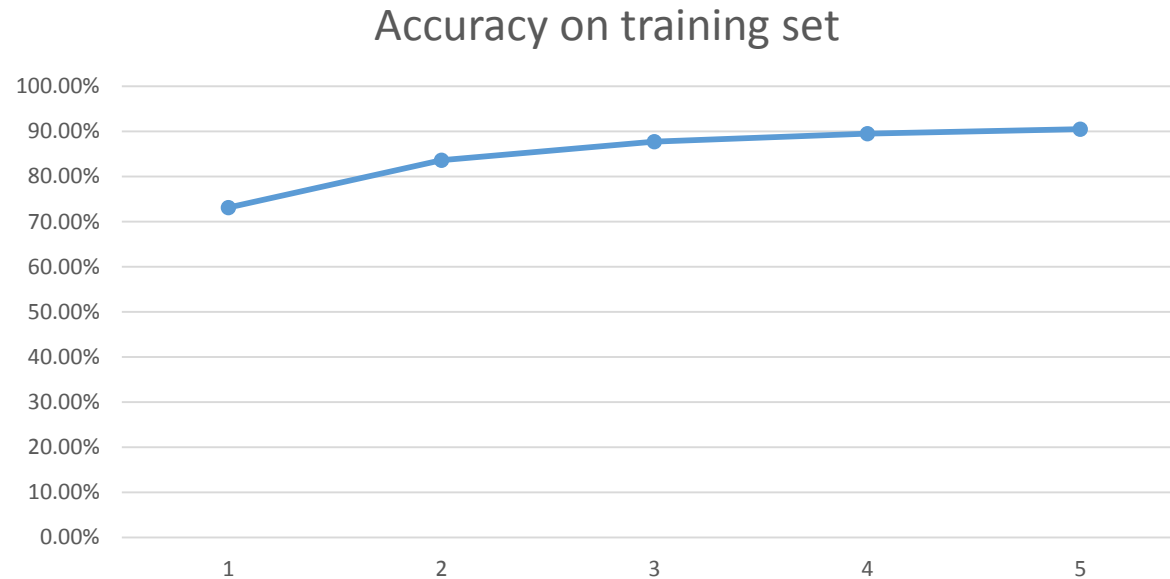
- NN structure



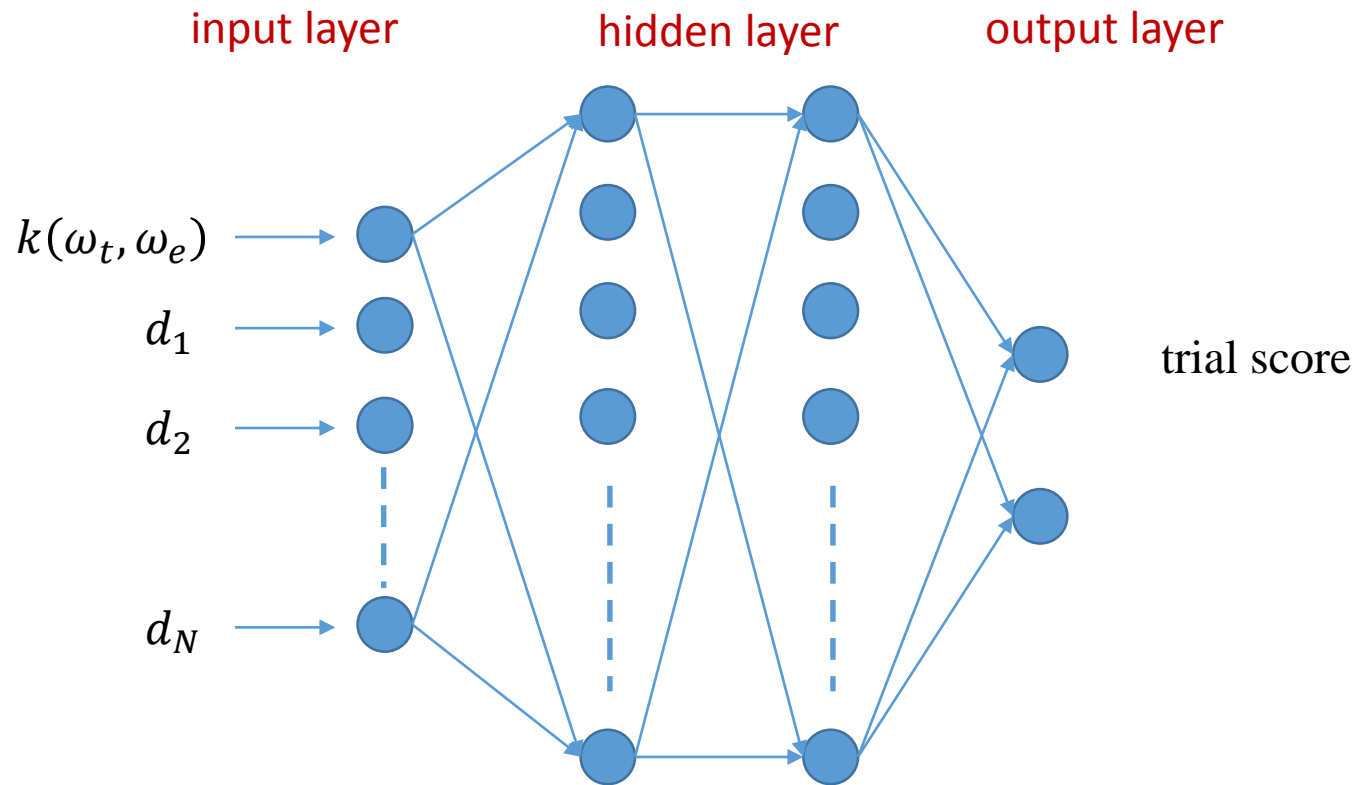
- `$layerdims="--layerdims N+1:200:200:2"`.
- `$epochs="--epochs 5:5:10"`.
- training data: 32500 pairs of utterances, 16250 for same speaker pairs.

- Accuracy on training set

- Different epochs.
- Using all frames for epoch frames, about 32500 frames for each epoch training.



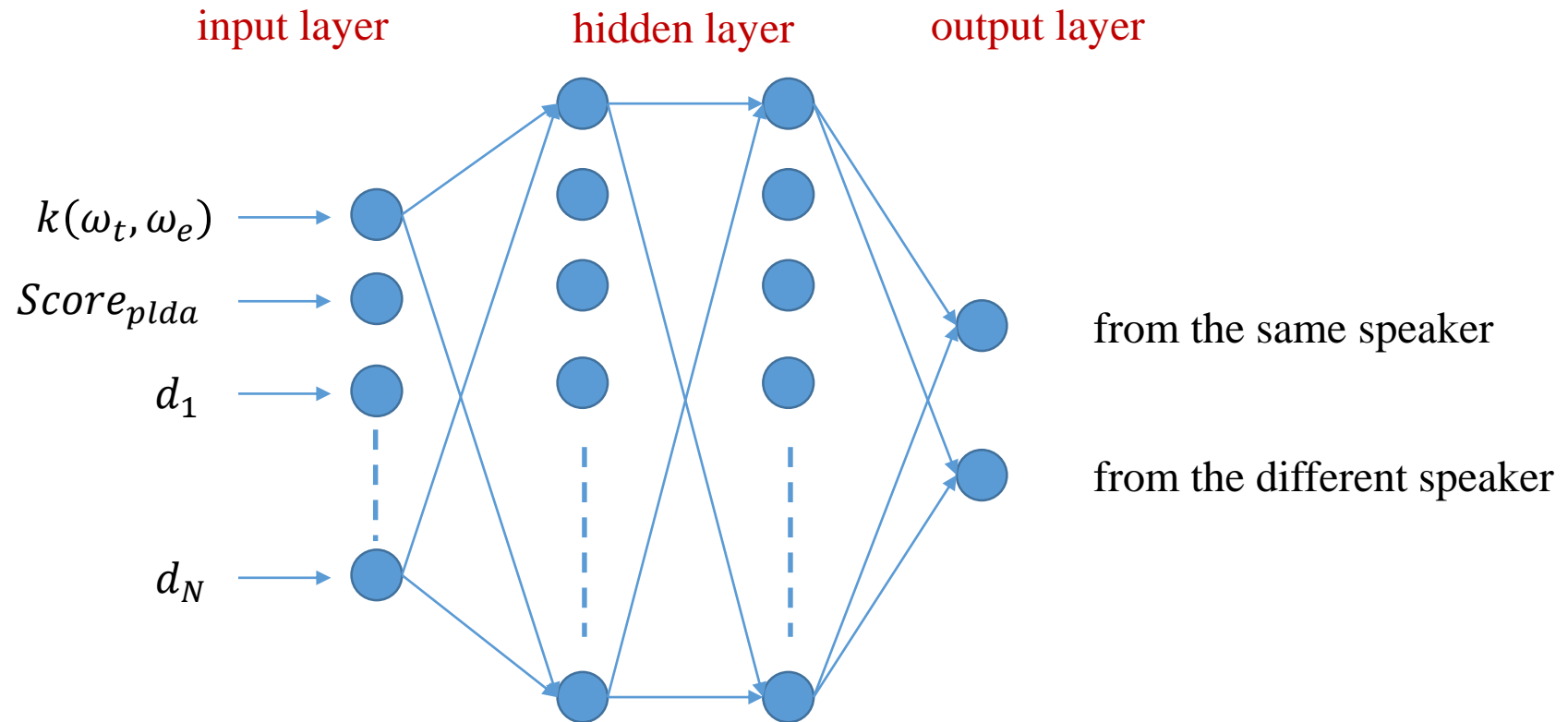
- NN testing



- The posterior probability that an i-vector pair belongs to the same person are read off from the NN output directly as the trial score.

## ◆ Two combination methods of MLP and PLDA

- c\_method 1:  $Score = \alpha Score_{nn} + (1 - \alpha) Score_{plda}$
- c\_method 2: Use PLDA score as one of the input nodes of NN



## ◆ Experiment

- Database

- Development database

- ✓ Fisher English part 1 and 2 as development dataset
- ✓ contains 7196 females (12837 utterances).

- We also define a cross-validation dataset

- ✓ select 100 speakers from SRE08 to build a cross-validation dataset
- ✓ contains about 3000 trials with all 8 common evaluation conditions.

➤ Test database

- ✓ NIST 2008 speaker recognition evaluation (SRE 2008) [NIST, 2008]:
- ✓ core test of SRE 2008 is named short2-short3
- ✓ contains 1997 females and 59343 trials.(including the cross-validation dataset)
- ✓ at least 2 minutes of speech for a given speaker
- ✓ 8 common evaluation conditions and define an all trials condition

	all trials	
	num of trials	proportion %
<b>c1</b>	<b>18898</b>	<b>34.83</b>
<b>c2</b>	<b>957</b>	<b>1.76</b>
<b>c3</b>	<b>17941</b>	<b>33.06</b>
<b>c4</b>	<b>6378</b>	<b>11.75</b>
<b>c5</b>	<b>4354</b>	<b>8.02</b>
<b>c6</b>	<b>22152</b>	<b>40.82</b>
<b>c7</b>	<b>10607</b>	<b>19.55</b>
<b>c8</b>	<b>4959</b>	<b>9.14</b>
<b>c9</b>	<b>54262</b>	<b>100</b>
<b>c10</b>	<b>3000</b>	<b>-</b>

Table 1: proportion of different conditions.

- Experiments setup

- Configurations of i-vector

- ✓ NIST 2008 speaker recognition evaluation (SRE 2008)
    - ✓ sampling rate of the audio signals is 8 kHz and the sample size is 16 bits
    - ✓ 20-dimensional mel-frequency cepstral coefficients, delta and delta-delta
    - ✓ 2048 Gaussian Mixtures
    - ✓ 400 total factors
    - ✓ 150-dimensional LDA, 400-dimensional PLDA

- MLP setup

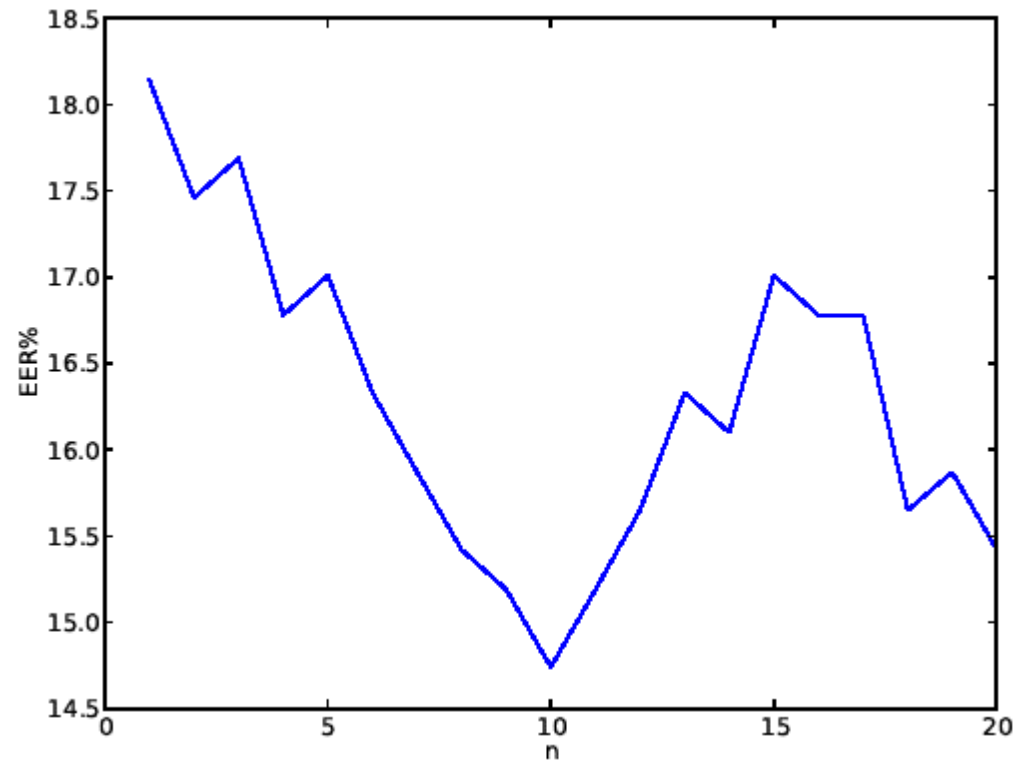
- ✓ 2 hidden layers with 200 nodes
    - ✓ Output layer: 2 nodes, 1 0 for the same speakers, 0 1 for the imposters
    - ✓ training data: 32500 pairs of speakers
    - ✓ epoch frames: using all training data in each epoch
    - ✓ input layer: number of nodes depends on the number of LDA dimensions we choose



- Experiment results

- NN test on the cross-validation dataset

- ✓ N=10 will get the best result



➤ NN test under different conditions

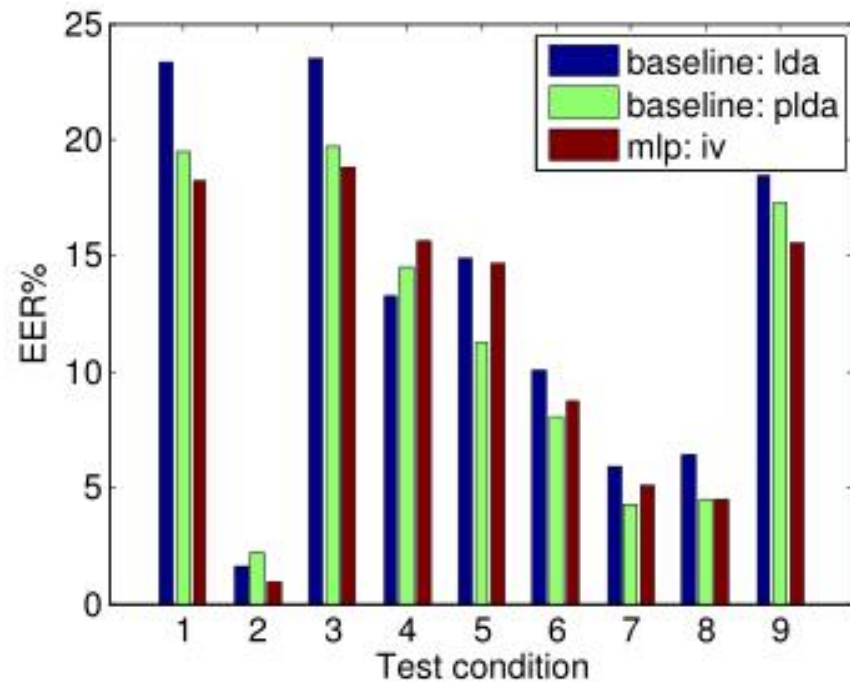


Figure 2: EER comparison under different conditions

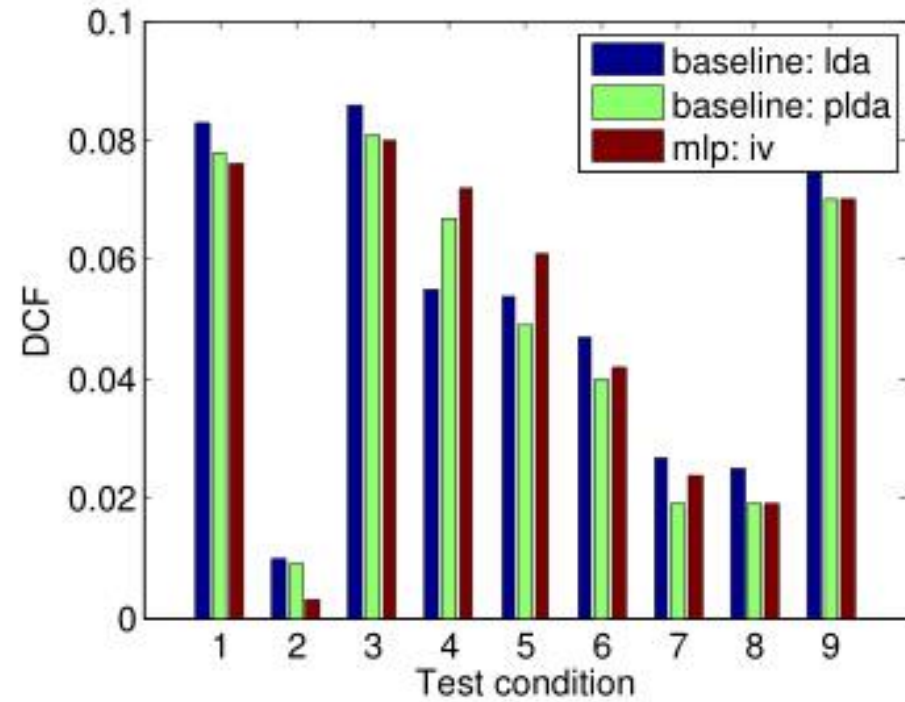


Figure 3: DCF comparison under different conditions

➤ NN test on all trials

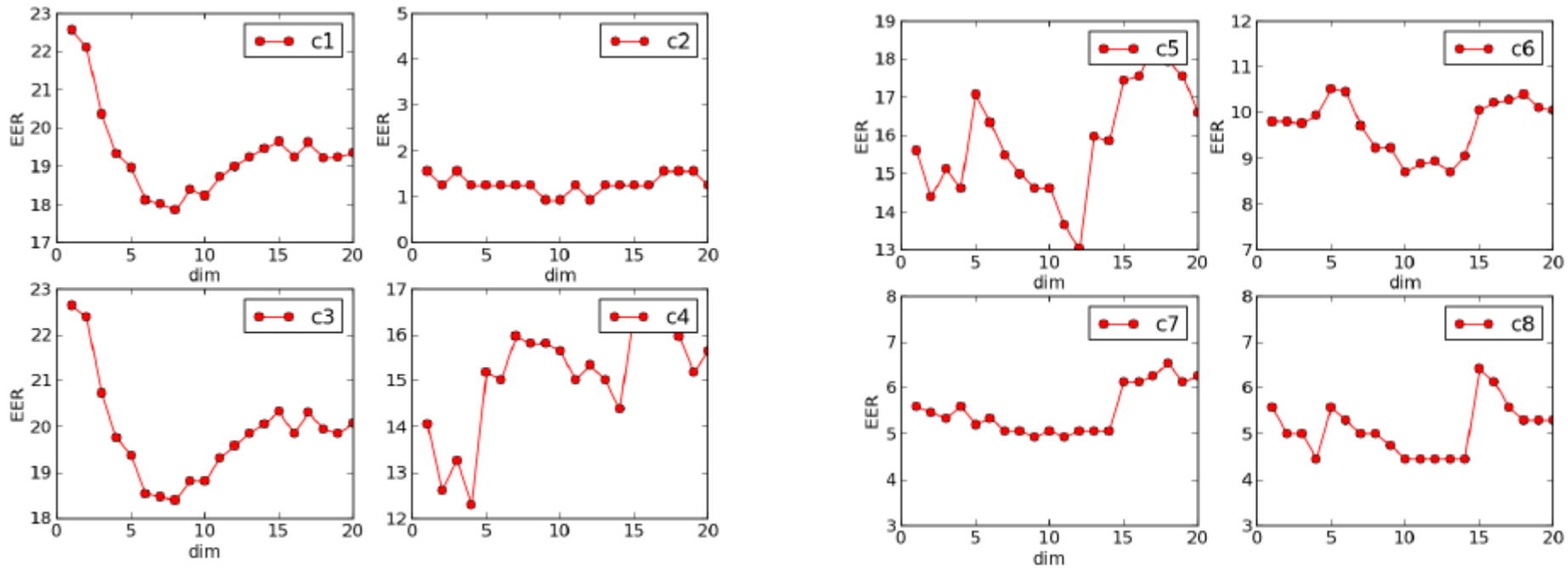
	all trials		
	LDA/MLP	PLDA/MLP	LDA/PLDA
20 classes	3.07e-13	3.11e-08	9.00e-10
30 classes	6.18e-14	1.94e-08	2.17e-07

Table 2: significant value of different methods.

	all trials	
	EER %	DCF
LDA	18.46	0.0797
PLDA	17.22	0.0703
MLP	15.56	0.0702

Table 3: Experiment results on all trials.

➤ NN test with different input dimensions (from N=1 to N=20)



➤ Com

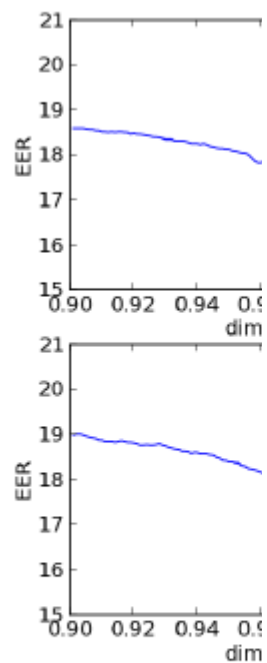
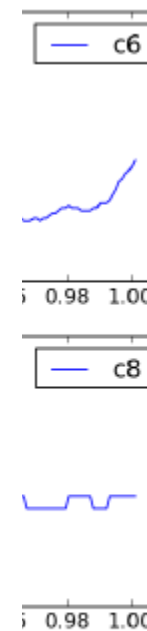
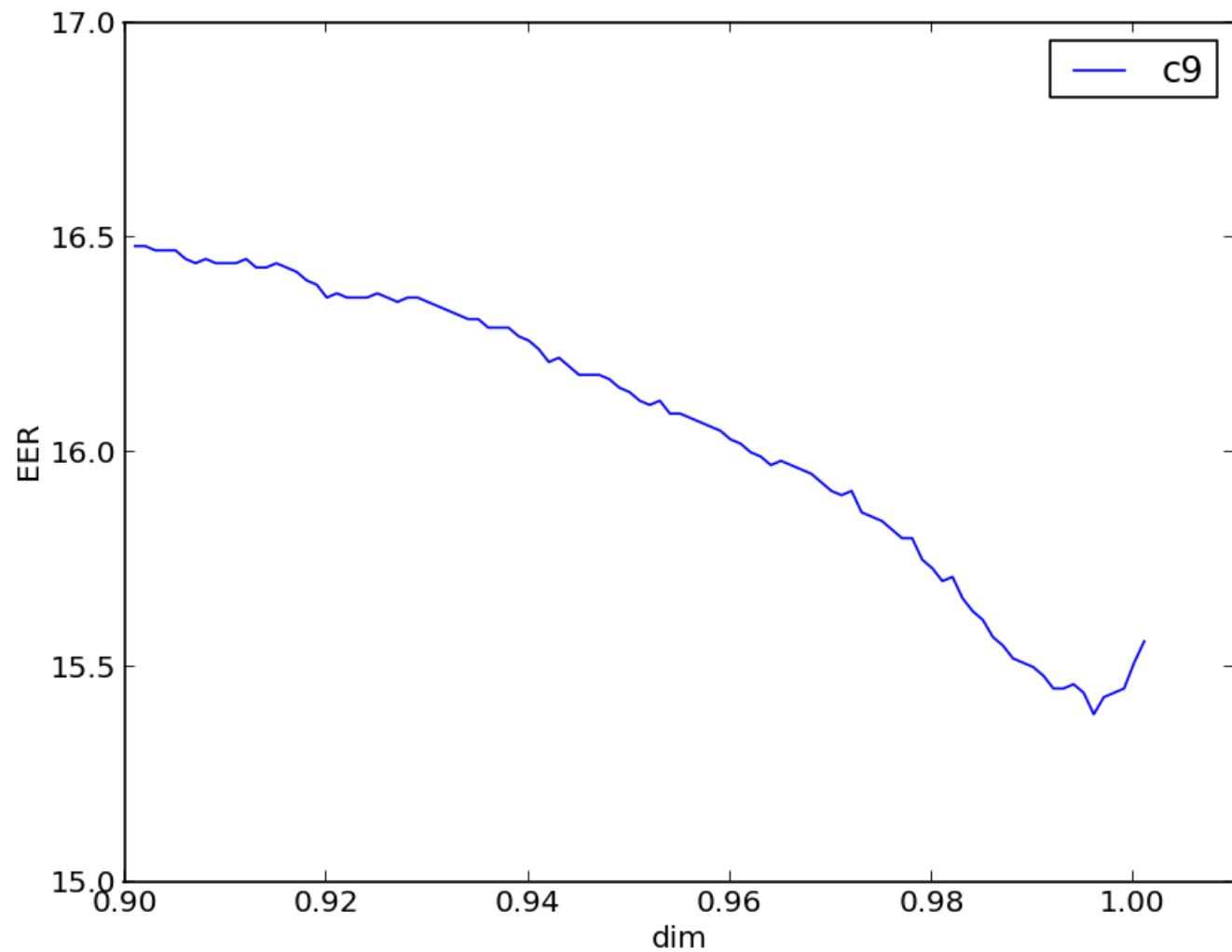


Figure 6: Combine



rent conditions

## ➤ Experiment results

EER%	c1	c2	c3	c4	c5	c6	c7	c8	c9
LDA	24.07	1.49	24.18	14.56	14.54	10.25	6.46	6.58	1.46
PLDA	19.50	2.18	19.71	14.54	11.22	8.06	4.27	4.46	17.22
NN	18.23	0.93	18.82	15.65	14.63	8.70	5.07	4.46	15.56
c_method1	17.57	0.93	17.82	13.26	12.07	7.77	4.40	4.18	15.47
c_method2	17.69	0.93	18.09	12.94	12.20	7.89	4.27	4.18	15.74

# References

- [1] N. Dehak, P. Kenny, R. Dehak, et al. Front-end factor analysis for speaker verification[J]. *Audio, Speech, and Language Processing*, IEEE Transactions on, 2011, 19(4): 788-798.
- [2] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.
- [3] A. Hatch and A. Stolcke, “Generalized linear kernels for one-versus-all classification: application to speaker recognition,” in to appear in *proc. of ICASSP*, Toulouse, France, 2006.
- [4] A. Hatch, S. Kajarekar, and A. Stolcke, “Within-class covariance normalization for SVM-based speaker recognition,” in *Proc. Int. Conf. Spoken Lang. Process.*, Pittsburgh, PA, Sep. 2006.
- [5] S. J. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.
- [6] The NIST Year 2008 Speaker Recognition Evaluation Plan, [http://www.nist.gov/speech/tests/spk/2008/sre-08\\_evalplan-v9.pdf](http://www.nist.gov/speech/tests/spk/2008/sre-08_evalplan-v9.pdf).
- [7] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.
- [8] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, “Within-class covariance normalization for SVM-based speaker recognition.” in *INTERSPEECH’06*, 2006.
- [9] A. Solomonoff, C. Quillen, and W. M. Campbell, “Channel compensation for SVM speaker recognition,” in *Proc Odyssey, Speaker Language Recognition Workshop 2004*, 2004, pp. 57–62.
- [10] S. Ioffe, “Probabilistic linear discriminant analysis,” in *ECCV 2006*, 2006, pp. 531–542.
- [11] M. McLaren and D. V. Leeuwen, “Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors,” *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 5456–5459, 2011.
- [12] N. Dehak, R. Dehak, P. Kenny, P. Ouellet, and P. Dumouchel, “Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification,” in *International Conference on Spoken Language Processing - ICSLP*. IEEE, 2009, pp. 1559–1562.
- [13] C. S. Greenberg, V. M. Stanford, A. F. Martin, M. Yadagiri, G. R. Doddington, J. J. Godfrey, and J. Hernandez-Cordero, “The 2012 NIST speaker recognition evaluation.” 2013.

THANKS