USC University of Southern California

USC Viterbi School of Engineering

Ming Hsieh Department of Electrical Engineering

SPEAKER VERIFICATION USING PLDA Modeling ON LASSO BASED SPARSE TOTAL VARIABILITY SUPERVECTORS

Ming Li, EE department, Signal Analysis and Interpretation Lab

Introduction & I-vector baseline

JFA and i-vector approaches

- JFA: powerful technique for compensating the variability caused by different channels and sessions
  (*Kenny el al. 2007*)
- I-vector: project on a low dimensional total variability space modeling both the speaker and channel variabilities
  (*Dehak el al. 2010*)
  - *front end processing, excellent performance, low complexity and small model size*
  - *Linear Discriminative analysis (LDA) and Within-Class Covariance Normalization (WCCN) are applied*
  - *Simple cosine distance scoring*

Sparse representation for representation

Sparse representation is still employed as a classification method, computational expensive for scoring and score normalization.

Sparse total variability supervectors (s-vectors)

- l^1 norm regularized least square based Lasso approach
- Probabilistic linear discriminant analysis

Difference between i-vector and s-vector

- I-vector: project into low rank subspace
- S-vector: project into large rank space with norm constraint

Complementary with i-vector and JFA subsystems

- Fusion improves the performance

I-vector (1) (*Dehak et al. 2010*)

Given a C component GMM UBM model λ ($\lambda_c = \{p_c, \boldsymbol{\mu}_c, \Sigma_c\}, c = 1, \cdots, C$) and a L frame feature sequence $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_L\}$, the 0^{th} and centered 1^{th} order Baum-Welch statistics on the UBM are calculated as follows:

$$N_c = \sum_{t=1}^{L} P(c|\boldsymbol{y}_t, \lambda) \quad (3)$$

$$\boldsymbol{F}_c = \sum_{t=1}^{L} P(c|\boldsymbol{y}_t, \lambda)(\boldsymbol{y}_t - \boldsymbol{\mu}_c) \quad (4)$$

Concatenating all the $\tilde{\boldsymbol{F}}_c$ together to generate centered mean supervector $\tilde{\boldsymbol{F}}$:

$$\tilde{\boldsymbol{F}}_c = \frac{\sum_{t=1}^{L} P(c|\boldsymbol{y}_t, \lambda)(\boldsymbol{y}_t - \boldsymbol{\mu}_c)}{\sum_{t=1}^{L} P(c|\boldsymbol{y}_t, \lambda)}. \quad (5)$$

I-vector (2) (*Dehak et al. 2010*)

The speaker and channel dependent centered GMM mean supervector $\tilde{\boldsymbol{F}}$ can be projected on a rectangular low rank total variability matrix $\boldsymbol{T}$:

$$\tilde{\boldsymbol{F}} = \boldsymbol{T}\boldsymbol{w} \quad (6)$$

The i-vector is computed by the Eigenvoice approach (*Dehak et al. 2010*):

$$\boldsymbol{w} = (\boldsymbol{I} + \boldsymbol{T}^t\Sigma^{-1}\boldsymbol{N}\boldsymbol{T})^{-1}\boldsymbol{T}^t\Sigma^{-1}\boldsymbol{N}\tilde{\boldsymbol{F}} \quad (7)$$

$\boldsymbol{N}$ is a CF × CF diagonal matrix whose diagonal blocks are $N_c\boldsymbol{I}, c = 1, \cdots, C$.
Σ is a CF × CF diagonal covariance matrix capturing residual variability.
Only 1^{th} order statistics adopted here, Σ is the concatenated version of Σ_c
The identity matrix is the prior of the i-vector $\boldsymbol{w}$.

S-vector & PLDA

S-vector (1)

Let's review equation (6) $\tilde{\boldsymbol{F}} = \boldsymbol{T}\boldsymbol{w}$ and (7) $\boldsymbol{w} = (\boldsymbol{I} + \boldsymbol{T}^t\Sigma^{-1}\boldsymbol{N}\boldsymbol{T})^{-1}\boldsymbol{T}^t\Sigma^{-1}\boldsymbol{N}\tilde{\boldsymbol{F}}$
The Maximum Likelihood (ML) part of equation (7) is:

$$\boldsymbol{w} = (\boldsymbol{T}^t\Sigma^{-1}\boldsymbol{N}\boldsymbol{T})^{-1}\boldsymbol{T}^t\Sigma^{-1}\boldsymbol{N}\tilde{\boldsymbol{F}}, \quad (9)$$

which is a weighted least square solution to equation (6).
Denote the normalized version of T and $\tilde{\boldsymbol{F}}$ as:

$$\acute{\boldsymbol{F}} = \tilde{\boldsymbol{F}}\Sigma^{-\frac{1}{2}}\boldsymbol{N}^{\frac{1}{2}} \quad (10)$$

$$\acute{\boldsymbol{T}}^k = \boldsymbol{T}^k\Sigma^{-\frac{1}{2}}\boldsymbol{N}^{\frac{1}{2}}, k = 1, \cdots, K. \quad (11)$$

Then, it becomes a standard least square estimation:

$$\boldsymbol{w} = (\acute{\boldsymbol{T}}^t\acute{\boldsymbol{T}})^{-1}\acute{\boldsymbol{T}}^t\acute{\boldsymbol{F}}. \quad (12)$$

S-vector (2)

The Lasso based l^1 norm regularized least square estimation ($\acute{\boldsymbol{w}}$ is S-vector):

$$\min \quad \|\acute{\boldsymbol{F}} - \acute{\boldsymbol{T}}\acute{\boldsymbol{w}}\|_2^2 \quad \text{subject to} \quad \|\acute{\boldsymbol{w}}\|^1 < \tau. \quad (13)$$

$$\min \quad \sum_{c=1}^{C} N_c(\tilde{\boldsymbol{F}}_c - \boldsymbol{T}_c\acute{\boldsymbol{w}}_c)^t\Sigma_c^{-1}(\tilde{\boldsymbol{F}}_c - \boldsymbol{T}_c\acute{\boldsymbol{w}}_c) \quad \text{subject to} \|\acute{\boldsymbol{w}}\|^1 < \tau. \quad (14)$$

Upper bound of KL divergence between GMM λ^1 and λ^2 (*Campbell 2006*).
$\lambda_c^1 = \{\frac{N_c}{\sum_{c=1}^{C} N_c}, \tilde{\boldsymbol{F}}_c, \Sigma_c\}, \lambda_c^2 = \{\frac{N_c}{\sum_{c=1}^{C} N_c}, \boldsymbol{T}_c\acute{\boldsymbol{w}}_c, \Sigma_c\}, c = 1, \cdots, C$

PLDA modeling

Assume the training data consists of J utterances from I speakers and the j^{th} s-vector of the i^{th} speaker ($\boldsymbol{x}_{ij}$) is generated by:

$$\boldsymbol{x}_{ij} = \boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{h}_i + \boldsymbol{G}\boldsymbol{w}_{ij} + \epsilon_{ij}. \quad (15)$$

- $\boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{h}_i$ is only dependent on the speaker index
- $\boldsymbol{G}\boldsymbol{w}_{ij} + \epsilon_{ij}$ is different for every s-vector and used to model the within-speaker variances
- Model parameters are estimated by Expectation Maximization (EM)
- hypothesis testing $P(S|H_1)/P(S|H_0)$ based log likelihood ratio test for a pair of s-vectors $S(\boldsymbol{w}_i, \boldsymbol{w}_j)$
  - H_1 a true trial
  - H_0 a false trial
- Our implementation is based on (*Prince et al. 2007*).

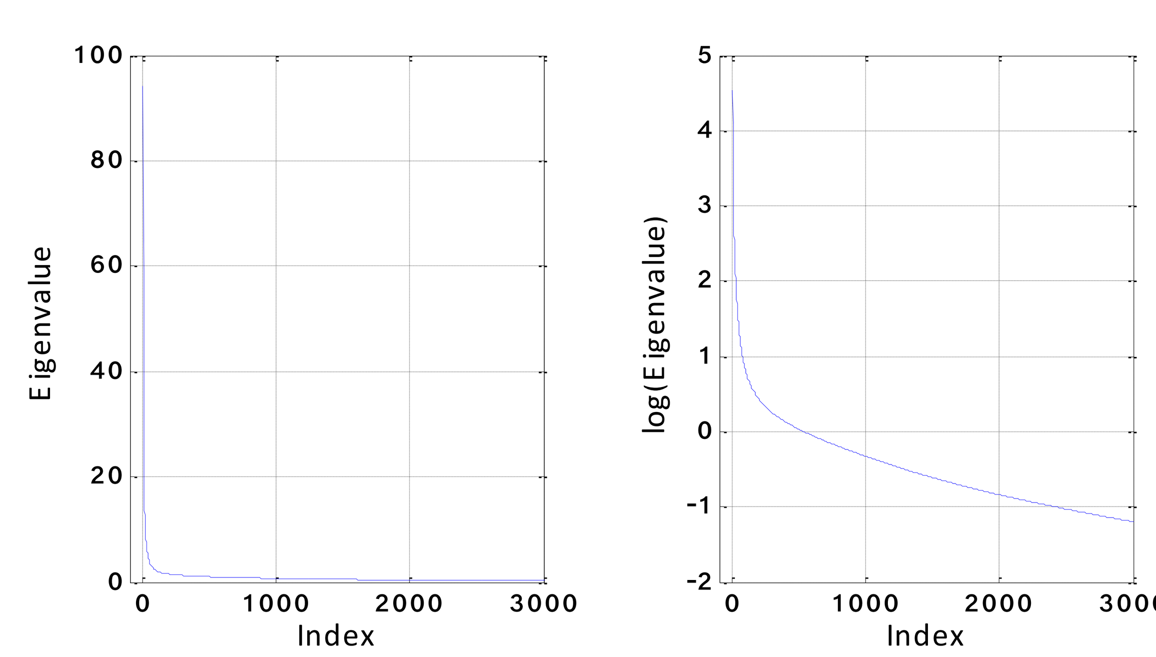


Fig. 1. The eigenvalues of the PCA on the centered GMM mean supervector space

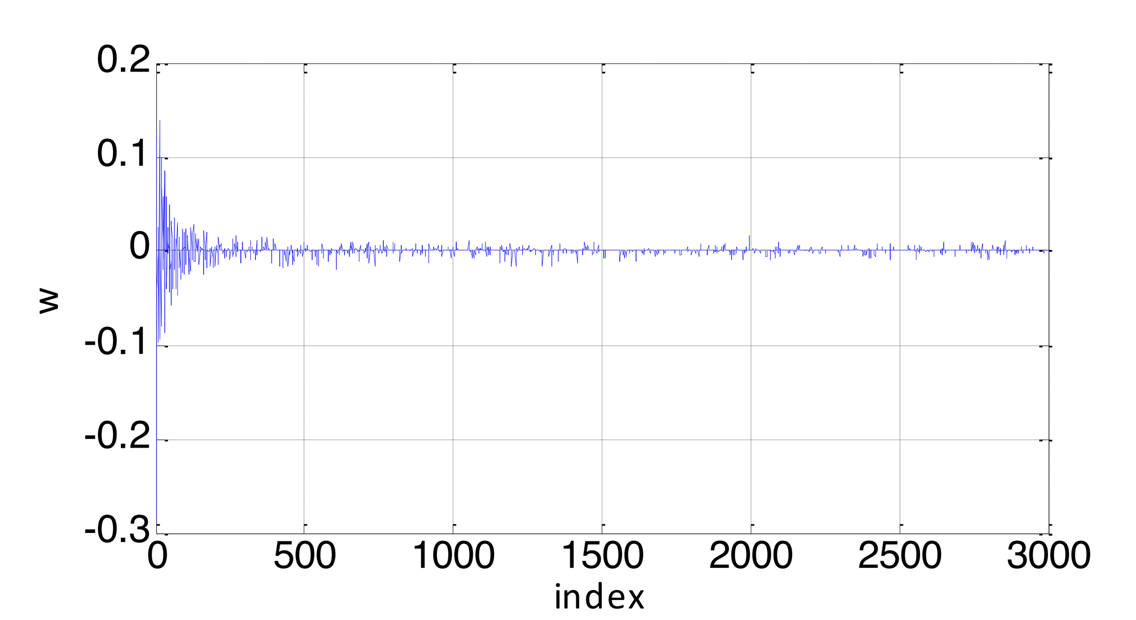


Fig. 2. The s-vector of utterance fzzhwB with $\tau = 6$. $\|\acute{\boldsymbol{w}}\|_0 = 753$, $\|\acute{\boldsymbol{w}}\|_1 = 6$

Experiments

Corpus

Test: the female part of common condition 5 in the NIST 2010 core task (female 05.nve-nve.phn-phn)

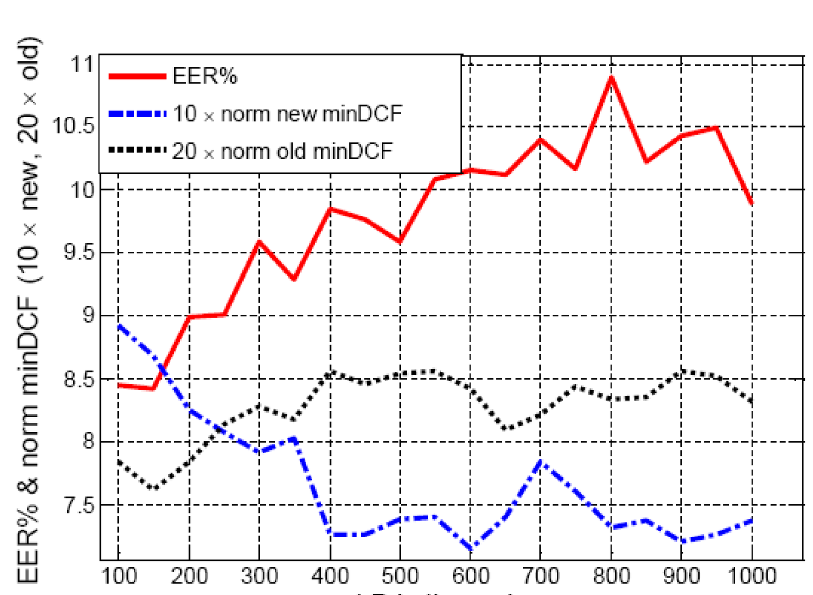
Table: Corpora used to estimate the UBM, total variability matrix, JFA factor loading matrix, WCCN, LDA, PLDA and the normalization data for tel-tel task.

	Switchboard	NIST04	NIST05	NIST06	NIST08
UBM		√	√		
T	√	√	√	√	√
JFA V	√				
JFA U		√	√	√	√
JFA D		√			
WCCN		√	√	√	√
LDA		√	√	√	√
PLDA		√	√	√	√
Znorm		√	√		
Snorm					√
Tnorm				√	

Parameters

- Czech phoneme recognizer based VAD (*Schwarz et al. 2006*)
- 18 MFCC coefficients and their first derivatives
- 25ms Hamming window with 10ms shifts
- Feature warping is applied
- 1024 GMM, diagonal
- JFA baseline uses linear scoring(*Glembek et al. 2009*)
  $LLR = (\boldsymbol{V}\boldsymbol{y}_{trn} + \boldsymbol{d}\boldsymbol{z}_{trn})^t\Sigma^{-1}(\boldsymbol{F}_{tst} - \boldsymbol{N}_{tst}\boldsymbol{U}\boldsymbol{x}_{tst})$
- JFA speaker factor size (300), channel factor size (100)
- I-vector baseline adopts LDA, WCCN for variability compensation and cosine distance for scoring
- I-vector factor size (400)
- SPGL toolkit for Lasso calculation.

S-vector LDA (female 05.nve-nve.phn-phn)



S-vector PLDA

Table: Performance of the s-vector system using PLDA modeling (female 05.nve-nve.phn-phn)

ID	τ	LDA	WCCN	PLDA			Snorm	EER%	norm minDCF	
				U	*G*	#EM			new	old
1	6	×	×	×	×	×	×	15.23	0.96	0.65
2	6	600	×	×	×	×	×	10.16	0.72	0.42
3	6	600	√	×	×	×	×	11.79	0.71	0.44
4	6	600	×	100	100	10	×	11.22	0.97	0.55
5	6	100	×	100	100	10	×	8.97	0.85	0.43
6	6	×	×	100	100	10	×	11.02	0.90	0.52
7	6	×	×	100	100	20	×	8.44	0.83	0.42
8	6	×	×	100	100	20	√	4.80	0.62	0.27
9	8	×	×	100	100	20	√	4.86	0.65	0.26
10	6	×	×	150	50	20	×	8.73	0.75	0.41
11	6	150	×	150	50	20	√	7.61	0.85	0.34
12	**6**	×	×	**150**	**50**	**20**	√	**4.83**	**0.55**	**0.28**

S-vector linear fusion (female 05.nve-nve.phn-phn)

Table: Performance of the s-vector system when fusing with the JFA and i-vector baseline systems

ID	Systems	EER(%)	norm minDCF	
			new	old
1	JFA linear scoring ZTnorm	3.62	0.41	0.193
2	I-vector LDA WCCN Cosine Snorm	5.04	0.52	0.241
3	S-vector PLDA Snorm	4.83	0.55	0.277
4	Fusion JFA + S-vector	3.37	0.38	0.177
5	Fusion I-vector + S-vector	4.14	0.47	0.215
6	Fusion JFA + I-vector + S-vector	3.09	0.37	0.157

S-vector linear fusion (male 05.nve-nve.phn-phn)

Table: Performance of the s-vector system when fusing with the JFA baseline systems

ID	Systems	EER(%)	norm minDCF	
			new	old
1	JFA linear scoring ZTnorm	3.69	0.33	0.148
2	S-vector PLDA Snorm	5.19	0.51	0.263
3	Fusion JFA + S-vector	2.85	0.31	0.114

Combine Female and Male, 05.nve-nve.phn-phn:
EER: 3.39, new cost 0.36, old cost 0.138

mingli@usc.edu. This work was supported in part by US Army and NSF.

Ming Hsieh Institute
Ming Hsieh Department of Electrical Engineering