I-VECTOR/PLDA SPEAKER RECOGNITION USING SUPPORT VECTORS WITH DISCRIMINANT ANALYSIS

Fahimeh Bahmaninezhad and John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
Department of Electrical Engineering
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.

IEEE ICASSP-2017
March 5-9, 2017 New Orleans, USA

Outline

Background

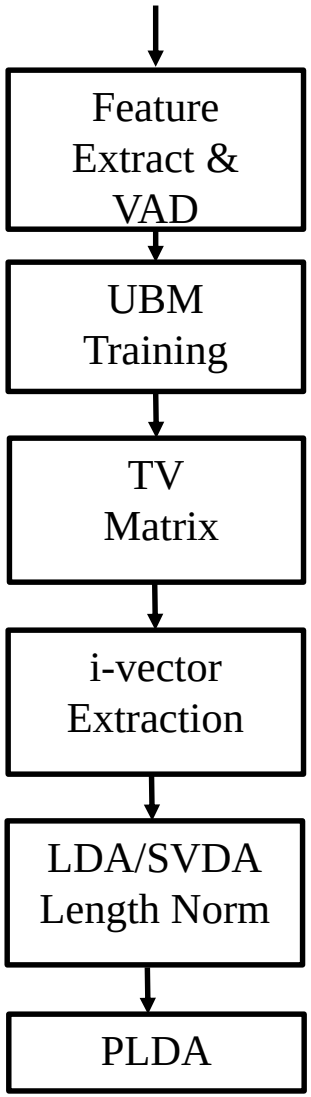
- Traditional i-Vector/PLDA speaker recognition uses LDA for dimension reduction.
- Discriminant analysis can be used to compensate mismatch.
- Our previous study of Generalized Discriminant Analysis (GDA) in speaker recognition [1] did not have consistent improvement over LDA (evaluation was on NIST SRE2010 task).

[1] **F. Bahmaninezhad, J.H.L. Hansen, "Generalized Discriminant Analysis (GDA) for Improved i-Vector Based Speaker Recognition,"** ***ISCA INTERSPEECH, 2016***.

Proposition

Discriminant analysis via support vectors (SVDA) is used instead of LDA for the NIST SRE2010 task.

Speaker Recognition



- MFCC Features and energy-based VAD are used.
- Speaker and channel dependent supervectors in the i-vector configuration are factorized as:

$$M = m + Tw$$

- UBM and TV matrix are trained using the Expectation Maximization (EM) algorithm.
- In the E-step, w is considered latent variable with normal prior distribution. The i-Vectors are calculated as posterior distributions of w:

$$\hat{w}(u) = (I + T^T \Sigma^{-1} N(u) T)^{-1} T^T \Sigma^{-1} S(u)$$

- In the recognition phase, given two i-Vectors, we need to determine if the two i-Vectors belong to the same speaker or not.

$$\log - likelihood = \log \frac{p(\hat{w}_1, \hat{w}_2 \mid \text{target})}{p(\hat{w}_1, \hat{w}_2 \mid \text{nontarget})}$$

Motivation

- Finds discriminatory directions based on the boundary structure of speaker classes.
  - LDA just focuses on the **CENTROID** of classes.
- Controls generalization in an easy way.
  - In solving SVM problem, we can adjust the tolerance of misclassification.
- Considers the small sample size problem.
  - For NIST2010 data, the number of samples in each class is smaller than dimension of data.
- Solves unbalanced problem partially.
  - Some classes have more than 90 samples, while some have less than 10.
- Compensates the domain-mismatch introduced in NIST SRE2016 challenge.
  - Using unlabeled in-domain data was effective in the challenge. With SVDA, we used them without any pseudo-labels.

Method

First, must train SVM classifier. Different strategies studied here:

- Traditional One-VS-One:
  - We need to train one SVM for each pair of speakers. Therefore, if C represents the number of speaker classes, we need C(C-1)/2 pairwise SVM models.
- Weighted One-VS-One:
  - In training the SVM, we observe that all samples from some classes are selected as the support vectors. This means these classes do not have enough samples. Therefore, we assign these classes a smaller weight for their contribution in calculating the between class covariance matrix.
- One-VS-Rest:
  - SVM will be trained for each class to discriminate it versus the remains. Therefore, we need to train “C” SVMs.
- **Linear SVM** is used to find the support vectors (LibSVM toolkit).

SVDA VS LDA

Similarities:

Class separation criterion (Fisher criterion) for both LDA and SVDA in direction of A is defined:

Between class covariance matrix

$$\lambda = \frac{A^T S_b A}{A^T S_w A}$$

Transformation matrix

Within class covariance matrix

The projection matrix A: contains the k eigenvectors corresponding to the k largest eigenvalues of: $S_w^{-1} S_b$

- T is the transpose operation.

SVDA VS LDA

Differences:

Definition of within and between class covariance matrices. SVDA only uses support vectors. **Within class and between class covariance matrices for LDA:**

Number of speaker classes

Number of samples in class c

$$S_b = \sum_{c=1}^{C} n_c (\mu_c - \mu)(\mu_c - \mu)^T$$

Overall mean of all samples

$$S_w = \sum_{c=1}^{C} \sum_{k \in c} (x_k - \mu_c)(x_k - \mu_c)^T$$

Mean of samples in class C

Sample in class c

SVDA

Within class and between class covariance matrices for SVDA:

$$S_b = \sum_{1 \le c_1 \le c_2 \le C} w_{c_1 c_2} w_{c_1 c_2}^T$$

$$= \sum_{i=1}^{l} y_i \alpha_i x_i$$

Optimal direction to classify two classes by linear SVM.

Target value: +1 for class one and -1 for class two.

Learning patterns

coefficients

$$S_w = \sum_{c=1}^{C} \sum_{i \in \hat{I}_c} (\hat{x}_i - \hat{\mu}_c)(\hat{x}_i - \hat{\mu}_c)^T$$

Mean of support vectors in class c.

Support vectors

Index of support vectors in class c.

System Description

System Setup		
Front-end	60-D MFCC (19 static + energy +Δ+ΔΔ)	
UBM and Total Variability (TV) matrix	**Data**	NIST SRE 2004, 2005, 2006, 2008, Switchboard II phase2 and phase3, Switchboard cellular part1 and part2.
	2048 mixture full-covariance UBMs.	
	600 dimensional TV space.	
Back-end	**Data**	Male speakers of NIST SRE 2004, 2005, 2006, 2008.
	LDA/SVDA, length normalization, Gaussian PLDA.	
	The dimension of i-Vectors is reduced to 400 with LDA/SVDA.	
Experiments	Male speakers of extended core (coreext) and core trials of NIST SRE 2010 (condition 5) are used for evaluations. Also, truncated coreext test utterances into 3s, 5s, 10s, 20s, 40s are also used for evaluation.	
i-Vector	600, reduced to 400 with LDA/SVDA	
Tools	Kaldi (sre10/v1)	

System Description

Statistics of data used for training models and system evaluations:

Enrolment/Test		Core/Core	Coreext/ 3,5,10,20,40,full
UBM-TV	Num. of **Speakers**	**5756**	**5756**
	Num. of **Segments**	**57273**	**57273**
LDA/SVDA/PLDA	Num. of **Speakers**	**1115**	**1115**
	Num. of **Segments**	**13605**	**13605**
Enrollment	Num. of **Models**	**2426**	**5237**
Trials	Num. of **Target**	**353**	**3465**
	Num. of **Non-Target**	**13707**	**175873**

UBM and TV matrix are trained using both male and female speakers, while LDA, SVDA, and PLDA are trained using just male speakers. Trials are also limited to the male speakers of NIST SRE2010.

SRE 2010 Results

EER (%) results comparing LDA and SVDA. Dimension of i-Vectors is reduced from 600 to 400.

Enrollment/Test	LDA	SVDA traditional 1-vs-1	SVDA weighted 1-vs-1	SVDA 1-vs-rest
Core/Core	1.66	1.13	1.25	1.42
Coreext/Coreext	1.5	1.35	1.3	1.39
Coreext/Coreext3s	14.5	14.22	14.23	14.2
Coreext/Coreext5s	9.71	9.64	9.55	9.81
Coreext/Coreext10s	5.61	5.58	5.6	5.72
Coreext/Coreext20s	3.17	3.12	3.17	3.35
Coreext/Coreext40s	2.48	2.4	2.37	2.42

EER is improved by **25%** and **32%** respectively with **weighted** and **traditional** 1-vs-1 SVDA.

SRE 2010 Results

minDCF (x 100) results comparing LDA and SVDA. Dimension of i-Vectors is reduced from 600 to 400.

Enrollment/Test	LDA	SVDA traditional 1-vs-1	SVDA weighted 1-vs-1	SVDA 1-vs-rest
Core/Core	3.7	3.99	3.64	3.68
Coreext/Coreext	2.97	3.08	2.87	2.9
Coreext/Coreext3s	9.84	9.74	9.74	9.75
Coreext/Coreext5s	9.24	9.15	9.09	9.2
Coreext/Coreext10s	7.59	7.49	7.37	7.6
Coreext/Coreext20s	5.85	5.74	5.73	5.93
Coreext/Coreext40s	4.48	4.23	4.07	4.11

minDCF is improved by **9%** and **5.6%** respectively with **weighted** and **traditional** 1-vs-1 SVDA.

SRE 2010 Results

Speaker recognition results comparing LDA and SVDA without dimension reduction.

- **EER (%):**

Enrollment/Test	LDA	SVDA traditional 1-vs-1	SVDA weighted 1-vs-1
Core/Core	1.58	1.45	1.46
Coreext/Coreext	1.46	1.37	1.36

- **minDCF (x100):**

Enrollment/Test	LDA	SVDA traditional 1-vs-1	SVDA weighted 1-vs-1
Core/Core	3.9	3.8	4
Coreext/Coreext	3.02	3.01	3.02

SRE 2016 Results

- **Single system submissions from CRSS: Sub-system 6 is the best single system which uses SVDA. Results for evaluation set:**

Sub-System	1	2	3	4	5	6	7	8	9	10	11
EER (%)	14.93	12.94	15.08	14.15	12.42	10.66	10.91	20.93	21.66	20.34	21.41
Min-Cprimary	0.846	0.766	0.837	0.826	0.797	0.698	0.719	0.895	0.918	0.956	0.96
Act-Cprimary Dev+Unlabeled	0.931	0.854	0.902	0.905	0.998	0.813	0.788	0.902	0.919	0.957	0.963
Act-Cprimary Dev	1.286	0.799	0.999	1.29	1.593	0.933	0.83				
Act-Cprimary Unlabeled	0.858	0.776	0.838	0.831	0.819	0.7	0.733				

DNN UBM Unlabeled-PLDA

Primary fused system (only 7 first sub-systems, with Dev+Unlabeled calibration): **EER: 9.37, min-Cprimary: 0.646, act-Cprimary: 0.708**.

C. Zhang, F. Bahmaninezhad, S. Ranjan, C. Yu, N. Shokouhi, J.H.L. Hansen, "UTD-CRSS Systems for 2016 NIST Speaker Recognition Evaluation," will be submitted to *ISCA INTERSPEECH, 2017*.

Discussions

- SVDA **consistently** improves LDA for both **dimension reduction** and **discrimination**.
- Weighted 1-vs-1 approximately outperforms traditional 1-vs-1. Therefore, removing classes that do not have enough samples are helpful.
- **Traditional 1-vs-1** improves equal error rate (EER) and minimum detection cost function (minDCF) by **32%** **and** **5.6%** respectively.
- **Weighted 1-vs-1** improves EER and minDCF by **25%** **and** **9%** respectively.
- 1-vs-1 performs better than 1-vs-rest; since, 1-vs-1 does not have imbalanced problem.

Conclusion

Outcomes

- SVDA has been applied to speaker recognition problem instead of traditional LDA.
- Results confirm consistent improvement for both short and long duration utterances for NIST SRE2010 task.
- Linear SVM with different strategies have been compared.
- Also used for NIST SRE2016, and shown to be effective for domain mismatch

Future work

- Improvement for short segment is negligible. Using some small samples in training SVM and giving them more weight in the calculation of between class scatter can be effective.
- SVDA for other conditions of NIST SRE 2010 (not just limited to condition5 and male speaker) will be evaluated later.