Unsupervised Domain Adaptation for Gender-Aware PLDA Mixture Models

Longxin Li and Man-Wai MAK

Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University

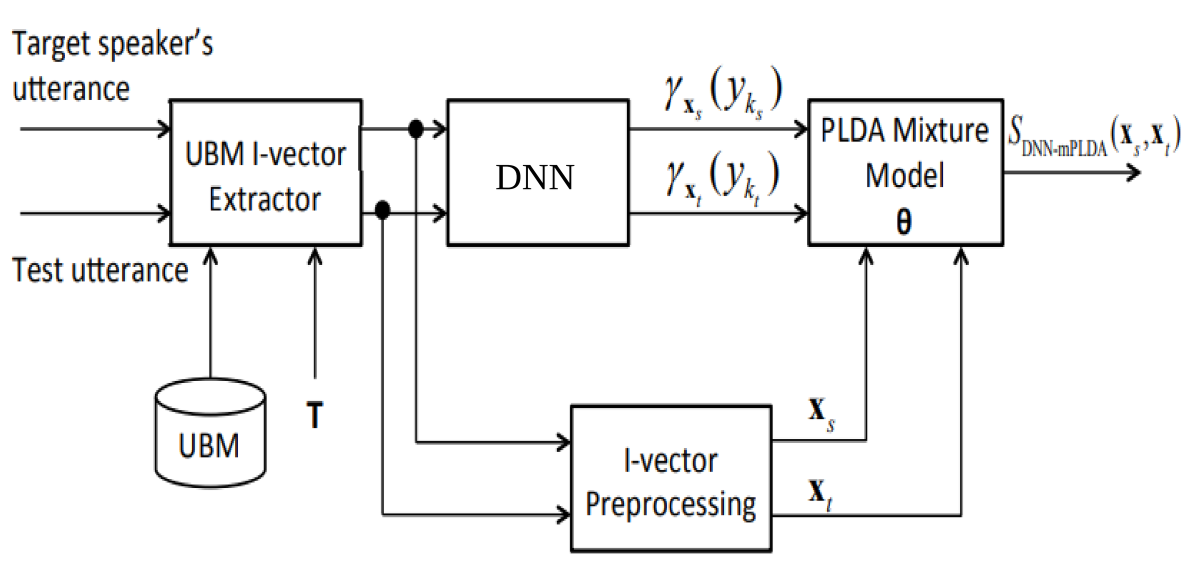
Introduction

- PLDA is still problematic when (1) the model is deployed to new environment (in-domain) that is very different from the training one (out-of-domain) and (2) there are insufficient labeled data from the new environment.
- This paper proposes using out-of-domain training data to pre-train a PLDA mixture model and applying the mixture model on the in-domain training data to compute a pairwise score matrix for spectral clustering. The hypothesized speaker labels produced by spectral clustering are then used for re-training the mixture model to fit the new environment.
- Experiments on NIST 2016 SRE demonstrate the effectiveness of the proposed framework compared with agglomerative hierarchical clustering (AHC).

Background

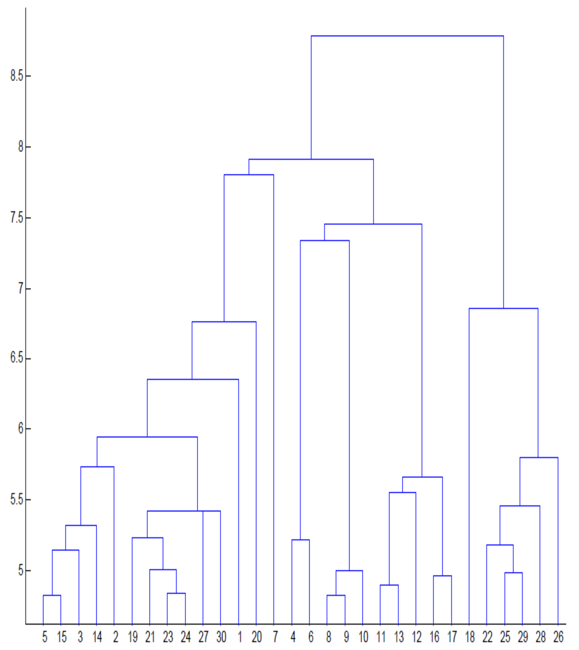
DNN-driven mixture of PLDA (DNN-mPLDA):

$$p(\mathbf{x}_{ij}) = \sum_{k=1}^{K} \gamma_{\mathbf{x}_{ij}}(\mathrm{y}_{ijk}) N(\mathbf{x}_{ij} \mid \mathbf{m}_k, \mathbf{V}_k\mathbf{V}_k^{\mathrm{T}} + \mathbf{\Sigma}_k)$$

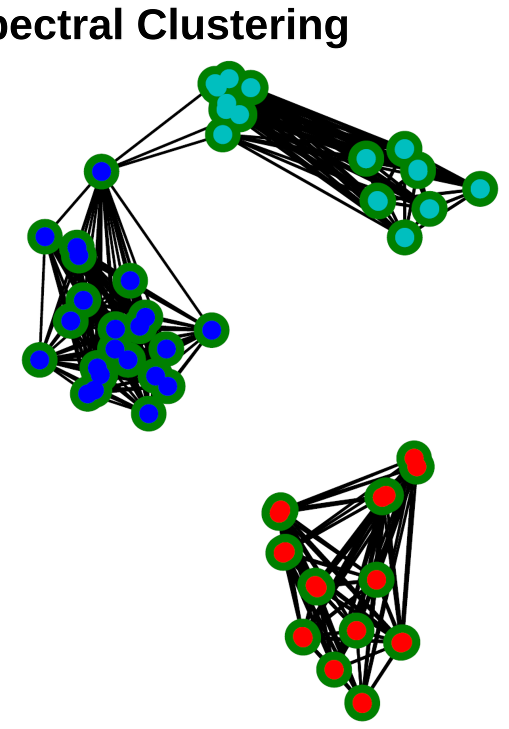


DNN-mPLDA.

AHC



Spectral Clustering



Spectral Clustering of I-Vectors

Step1 Compute a pairwise PLDA score matrix **S** from n training i-vectors:

$$s_{ij} = S_{\mathrm{mPLDA}}(\boldsymbol{x_i}, \boldsymbol{x_j}), \qquad i,j = 1, \dots, n.$$

Step 2 Convert **S** to a adjacency matrix **A** with elements:

$$a_{ij} = \begin{cases} \exp\left\{-\frac{(s_{\mathrm{amax}} - s_{ij})^2}{2\sigma^2}\right\} & i \neq j \\ 1 & \text{otherwise} \end{cases}$$

where s_{amax} is the absolute maximum in **S**.

Step 3 Compute a Laplacian matrix:

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$$

where **D** is a diagonal matrix with elements $d_{ii} = \sum_{j=1}^{n} a_{ij}$.

Step 4 Pack K eigenvectors of **L** with the smallest eigen-values to form $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_K] \in \mathbb{R}^{n \times K}$.

Step 5 Normalize the row of **V**:

$$v_{ij} \leftarrow \frac{v_{ij}}{\sqrt{\sum_j v_{ij}^2}}$$

Step 6 Apply K-means to the n rows of **V**.

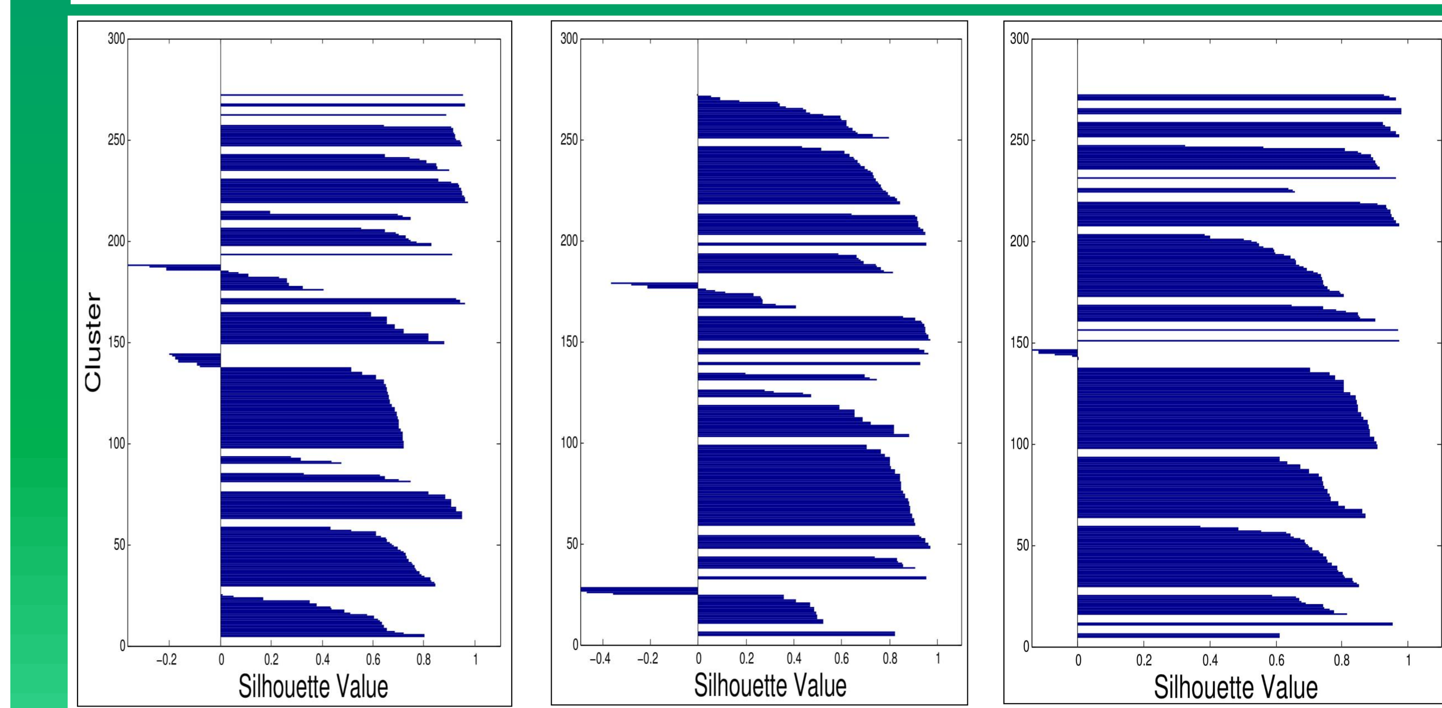
Cluster Quality

- Silhouette values is used to quantify the quality of clusters. Each sample has a Silhouette value

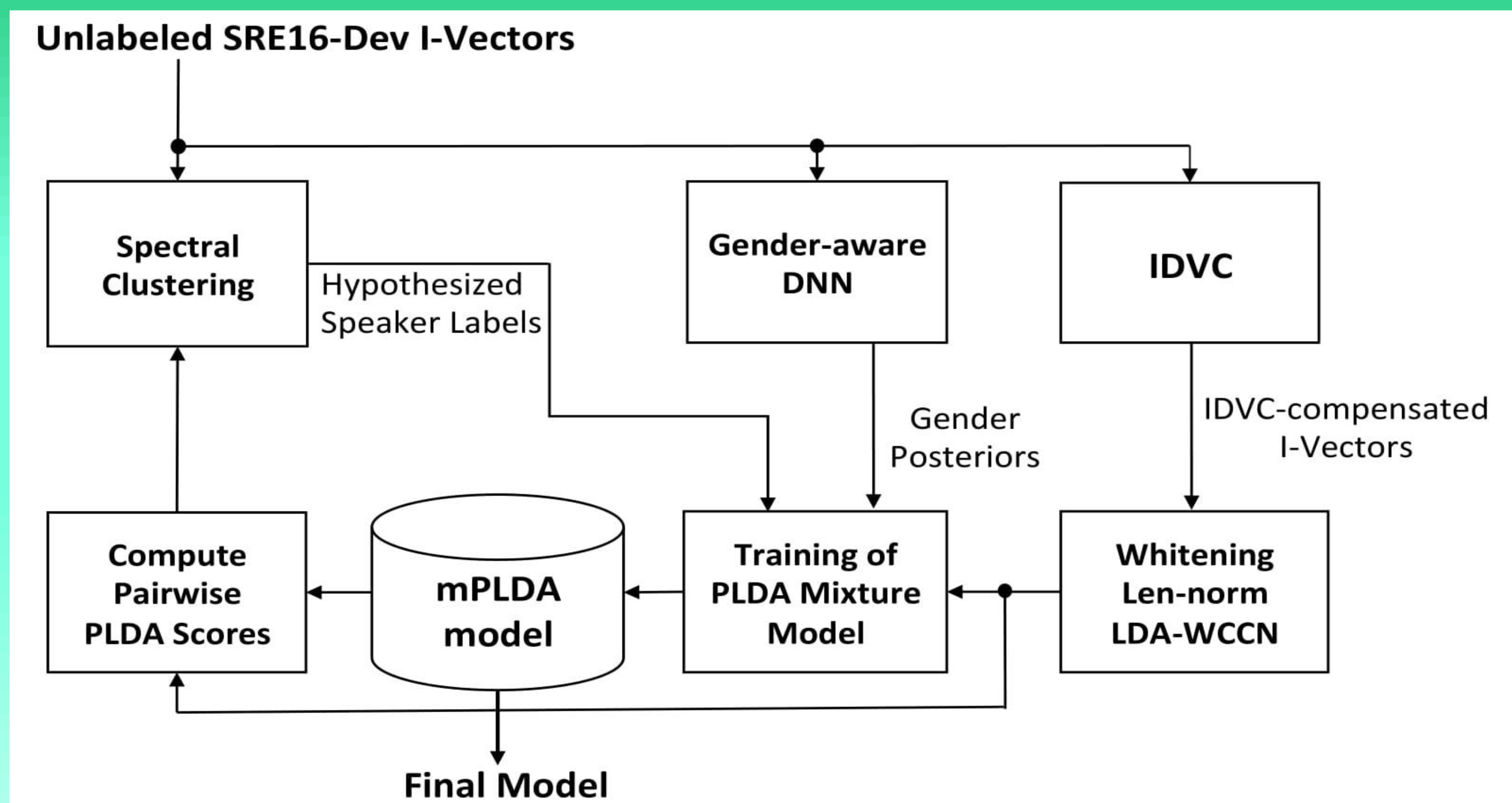
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average dissimilarity of sample i with respect to other samples in the same cluster and $b(i)$ is the lowest average dissimilarity of sample i with respect to any other cluster not containing i.

- $s(i) = +1$ ⇨ sample i is well matched to its own cluster
- $s(i) = -1$ ⇨ sample i is assigned to the wrong cluster
- Results show that Iterative-SC has
  - **Highest average** Silhouette score
  - **Less negative** Silhouette scores
- So, Iterative-SC produces clusters with better quality



AHC-Euclid AHC-Cosine Iterative-SC



Results

- *Performance of the iterative retraining method for different numbers of iterations on SRE16-dev and SRE16-eval*

	SRE16-Dev		SRE16-Eval	
Iteration	EER(%)	minDCF	EER(%)	minDCF
1	17.12	0.812	18.72	0.952
2	16.31	0.789	15.32	0.883
3	15.79	**0.751**	13.62	0.829
4	15.68	0.774	12.79	0.798
5	**15.04**	0.799	**12.73**	**0.779**
6	15.74	0.782	13.03	0.792
7	15.79	0.788	13.34	0.801

- *Performance of PLDA mixture models on SRE16 using different speaker clustering methods and with and without covariance matrix interpolation (Cov. Interp.)*

Row	Clustering Method	Followed by Cov. Interp.	SRE16-Dev		SRE16-Eval	
			EER(%)	minDCF	EER(%)	minDCF
1	Euclid-AHC	N	19.54	0.937	18.68	0.932
2	Cosine-AHC	N	18.23	0.862	16.37	0.846
3		Y	16.36	0.818	14.12	0.832
4	Iterative-SC	N	**15.04**	**0.799**	12.73	**0.779**
5		Y	15.21	0.809	**12.60**	0.816

References:

- N. Li, M. W. Mak, and J. T. Chien, "DNN-driven mixture of PLDA for robust speaker verification," IEEE/ACM Transactions on Audio, Speech, and Language Processing, no. 6, pp.1371–1383, 2017.
- M. W. Mak, X. M. Pang, and J. T. Chien "mixture of PLDA for noise robust i-vector speaker verification," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 24, no. 1, pp. 130–142, 2016.
- Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in Proc. ICASSP, pp. 1695-1699, 2014.
- H. Aronowitz, "Inter dataset variability modeling for speaker recognition," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2017, pp. 5400–5404.