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, Hong Kong SAR

lkqllx@gmail.com, enmwmak@polyu.edu.hk

ABSTRACT

Probabilistic linear discriminant analysis (PLDA) is a state-of-art back-end for i-vector based speaker verification. However, this back-end 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. To address these problems, 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. To refine the mixture model, the spectral clustering and re-training processes are repeated a number of times. To make the mixture model amenable to both genders, a deep neural network (DNN) is trained to produce gender posteriors given an i-vector. The gender posteriors then replace the posterior probabilities of the indicator variables in the PLDA mixture model. Evaluations based on NIST 2016 SRE suggest that at the end of the iterative re-training, the PLDA mixture model becomes fully adapted to the new domain. Results also show that the PLDA scores can be readily incorporated into spectral clustering, resulting in high quality speaker clusters that could not be possibly achieved by agglomerative hierarchical clustering.

Index Terms— I-vectors; DNN-driven mixture of PLDA; spectral clustering; domain adaptation; speaker verification

1. INTRODUCTION

I-vectors [1] have been regarded as the best feature representation for speaker verification. To achieve good performance, a robust back-end that can minimize the effect of the unwanted variabilities in i-vectors is essential. So far, probabilistic linear discriminant analysis (PLDA) [2] is still the best back-end for this purpose. Given the i-vectors of a target speaker and a claimant, the likelihood ratio between the same-speaker hypothesis and different-speaker hypothesis is computed from a PLDA model. During the computation of the marginal likelihood of these two hypotheses, the unwanted variabilities in the i-vectors are marginalized out.

Despite its remarkable performance, PLDA models require a large amount of speech data with speaker labels for training. In particular, to model the speaker subspace reliably, each speaker in the training set should have multiple sessions, preferably collected by different microphones. Most of the current speech corpora (e.g., Switchboard, Fisher, and Mixer) focus on English telephone speech. Therefore, training a reliable PLDA model for English telephone speech is not an issue. However, other languages or acoustic environments may not have such rich resources. Even if we have the speech data of other languages, we may not have the speaker labels. The NIST 2016 SRE [3] has exactly such situation. In this evaluation, participants were given *unlabeled* speech data for training whatever models for suppressing the channel, language and gender variabilities.

As training of a PLDA model requires speaker labels, one sensible approach is to apply unsupervised clustering on the i-vectors derived from the in-domain data to produce some hypothesized speaker labels. Agglomerative hierarchical clustering [4] can be used for such purpose. Alternatively, spectral clustering [5–7], which utilizes the eigenvectors of a similarity matrix, can be used. The similarity matrix can be derived from the pairwise PLDA scores of in-domain i-vectors.

Beside speaker information, genders and languages are another two crucial characteristics of human voice. Male and female possess different vocal-tract structures, which induce different voice characteristics for the two genders [8]. If gender information is not available during scoring, a gender classifier can be used as a front-end for the gender-dependent systems. Again, i-vectors can be used as the features of this classifier because they contain gender information. A better approach is to jointly train the gender-dependent PLDA models using the data from both genders. This leads to a gender-aware PLDA mixture model, which is the key contribution of this paper.

Speaker verification systems also need to deal with language mismatch. In particular, a system trained on one language (e.g. English) will have difficulty in distinguishing speakers speaking another language (e.g. Mandarin). To suppress the effect of language differences on i-vectors, an approach called inter dataset variability compensation (IDVC) [9–11] can be used to estimate the nuisance subspace and remove the subspace from all i-vectors. To apply IDVC, we may separate the training dataset into disparate groups according to genders and languages and compute the mean i-vector of each group. A low-dimension subspace is then obtained by applying principle component analysis (PCA) on the means. The variability caused by gender and language differences is then projected out based on this low-dimensional subspace.

In light of the promise of spectral clustering and mixture of PLDA [12, 13], this paper proposes a novel method for adapting a gender-aware PLDA mixture model to a new domain using a small amount of unlabelled data in the new domain. The key idea is to incorporate the pairwise PLDA scores produced by an initial mixture model into the spectral clustering process so that the resulting hy-

This work was in part supported by The RGC of Hong Kong SAR, Grant No. PolyU 152518/16E and PolyU 152137/17E.

pothesized speaker labels can be used for iteratively refining the mixture model. The method was evaluated on the NIST 2016 Speaker Recognition Evaluation corpus (SRE16-eval) using its development set (SRE16-dev) as the unlabelled data for domain adaptation. Surprisingly, it was found that only the SRE16-dev data are enough for this iterative refinement process to achieve good performance. By analyzing the Silhouette values of the clusters produced by spectral clustering (SC) and agglomerative hierarchical clustering (AHC), we found that SC is much better than AHC for hypothesizing the speaker labels. In particular, SC not only provides an efficient way of using the PLDA scores for speaker clustering, it also produces consistent clusters as compared to AHC.

2. BACKGROUND

2.1. Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering (AHC) is a bottom-up clustering algorithm. Given a collection of data points, the AHC algorithm starts by considering each data point in a dataset as a cluster. Then, at each step, it finds a pair of clusters that are *closest* to each other and merges the two clusters to form a bigger one. The process is repeated until there is only one cluster remains. AHC is a popular approach to hypothesizing the speaker labels of in-domain datasets for unsupervised domain adaptation. For example, Garcia-Romero [14] proposed using the scores derived from an out-of-domain PLDA model as the similarity metric and merging two clusters if their average pairwise score is larger than a threshold. The merging process stops when the evidence in favour of the same-speaker hypothesis is higher than that of the different-speaker hypothesis. The hypothesized speaker labels are then used for estimating the covariance matrices of an in-domain PLDA model, which are then interpolated with the covariance matrices of the out-of-domain PLDA model [15]. Alternatively, Torres-Carrasquillo *et al*. [16] suggested using AHC to cluster the unlabeled data in the SRE16-dev and applied the hypothesized speaker labels for supervised domain adaptation similar to [15].

2.2. Silhouette Values

In cluster analysis, silhouette values are measures that quantify the coherence of data within a cluster. It was first introduced by Rousseeuw [17] as a graphical tool for interpreting and validating clusters in cluster analysis. Using the notations in silhouette's literature, we denote $a(i)$ as the average dissimilarity (distance) of sample i with respect to all other samples within the same cluster. Also, we denote $b(i)$ as the lowest average dissimilarity of sample i with respect to any other cluster not containing i. Then, the silhouette value of sample i is given by

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \tag{1}$$

Obviously, $s(i)$ ranges from -1 to $+1$. A value of $+1$ means that the corresponding data sample is well matched to its own cluster and is poorly matched to its rival cluster. On the other hand, a value of -1 means that the data sample is assigned to the wrong cluster. To ensure proper clustering of data, we strive for having positive silhouette values for all data samples or having an average silhouette value close to +1.

The silhouette values depend on the similarity/dissimilarity metrics. In speaker clustering, these metrics can be derived from the Euclidean distances, cosine distances, and PLDA scores (see Section 3.1) between pairs of i-vectors. For example, in [18], silhouette values based on the Euclidean distance between i-vectors were used for determining which clusters should be merged during the speaker clustering process.

3. PROPOSED FRAMEWORK

3.1. Hypothesized Speaker Labels

In the proposed domain adaptation method, spectral clustering is the key step for hypothesizing the speaker labels in the in-domain data for iterative retraining of the PLDA mixture model. To perform spectral clustering, a similarity matrix comprising the similarity scores between each pair of the training i-vectors is needed. The similarity matrix can be obtained from the PLDA scores of training utterances. As PLDA scores are log-likelihood ratios, they can be negative. Therefore, we need to convert the PLDA scores to similarity scores that are amenable to spectral clustering.

Given a dataset $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ comprising n i-vectors, we compute a PLDA score matrix $\mathbf{S} \in \Re^{n \times n}$, where the element s_{ij} of $\mathbf{S}$ is the score of $\mathbf{x}_i$ and $\mathbf{x}_j$ based on a PLDA mixture model [12]:

$$s_{ij} = S_{\text{mPLDA}}(\mathbf{x}_i, \mathbf{x}_j).$$

Then, we convert $\mathbf{S}$ to a distance matrix $\mathbf{M}$ with elements:

$$m_{ij} = \begin{cases} s_{\text{amax}} - s_{ij} & i \neq j \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where

$$s_{\text{amax}} = \max_{i,j;\ i \neq j} |s_{ij}|. \tag{3}$$

Then, we convert the distance matrix $\mathbf{M}$ to a similarity matrix $\mathbf{A}$ that is suitable for spectral clustering. Specifically, the element of $\mathbf{A}$ is

$$a_{ij} = \exp\left\{-\frac{m_{ij}^2}{2\sigma^2}\right\}, \tag{4}$$

where σ is a scaling parameter that controls how fast the similarity drops with the distance m_{ij}. The similarity of two i-vectors reflects the "distance" or difference between the two utterances. A negative s_{ij} means that the two i-vectors are very dissimilar, which results in a large m_{ij} in Eq. 2 and small a_{ij} in Eq. 4. On the other hand, a large s_{ij} means that the two i-vectors are very similar, which results in $m_{ij} \approx 0$ and $a_{ij} \to 1$.

With the simiarlity matrix $\mathbf{A}$, we may divide $\mathcal{X}$ into K clusters as follows. First, we compute the Laplacian matrix

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

where $\mathbf{I}$ is an $n \times n$ identity matrix, $\mathbf{D}$ is a diagonal matrix with diagonal elements $d_{ii} = \sum_{j=1}^{n} a_{ij}$, and $\mathbf{D}^{-\frac{1}{2}}$ stands for the inverse of the square root of $\mathbf{D}$. Then, we compute the K eigenvectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$ of $\mathbf{L}$ with the smallest eigenvalues and pack the K eigenvectors to form a matrix $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_K] \in \Re^{n \times K}$, followed by re-normalization of the rows:

$$v_{ij} \leftarrow \frac{v_{ij}}{\sqrt{\sum_j v_{ij}^2}}. \tag{5}$$

Then, we consider the rows of the normalized $\mathbf{V}$ as K-dimensional vectors and the n row vectors can be clustered by K-means to form

K clusters. The row vectors and their corresponding utterances in the c-th cluster ($c = 1, \ldots, K$) are considered to be associated with the c-th hypothesized speaker.

3.2. Gender and Language Mismatch Compensation

To suppress the effect of gender and language mismatch in the i-vectors, we applied inter-dataset variability compensation (IDVC) [11]. IDVC aims to find a low-dimensional subspace that is sensitive to the mismatches and remove this subspace from both the development and evaluation i-vectors. To this end, we partitioned SRE16-dev into 4 subsets (two per gender).[1] Principal component analysis was then applied to the mean i-vectors of these subsets to find the first 3 eigenvectors $\{\mathbf{u}_r\}_{r=1}^{3}$ with the largest eigenvalues. All i-vectors were than projected by the transformation matrix $(\mathbf{I}-\mathbf{U}\mathbf{U}^{\mathsf{T}})$, where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]$.

3.3. DNN-Driven PLDA Mixture Model

While IDVC is capable of projecting out the gender variability in i-vectors, using only the mean i-vectors to estimate the nuisance directions is rather crude in that other statistical properties of i-vectors are simply ignored. We propose using a DNN-driven PLDA mixture model [12] to model the remaining gender information in the i-vectors. Because there are only two genders, the number of mixtures was set to 2. During the training phase, the gender (mixture) posteriors are provided by a gender-aware DNN that receives i-vectors as input and the *hypothesized* speaker labels are provided by spectral clustering. Fig. 1 shows the block diagram of the training process.

As shown in Fig. 1, at the beginning of training, we need a PLDA mixture model to compute the first set of pairwise scores for spectral clustering. This can be accomplished by using SRE05–SRE12 data (with real speaker labels) to train an initial PLDA mixture model. Once we have the first set of hypothesized speaker labels, the mixture model can be retrained by using the gender posteriors and IDVC-compensated vectors, and the process can be repeated for a number of iterations. The EM algorithm for training the PLDA mixture model can be found in [12]. Note that while the mixture model is gender-aware, the two mixture components are gender-dependent. Therefore, each component has its own mean vector $\mathbf{m}_k$, speaker subspace $\mathbf{V}_k$ and residue covariance matrix $\mathbf{\Sigma}_k$, i.e., the mixture model is parameterised by $\theta = \{\mathbf{m}_k, \mathbf{V}_k, \mathbf{\Sigma}_k\}_{k=1}^{2}$. Note also that the SRE05–SRE12 data are only used for initializing the mixture model; during the iterative training process, only the SRE16-dev data are used. Therefore, the training process is efficient.

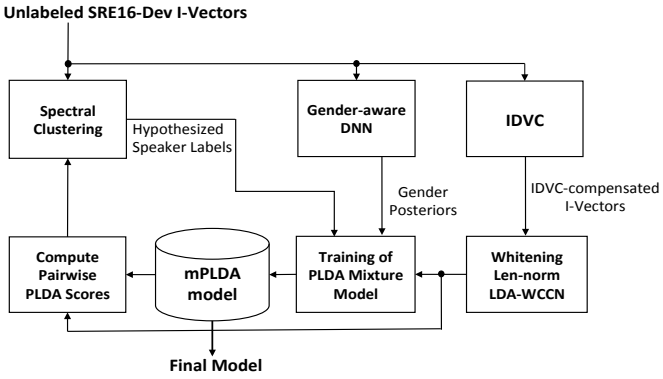


Fig. 1. Iterative retraining of the gender-aware PLDA mixture model

[1] As there is no language label except for the type "major" and "minor" in the directory structure, each gender can only be divided into two groups.

After the iterative training process, the PLDA mixture model is ready for scoring. Given the i-vectors of a target-speaker and a claimant, their gender posteriors are computed by the gender-aware DNN. The i-vectors are subject to IDVC, followed by i-vector preprocessing (whitening + len-norm + LDA-WCCN) [19, 20] before presenting to the mixture model to compute the PLDA score. The scoring formula can be found in [12]. In this work, we applied WCCN transformation to whiten the i-vectors [21].

4. EXPERIMENTS

4.1. Evaluation Protocol and Speech Data

Evaluations were performed on the evaluation set of NIST 2016 SRE (SRE16-eval) [3]. Data from the development set of SRE16 (SRE16-dev) and from SRE05–SRE12 were used for development. The data were divided into the following parts:

- *Enrollment and Test Data*: SRE16-dev has 120 enrollment segments, each with approximately 60s. It also contains 1,207 test segments with duration ranging from 10s to 60s. All segments contain telephone conversations spoken by 20 subjects in either Mandarin or Cebuano. Each target speaker has one or three enrollment segments. The evaluation protocol in SRE16-dev defines which target-speaker models should score against which test segments, with a total of 4,829 target trials and 19,312 non-target trials. SRE16-eval has the same structure as SRE16-dev, excepting that the numbers of enrollment segments and test segments increase to 1202 and 9,294, respectively. The number of subjects also increases to 201. The evaluation protocol defines 37,063 target trials and 1,949,666 non-target trials. Also, unlike SRE16-dev, all enrollment and test segments in SRE16-eval were spoken in either Cantonese or Tagalog, which causes language mismatch for systems trained on SRE16-dev data.
- *Development Data*: Telephone segments from SRE05–SRE12 were used for training the gender-aware DNN and the initial PLDA mixture model in Fig. 1. The unlabelled data in SRE16-dev, including the major and minor languages, were used for training the subspace projection matrices (LDA and WCCN), a 512-mixture UBM, and a 300-factor total variability matrix. They were also used for the iterative retraining of the PLDA mixture model in Fig. 1.

For each speech segment, a 2-channel voice activity detector [22] was applied to remove silence regions. Then, the speech regions were segmented into 25-ms Hamming windowed frames with 10ms frame shift. For each frame, 19 Mel frequency cepstral coefficients and log energy together with their first and second derivatives are packed to form a 60-dimensional acoustic vector, followed by cepstral mean normalization and feature warping [23] with a window size of 3 seconds.

4.2. Training of Gender-Aware DNN

The DNN was constructed by stacking a number of restricted Boltzmann machines (RBMs) [24], which were initialized layer-wise by the contrastive divergency algorithm [25]. After that, a softmax layer was placed on top of the network to ensure that the network can produce gender posteriors. Then, backpropagation was applied to minimize the cross-entropy between desired and actual outputs. In this work, we used the utterances in SRE05–SRE12 and their gender labels to train the gender-aware DNN.

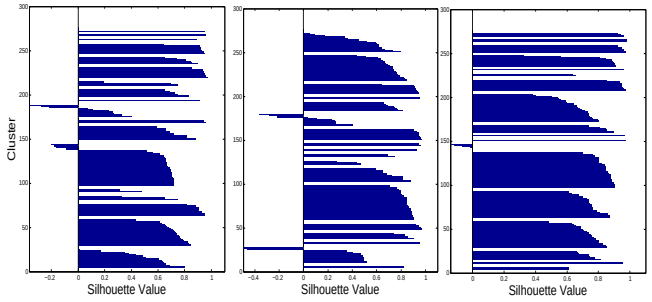
4.3. Score Normalization

As suggested by [26], adaptive score normalization can improve the performance of i-vector/PLDA systems on NIST 2016 SRE signifi-

cantly. To reduce scoring time, we applied adaptive z-norm instead of the more computationally demanding adaptive s-norm as a compromise. Specifically, we used the unlabelled utterances in SRE16-dev as the candidate cohorts for the enrollment utterances. For each enrollment utterance, its PLDA scores with respect to the unlabelled i-vectors in SRE16-dev were computed and ranked; then, the top-200 i-vectors were selected as the cohort set for computing the z-norm parameters of the utterance.

5. RESULTS AND DISCUSSIONS

To compare the quality of the i-vector clusters produced by agglomerative hierarchical clustering (AHC) and iterative spectral clustering (Iterative-SC), we computed the silhouette values from the clusters produced by these two methods and displayed them as silhouette plots in Fig. 2. As AHC can use Euclidean or cosine distance as its distance metric, we refer to the resulting methods as Euclidean-AHC and Cosine-AHC, respectively. Fig. 2 shows that Iterative-SC has the highest average silhouette score and has less negative silhouette values. This suggests that Iterative-SC produces clusters with better quality.



(a) avgscore = 0.650 (b) avgscore = 0.668 (c) avgscore = 0.752

Fig. 2. Silhouette plots showing the quality of i-vector clusters produced by (a) Euclidean-AHC, (b) Cosine-AHC and (c) Iterative SC. Each silhouette pattern represents a cluster, and the silhouette values of individual samples are shown on the horizontal axis.

We used equal error rate (EER) and minimum decision cost function (minDCF) defined in NIST 2016 SRE to evaluate the performance of different systems. Unless stated otherwise, the number of clusters (hypothesized speakers) is 180.

	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

Table 1. Performance of the iterative retraining method in Fig. 1 for different numbers of iterations on SRE16-dev and SRE16-dev.

Table 1 shows that the performance generally improves after a few iterations on both datasets. Because of the mismatch between pre-SRE16 and SRE16 data, the performance in the first iteration is the worst. However, when the number of iterations increases, the PLDA mixture model gradually adapts to the new domain and both the EER and minDCF drop. We observed that increasing the number of iterations beyond 7 does not bring any further peformance improvement.

In the next experiment, we compared different speaker clustering methods and used AHC as the baseline. Also, we used covariance matrix interpolation [14–16] as the baseline for domain adaptation. Specifically, we interpolated the covariance matrices of the in-domain PLDA mixture model with the covariance matrices of the out-of-domain PLDA mixture model using an interpolation weight of 0.5. Table 2 shows the speaker verification performance using the 3 speaker clustering methods. Note that iterative retraining (Fig. 1) is meaningful to Iterative-SC only because the distance metrics of AHC is independent of the PLDA model. Results show that iterative-SC together with the retraining strategy can leverage the limited amount of unlabelled in-domain data to achieves superior performance. Rows 2 and 3 in Table 2 suggest that without iterative re-training, covariance interpolation helps to lower the EER and minDCF. However, when iterative re-training is applied (Row 4 and Row 5), the benefit of covariance interpolation diminishes.

Row	Clustering Method	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

Table 2. Performance of PLDA mixture models on SRE16 using different speaker clustering methods and with and without covariance matrix interpolation (Cov. Interp.).

In the covariance interpolation method [14–16], the out-of-domain data have a *direct* influence on the adapted model and the degree of influence is controlled by an interpolation weight. The problem is that this weight should be set according some prior knowledge about the two domains, which may not be easily quantified. In our method, however, such influence will be progressively diminished during the iterative training process. As shown in Table 1, the PLDA model can be fully adapted to the new domain after 5 iterations.

6. CONCLUSIONS

This paper demonstrates the capability of an iterative training procedure that leverages spectral clustering, inter-dataset variability compensation (IDVC), mixture of PLDA and DNNs for gender-aware speaker verification. Evaluations on NIST 2016 SRE reveal that spectral clustering outperforms traditional clustering methods such as agglomerative hierarchical clustering because the PLDA scoring intrinsically requires i-vector pairs, which can be easily incorporated into the similarity matrix of spectral clustering. Results also show that despite the limited amount of development data and the unavailability of speaker and gender labels in the development data, the proposed method can achieve superior performance. A number of factors contribute to this superior performance. Firstly, the IDVC helps to reduce the gender and language mismatch in the development data of NIST 2016 SRE. Secondly, spectral clustering can effectively find the hypothesized speaker labels for training the PLDA mixture model. Thirdly, the gender-aware DNN provides the gender posteriors for the PLDA mixture model to capture whatever gender information remains in the IDVC-compensated i-vectors.

7. REFERENCES

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. of Odyssey: Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.

[3] NIST, "The NIST year 2016 speaker recognition evaluation plan," in *https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016*, 2016.

[4] K. J. Han and S. S. Narayanan, "Agglomerative hierarchical speaker clustering using incremental Gaussian mixture cluster modeling," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[5] J. Shi and M. J., "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[6] W. Chen, Y. Song, H. Bai, C. Lin, and E. Chang, "Parallel spectral clustering in distributed systems," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 568–586, 2011.

[7] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the nystrom method," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 214–225, 2004.

[8] R. R. Patel, K. Forrest, and D. Hedges, "Relationship between acoustic voice onset and offset and selected instances of oscillatory onset and offset in young healthy men and women," *Journal of Voice*, vol. 31, no. 3, pp. 389–e9, 2017.

[9] 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.

[10] ——, "Compensating inter-dataset variability in PLDA hyperparameters for robust speaker recognition," in *Speaker Odyssey: Speaker and Language Recognition Workshop*, 2014, pp. 282–286.

[11] ——, "Inter dataset variability compensation for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4002–4006.

[12] 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.

[13] 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. 132–142, 2016.

[14] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, "Improving speaker recognition performance in the domain adaptation challenge using deep neural networks," in *IEEE Workshop on Spoken Language Technology (SLT)*, 2014, pp. 378–383.

[15] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 4047–4051.

[16] P. A. Torres-Carrasquillo, F. Richardson, S. Nercessian, D. Sturim, W. Campbell, Y. Gwon, S. Vattam, N. Dehak, H. Mallidi, P. S. Nidadavolu, R. Li, and R. Dehak, "The MIT-LL, JHU and LRDE NIST 2016 speaker recognition evaluation system," *Proc. Interspeech 2017*, pp. 1333–1337, 2017.

[17] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[18] D. E. Colibro, C. Vair, and K. R. Farrell, "Method and apparatus for automatic speaker-based speech clustering," Jun 2016, US Patent 9,368,109.

[19] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, 2011, pp. 249–252.

[20] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition." in *Interspeech*, 2006.

[21] M. McLaren, M. Mandasari, and D. Leeuwen, "Source normalization for language-independent speaker recognition using i-vectors," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, 2012, pp. 55–61.

[22] M. W. Mak and H. B. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations," *Computer, Speech and Language*, vol. 28, no. 1, pp. 295–313, Jan 2014.

[23] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Crete, Greece, Jun. 2001, pp. 213–218.

[24] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[25] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[26] P. Matejka, O. Novotnỳ, O. Plchot, L. Burget, and J. Cernockỳ, "Analysis of score normalization in multilingual speaker recognition," in *Proceedings of Interspeech*, 2017.