A CHANNEL-BLIND SYSTEM FOR SPEAKER VERIFICATION

Najim Dehak[1], *Zahi N. Karam*[2,3], *Douglas A. Reynolds*[3],*Réda Dehak*[4], *William M. Campbell*[3], *James R. Glass*[1]

[1] CSAIL at MIT, Cambridge, MA, USA; [2] DSPG, RLE at MIT, Cambridge, MA, USA;
[3] MIT Lincoln Laboratory, Lexington, MA, USA; [4] LRDE, Paris, France

ABSTRACT

The majority of speaker verification systems proposed in the NIST speaker recognition evaluation are conditioned on the type of data to be processed: telephone or microphone. In this paper, we propose a new speaker verification system that can be applied to both types of data. This system, named blind system, is based on an extension of the total variability framework. Recognition results with the proposed channel-independent system are comparable to state of the art systems that require conditioning on the channel type. Another advantage of our proposed system is that it allows for combining data from multiple channels in the same visualization in order to explore the effects of different microphones and collection environments.

Index Terms— Total variability space, PLDA, LDA, WCCN.

1. INTRODUCTION

Over the last five years, several channel compensation approaches were proposed for speaker verification. Hoverer, Joint Factor Analysis (JFA) [1] became one of the more popular approaches. This technique was proposed in the context of the Gaussian Mixture Model (GMM) framework in order to model between speaker variability and to compensate for channel effects. The basic assumption of the JFA approach is that a high dimensional GMM supervector for a given utterance can be decomposed into the addition of two parts: The first part depends on the speaker, which contains the useful information, and the second depends on the channel, which models the information that we need to compensate for.

Recently, in [2], we proposed a new speaker verification system that uses factor analysis techniques for feature extraction rather than separate speaker and channel modeling, as is done in JFA [1]. In this new approach, every speech recording is mapped into a single low-dimensional total variability vector named *total factors*. Unlike JFA, there is no distinction between the speaker and intersession variabilities in the GMM supervector space. The channel compensation in the new approach is carried out in the low-dimensional total variability space instead of the GMM supervector space. It is comprised of a combination of Linear Discriminant Analysis (LDA) and Within Class Covariance Normalization (WCCN) [2]. The speaker verification decision score is obtained using the cosine similarity computed between the target and test total factors. The total variability space was first applied in the context of telephone data of the NIST speaker recognition evaluation. However, an extension of this approach was also proposed in the context of the microphone data as well [3]. This approach consists of stacking extra total factors estimated on the microphone data to the original telephone total factors. An extension of the LDA and WCCN combination was also proposed to handle the interview condition.

In the context of NIST Speaker Recognition Evaluations (SRE) [4], all proposed speaker verification systems are conditioned on the type of data (telephone or interview) to be used. In this paper we propose a new single total variability system, that can be applied simultaneously for both telephone and microphone data without prior knowledge about the data type being processed. This new system is also based on the total variability space stacking for both telephone and interview data as proposed in [3]. However, we will show how Probabilistic Linear Discriminant Analysis (PLDA) [5] can be used to project both telephone and interview total factors into a common space. In this new space, an LDA and WCCN combination was also applied to further compensate for remaining channel effects.

A data visualization technique, first proposed in the context of speaker verification in [6], is used as both an exploratory and analysis tool. The technique uses graph embedding, graph layout and visualization software [7] to visualize all speech utterances within a data-set of interest. This is done in a manner that groups similar, as set by the score of the blind system, utterances together. With this tool we are able to highlight the efficacy of the blind system, as well as the crucial role of WCCN/LDA in removing channel variability.

2. TOTAL VARIABILITY SPACE

The total variability space proposed in [2] models both the speaker and channel variabilities simultaneously. It is defined by the total variability matrix, which contains the eigenvectors with the largest eigenvalues of the total variability covariance matrix. In this new model, we make no distinction between the speaker and the channel effects in the GMM supervector space, as compared to JFA [1] which does. For a given speech utterance, the speaker- and channel-dependent GMM supervector is represented by the following equation

$$M = m + T_{tel}w \tag{1}$$

where m is the Universal Background Model (UBM) supervector, the low rank matrix T_{tel} defines the total variability space estimated on telephone speech, and the vector w is the speaker- and session-dependent factors in the total variability space. The w vectors are random variables distributed according to the Normal distribution $\mathcal{N}(0, I)$.

The large success of this new approach on the telephone data of the NIST-SRE is mainly due to the large amount of telephone data used to train the total variability matrix T_{tel}. An extension of the total variability space to the interview data of the NIST evaluation is proposed in [3]. It is based on estimating extra total variability

This work was sponsored by the Department of Defense under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

components on the interview data and stacking them with the original telephone components. The new space is composed of the concatenation of both telephone and interview matrices. Note that the total variability components of the interview data are complementary to the telephone data (the two spaces are not independent). The stacking approach was proposed in order to solve the problem of imbalance in the quantity of telephone and microphone data. The new speaker- and channel-dependent supervector for a given utterance can be obtained as follows

$$M = m + N\hat{w} \tag{2}$$

where $N = [T_{tel}, T_{int}]$ is the new total variability matrix composed of the concatenation of both the telephone and microphone data total variability matrices. The vector $\hat{w}$ is the speaker- and session-dependent factor in the new telephone and microphone total variability space. An important characteristic of this new modeling is that the microphone data lives in the full total variability space defined by both telephone and interview matrices $[T_{tel}, T_{int}]$. However, the telephone data lives only in the original telephone space obtained by the telephone total variability matrix T_{tel}. To project both telephone and microphone data in the same space, we used Probabilistic Linear Discriminant Analysis (PLDA) [5], which is described in the next section.

3. PROBABILISTIC LDA

PLDA is similar to the JFA approach but applied in the low-dimensional total variability space rather than the GMM supervector space [1]. It was introduced and used in face recognition [5]. The new total factor $\hat{w}$ for a given utterance can be generated using the following process:

$$\hat{w} = \mu + Vy + Ux + \epsilon, \tag{3}$$

where μ is the mean over all training examples. The matrix V defines the speaker subspace (eigenvoices matrix), and U defines a session subspace (eigenchannels matrix). The vectors y and x are the speaker and session dependent factors in the respective subspace and each is assumed to be a random variable with a Normal distribution $\mathcal{N}(0, I)$. The term ϵ models the residual noise not captured with the matrix U, and is modeled by a full covariance matrix Σ. To apply PLDA for speaker verification, we first need to estimate the PLDA hyper-parameters $\theta = (\mu, V, U, \Sigma)$ based on a maximum likelihood approach given appropriate labeled development corpora [5].

3.1. Hyper-Parameter Training

We use a maximum likelihood approach, divided into two steps, to train the PLDA parameters that is quite different from the approach in [5]. The first step consists of estimating μ, V and Σ under the assumption that $U = 0$. In the second step, we estimate separately the channel matrix U by fixing the parameters μ, V and Σ. This training regime is similar to JFA training as proposed in [1]. The reason for this training split is that the first set of parameters (μ, V, Σ) are trained only on telephone data, while the matrix U is trained on microphone data.

3.1.1. Training (μ, V, Σ)

To estimate these parameters, we used the same EM algorithm as proposed in [5]. We started our training with a random initialization of the three parameters (μ, V, Σ). In the E-step of the EM algorithm, we need to estimate the posterior distribution of the hidden variable y for a given speaker i. This distribution has a Gaussian form with mean vector and covariance matrix defined as follows

$$E[y_i] = \left(J\left(V^t\Sigma^{-1}V\right) + I\right)^{-1} \sum_{j=1}^{J} V^t\Sigma^{-1}\left(\hat{w}_{ij} - \mu\right) \tag{4}$$

$$E\left[y_i y_i^t\right] = \left(J\left(V^t\Sigma^{-1}V\right) + I\right)^{-1} + E[y_i]\,E[y_i]^t\,, \tag{5}$$

where J corresponds to the number of recordings for a given speaker i. The M-step consists of updating the values of the three PLDA parameters based on the means and covariance matrices for all the speakers' hidden variables as evaluated in the E-step. The new parameter values that maximize the likelihood, are given as follows

$$\mu = \frac{1}{IJ}\sum_{i,j} x_{i,j} \tag{6}$$

$$V = \left(\sum_{i,j}\left(\hat{w}_{i,j} - \mu\right)E[y_i]^t\right)\left(\sum_{i,j} E\left[y_i y_i^t\right]\right)^{-1} \tag{7}$$

$$\Sigma = \frac{1}{IJ}\sum_{i,j}\left(\hat{w}_{i,j} - \mu\right)\left(\hat{w}_{i,j} - \mu\right)^t - VE[y_i]\left(\hat{w}_{i,j} - \mu\right)^t , \tag{8}$$

where I is the number of speakers in the training corpora and J corresponds to the number of recordings for each speaker i.

3.1.2. Training U

Training the channel matrix U on the microphone data is quite similar to the matrix V training. Before estimating the channel matrix, we need to remove the telephone variability already captured by the first set of parameters (μ, V, Σ) from the total factors of the microphone data. For a given speaker's microphone recordings, we used the first set of parameters already trained in telephone data to estimate a single speaker factor vector for all these recordings using a MAP point estimate. This speaker factor vector is then used to centralize the entire microphone total factors of the same speaker. The matrix U re-estimation equation corresponds exactly to the equation (7) except that we modify the term μ corresponding to the mean of the entire speakers population by y_i which is the speaker factor vector of a given speaker i.

3.2. Score Evaluation

Similar to [2], the decision score is evaluated using a modified version of the cosine similarity between the target and test speaker-dependent factors based on the LDA and WCCN projection matrices. It is given by the following equation.

$$score\left(y_{target}, y_{test}\right) = \frac{\left(A^t y_{target}\right)^t W^{-1}\left(A^t y_{test}\right)}{\sqrt{\left(A^t y_{target}\right)^t W^{-1}\left(A^t y_{target}\right)}.\sqrt{\left(A^t y_{test}\right)^t W^{-1}\left(A^t y_{test}\right)}} \tag{9}$$

where A is the LDA projection matrix and W is the within class covariance matrix. Unlike our previous system configuration [2], both matrices are trained on both telephone and microphone data.

4. DATA VISUALIZATION

The visualization begins by embedding the speech utterances in a given data-set in a nearest neighbor (NN) graph. The embedding

Table 1. Corpora used to estimate the UBM, total variability matrix (T), PLDA, LDA and WCCN.

	UBM	T	PLDA		LDA	WCCN
			μ, V, Σ	U		
Switchboard II	X	X	X		X	
Switchboad Cellular	X	X	X		X	
Fisher English		X				
NIST 04, 05, 06 Tel	X	X	X		X	X
NIST 05, 06 Mic		X		X	X	X
NIST 08 dev, test		X		X	X	X

creates a graph of vertices and edges, where the vertices represent the utterances and undirected edges represent connections, based on similarity, between a pair of points. The embedding places an edge between two vertices if they are among the K-NNs of each other, where the distance is the cosine distance obtained by comparing all pairs of utterances using the blind system. It is important to note that the full comparison of all pairs using the blind system is a computationally cheap step since each comparison corresponds to an inner product in a 600 dimensional space. In the resulting NN-graph the location of the vertices is not important, only the existence of the edges between them does. The graph is then "laid out"; the process of choosing vertex locations, in a manner that would result in good visualization, specifically grouping highly connected utterances with each other. We use the GUESS [7] software package to perform both the visualization and the layout using the GEM algorithm [8]. Meta data can then be overlaid on the graph by varying the colors and shapes of the vertices, e.g. coloring all utterances from the same speaker in the same manner. This overlaid information allows for understanding structures that emerge in the visualization.

5. EXPERIMENTS

Our experiments operate on cepstral features, extracted using a $25ms$ Hamming window. 19 mel frequency cepstral coefficients together with log energy are calculated every 10 ms. Delta and double delta coefficients were then calculated using a 5 frame window to produce 60-dimensional feature vectors. This 60-dimensional feature vector was subjected to feature warping using a $3s$ sliding window. We used gender dependent UBMs containing 2048 Gaussians. Table 1 summarizes all corpora used to estimate the UBM, total variability matrix, PLDA, LDA and WCCN.

To compare our new approach (blind system) with condition-dependent systems in the telephone data, we built two different systems. The first system is a total variability system trained only on telephone data similar to [2]. We estimated a total variability matrix of dimension 600. LDA was used to reduce the dimension to 250, and then WCCN is applied to normalize the cosine kernel. The scores were zt-normalized based on a set of impostors taken from the telephone data. The blind system is based on 600 total factors trained on telephone speech and 200 total factors trained on microphone data. PLDA is used to project both telephone and microphone total factors into the same space of dimension 600. Another dimension reduction based on classical LDA is applied to reduce the space to 250 dimensions. The WCCN technique is then used to normalize the cosine kernel in the reduced space (250).

Table 2. Comparison of results between conditioned and blind system on the core telephone condition of the NIST 2010 SRE (det5). The results are given on Equal Error Rate (EER) and minimum Detection Cost Function (DCF)

Systems	Female		Male	
	EER	DCF	EER	DCF
Conditioned	**2.56%**	0.608	**1.73%**	0.384
Blind	2.96%	**0.481**	1.92%	**0.352**

All our experiments were carried out on the extended trials of the core condition of the NIST 2010 SRE. It is composed by telephone conversation data of 5 minutes and interview data of 3 minutes. A comparison of the results between the conditioning and blind system in the tel-tel condition (det5) is reported in Table 2. These results are given on the EER and the new MinDCF point [4].

The results reported in Table 2 show that both systems achieved equivalent results for male trials. However, the blind system obtained better MinDCF compared to the conditioned system. Adding microphone data to train the blind system did not hurt the performance of the system in telephone data. The DET curve given in Figure 1 presents the performance of the blind system on the interview data of the core condition of the NIST 2010.

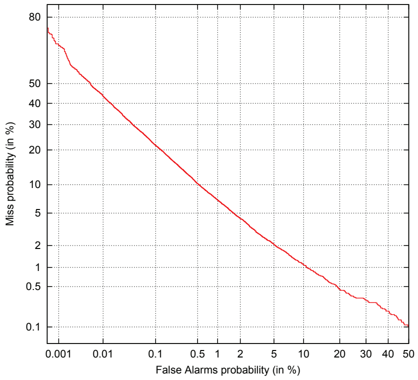

Fig. 1. The performance of the blind system on the interview data.

5.1. Data Visualization

In this section, A data analysis based on visualization is proposed in order to study the channel effects in the context of the blind system. We will present only male utterances as they paint the clearest picture, however similar results are observed with female utterances. The graphs show all male utterances of the core conditions of the 2010 extended NIST SRE, and the number of NNs is set to $K = 3$.

We will begin by showing the efficacy of the blind system by using the system in building the NN-graph. Figure 2 shows the resultant visualization with speaker meta-data overlaid such that utterances of the same speaker are colored alike. The clusters of similar color, representing clusters of utterances of the same speaker, show that the system is indeed assigning lower distance values to pairs of utterances of the same speaker.

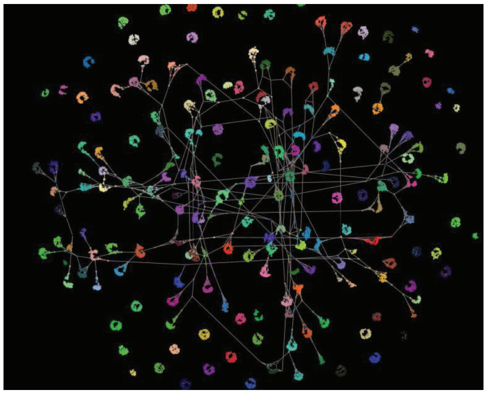

Fig. 2. Graph visualization of all Male utterances of the NIST SRE 2010. It is based on using the full blind system with speaker meta data overlaid.

Next we examine the importance of the channel compensation performed by the combination of WCCN/LDA. To do this we build a NN-graph using the blind system without the WCCN/LDA step, the corresponding visualization is in Figure 3. We notice that the speaker clustering observed with the full blind system is no longer visible, however there does seem to be some structure to the graph.

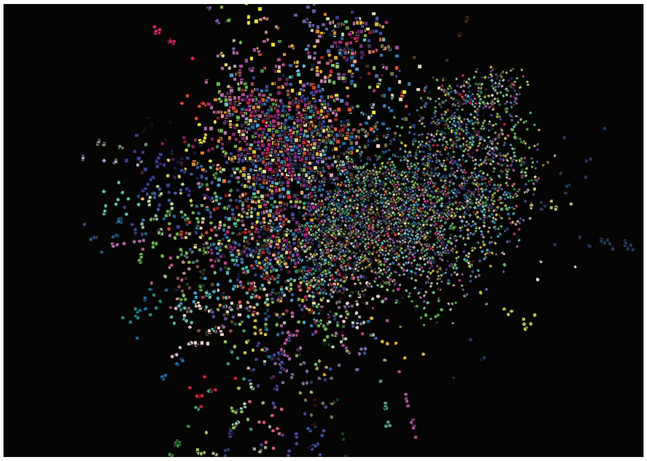

Fig. 3. Graph visualization of all Male utterances of the NIST SRE 2010. It is based on using the blind system without LDA/WCCN channel compensation with speaker meta data overlaid.

Further exploration, by overlaying channel meta-data, shows that the structure can be attributed to channel variability. Figure 4 shows the layout of the NN-graph using the blind system without WCCN/LDA with: colors representing different telephone and microphone channels, the node shape representing the two different rooms the interview data was collected in. Upon careful inspection of the graph, one notices that the room accounted for more variability than the interview microphones, specifically for the far talking microphones: MIC CH 05/07/08/12/13. Another observation is that the two phone numbers corresponding to the land-line phones located in each of the rooms (215573qqn and 215573now) cluster near the interview data, and specifically near the close talking and desk microphones: MIC CH 02/04.

This ability to visualize and explore the dominant variability within a data-set may prove to be a useful tool when dealing with newly collected data-sets, and the relatively low computation cost of the cosine distance scoring allows for handling large corpora.

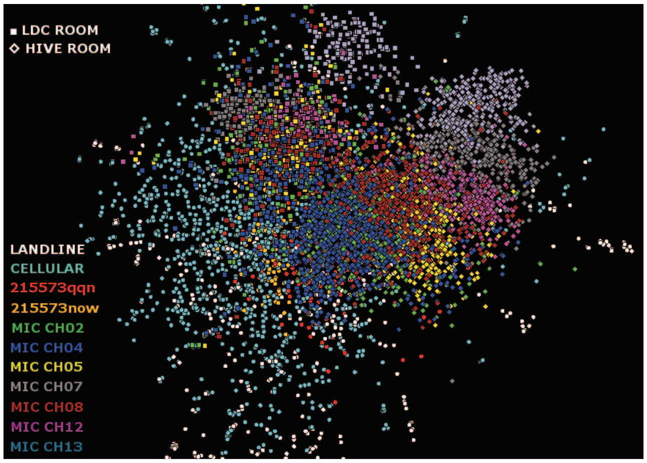

Fig. 4. Graph visualization of all Male utterances of the NIST SRE 2010. It is based on using the blind system without LDA/WCCN channel compensation with channel meta data overlaid.

6. CONCLUSION

This paper presents a new single speaker verification system that can be applied simultaneously, without conditioning, to both telephone and interview data of the NIST SRE, and achieves state of the art performance in both. This system, which is based on the cosine similarity, allowed us to use data visualization to show the channel effects in the data and how the LDA and WCCN combination can compensate for them.

7. REFERENCES

[1] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A Study of Interspeaker Variability in Speaker Verification,” *IEEE Transaction on Audio, Speech and Language*, vol. 16, no. 5, pp. 980–988, july 2008.

[2] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, “Front End Factor Analysis for Speaker Verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. to appear, 2010.

[3] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, “An i-Vector Extractor Suitable for Speaker Recognition with Both Microphone and Telephone Speech,” in *Odyssey*, Brno, Czech Republic, 2010.

[4] `http://www.itl.nist.gov/iad/mig/tests/sre/2010/index.html`.

[5] S.J.D. Prince and J.H. Elder, “Probabilistic Linear Discriminant Analysis for Inferences about Identity,” in *11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[6] Z.N. Karam and W.M. Campbell, “Graph-Embedding for Speaker Recognition,” in *Interspeech*, 2010.

[7] Eytan Adar, “Guess: A Language and Interface for Graph Exploration,” in *CHI*, 2006.

[8] Di Battista, P. Eades, R. Tamassia, and I.G. Tollis, *Graph Drawing: Algorithms for Visualization of Graphs*, Prentice Hall, 2002.