Exploiting Intra-Conversation Variability for Speaker Diarization

Stephen Shum[1]*, Najim Dehak*[1]*, Ekapol Chuangsuwanich*[1]*, Douglas Reynolds*[2]*, Jim Glass*[1]

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA
[2]MIT Lincoln Laboratory, Lexington, MA
{sshum,najim,ekapolc,glass}@csail.mit.edu, dar@ll.mit.edu

Abstract

In this paper, we propose a new approach to speaker diarization based on the Total Variability approach to speaker verification. Drawing on previous work done in applying factor analysis priors to the diarization problem, we arrive at a simplified approach that exploits intra-conversation variability in the Total Variability space through the use of Principal Component Analysis (PCA). Using our proposed methods, we demonstrate the ability to achieve state-of-the-art performance (0.9% DER) in the diarization of summed-channel telephone data from the NIST 2008 SRE.

Index Terms: speaker diarization, factor analysis, Total Variability, principal component analysis

1. Introduction

Audio diarization is defined as the task of marking and categorizing the different audio sources within an unmarked audio sequence. The types and details of the audio sources are application specific, but can include particular speakers, music, background noise sources, et cetera. This paper concerns speaker diarization, or "who spoke when", the problem of annotating an unlabeled audio file where speaker changes occur (*segmentation*) and then associating the different segments of speech belonging to the same speaker (*clustering*). [1]

We develop an approach to diarization based on the successes of factor analysis-based methods in speaker recognition [2], as well as diarization [3], [4]. Inspired by the ability of the Total Variability subspace to extract speaker-specific features on short segments of speech [2], [5], we propose a method for performing speaker clustering directly in the low-dimensional Total Variability subspace. By evaluating the performance of our system on the same summed-channel telephone data from the 2008 NIST Speaker Recognition Evaluation (SRE), we show that our resulting work is not only simpler than the Variational Bayes system formulated previously in [3], but can also achieve the same state-of-the-art performance.

The rest of this paper is organized as follows: Section 2 reviews the Total Variability approach as a factor analysis-based front-end for extracting speaker-specific features. Section 3 then motivates the use of PCA to exploit intra-conversation variabilities for speaker clustering before Section 4 outlines the remaining details of our system. The results of our experiments are explained in Section 5, and Section 6 concludes with a discussion of possible directions for future work.

2. A Review of Total Variability

At the heart of speaker diarization lies the problem of speaker modeling. In an effort to enhance the classical method of modeling speakers using Gaussian Mixture Models (GMMs) [6], recently developed methods apply factor analysis to supervectors - a vector consisting of stacked mean vectors from a GMM - in order to better represent speaker variabilities and compensate for channel (or session) inconsistencies [2]. One such approach is Total Variability, which decomposes a speaker- and session-dependent supervector M as follows:

$$M = m + Tw + \epsilon \tag{1}$$

where m is the speaker- and session-independent supervector commonly taken from a large GMM, known as the Universal Background Model (UBM), trained to represent the speaker-independent distribution of acoustic features [6]. T is a rectangular matrix of low rank that defines the Total Variability subspace, w is a low-dimensional random vector with a standard normal prior distribution $\mathcal{N}(0, I)$, and the residual noise term $\epsilon \sim \mathcal{N}(0, \Sigma)$ covers the variabilities not captured by T [7]. The vector w will be referred to as a *total factor vector* or an *i-vector*.

The cosine similarity metric has been applied successfully in the Total Variability subspace to compare two i-vectors [2]. Given any two total factor vectors w_1 and w_2, the cosine similarity score is given as

$$\text{score}\,(w_1, w_2) = \frac{(w_1)^t\,(w_2)}{\|w_1\| \cdot \|w_2\|} \tag{2}$$

By working within the Total Variability subspace instead of projecting back into the GMM-supervector space, this scoring function is considerably less complex than the log-likelihood ratio scoring operations used in the past [6].

3. Intra-Conversation Variability

The Total Variability approach has achieved state of the art results in the task of speaker verification [2]; it is therefore natural to try to adapt these methods for the problem of speaker diarization. We began by recognizing the shortcomings of standard (speaker verification-based) inter-session compensation techniques when applied to speaker diarization: the use of eigenchannels was ineffective in [3], as was the rote application of LDA+WCCN for the Total Variability-based i-vectors [2]. This gave way to the realization that compensating for inter-session variability was wholly unnecessary in the problem of diarization; because we were working on summed-channel telephone conversations, there was really no *inter*-session. What we really cared about were *intra*-session (or intra-conversation) variabilities within each audio file. Such insight paved the way for the rest of this work.

Assuming we have some initial segmentation in place, we can extract an i-vector for each segment. Then to associate

each i-vector with a corresponding speaker, we generate clusters from them. For our experiments, we assume that there are exactly two speakers in the given conversation. Of course, it is not known *a priori* where our two respective speakers lie in the Total Variability space, but because i-vectors were designed to contain primarily speaker-specific information, the most prominent source of variability between these i-vectors ought to be attributed to differences between the speakers' voices.

We can find the directions of maximum variability within our Total Variability space by using simple Principal Component Analysis (PCA). Figure 1 shows the first two principal components of a set of Total Factors extracted from a male/female conversation The plot also includes, in black x's, the i-vectors corresponding to overlapped speech segments. To be sure, the PCA projection was calculated on all i-vectors including these overlapped speech segments, as we have not yet explored ways to distinguish between overlapped and non-overlapped speech.

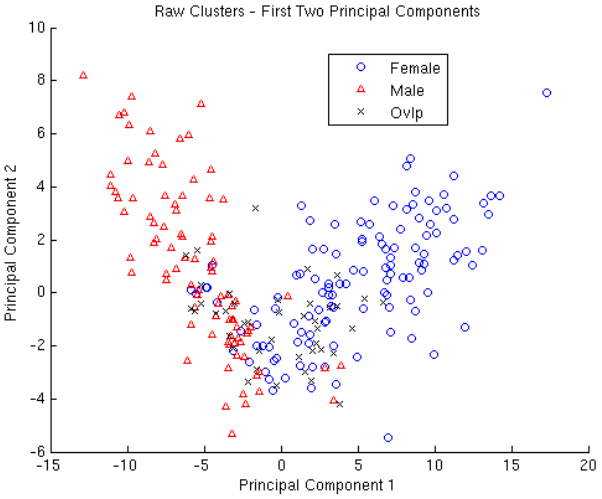

Figure 1: *Plot of the first two dimensions (principal components) of PCA-projected speaker i-vectors. The triangles in red represent i-vectors of a male speaker, while the blue circles represent i-vectors of a female speaker in the same conversation. The black x's correspond to i-vectors representing overlapped speech.*

Though this is a visualization of only the first two principal components from an initial i-vector dimension of 400, we can already see a distinct separation between the sets of total factor vectors corresponding to different speakers. Furthermore, it can be observed that the separation between the two clusters is primarily directional; this is because a PCA projection centers the mean of the dataset at the origin and also because each i-vector has a standard normal prior distribution. This suggests that the most important information may be contained not in the magnitude of the i-vector, but in its relative orientation. Figure 2 shows a length-normalized version of the first two principal components for the same two speakers seen in Figure 1. Notice how the majority of each cluster can be found in distinctly different regions along the unit circle. This further motivates the use of the cosine similarity as a metric for comparing i-vectors.

To even further emphasize the importance of the PCA directions with the most variability (i.e. largest eigenvalues), we introduce the following weighted modification to our cosine similarity score

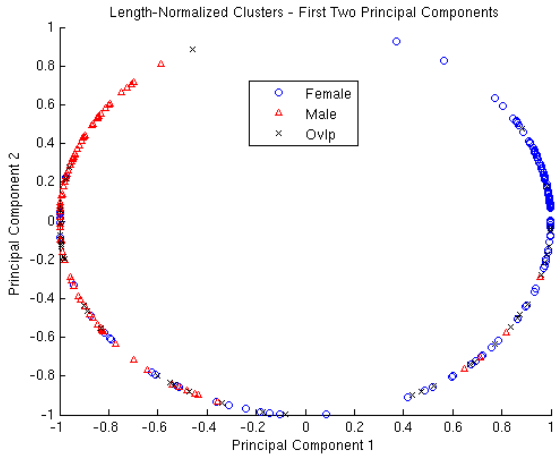

Figure 2: *Plot of the length-normalized speaker i-vectors after applying a two dimensional PCA-projection across the entire conversation. Notice also the random scatter of the black x's corresponding to overlapped speech segments.*

$$\text{score}\left(w_1', w_2'\right) = \frac{\left(w_1'\right)^t \Lambda \left(w_2'\right)}{\|\Lambda^{\frac{1}{2}} w_1'\| \cdot \|\Lambda^{\frac{1}{2}} w_2'\|} \tag{3}$$

where w_i' is the PCA-projected i-vector and Λ is the corresponding diagonal matrix of the eigenvalues. Additionally scaling our PCA-projected i-vector components by the square root of the eigenvalues $\Lambda^{\frac{1}{2}}$ gives us added emphasis on the directions of higher variability (i.e. the "most" principal components). Though PCA naturally gives more scoring weight to the larger principal components, our experiments showed that increasing this effect artificially had a positive impact on performance.

4. The Speaker Diarization System

This section describes the various parts of our proposed diarization system.

4.1. Segmentation

To obtain an initial segmentation on the summed-channel telephone data, we use a Harmonicity and Modulation Frequency-based Voice Activity Detector (VAD) described in [8]. Its output gives us the start/stop times for segments that are classified as speech. Over the entire test set, the average length of these segments is 1.09s with a standard deviation of 0.648s. Though the segment lengths range widely between 0.03s and 11.31s, we chose to use this VAD without any additional refinements.

4.2. PCA-based Dimensionality Reduction

After extracting an i-vector for each speech segment in our conversation, we apply PCA-based projection as described in Section 3. Rather than forcing the system to adhere to a specific number of principal components (dimensions), however, we specified a proportion p of eigenvalue mass to use instead. That is, we use the dimensions corresponding to the n largest eigenvalues such that

$$\min_n \frac{\sum_{i=1}^{n} \lambda_i}{\sum_{j=1}^{D} \lambda_j} \geq p \tag{4}$$

where we assume that our set of eigenvalues $\{\lambda_i\}$ is indexed in decreasing order and D is the initial i-vector dimension. For some additional insight, Table 1 provides some statistics regarding the number of dimensions used for different values of p given an initial i-vector dimension of $D = 400$. Ultimately, the proportion that provided the best empirical results was $p = 0.5$ (i.e. 50% eigenvalue mass), which we will use for our subsequent experiments.

Pct. Eig. Mass (p)	Avg Dim (n)	Min n	Max n
30%	10.3	5	10
50%	25.5	16	33
80%	70.1	52	84

Table 1: *Comparison of the number of PCA-dimensions needed for different proportions of eigenvalue mass. These statistics were computed over 200 randomly selected test files from the NIST 2008 SRE.*

4.3. First Pass Clustering

To perform the clustering step with our new set of PCA-projected and dimensionality-reduced i-vectors, we simply use K-means ($K = 2$) clustering based on the cosine distance. The iterative nature of this algorithm allows it to self-correct poor initializations, whereas other methods such as the bottom-up approach of agglomerative hierarchical clustering used in [3] uses only one iteration to make hard decisions.

4.4. Re-segmentation

After an initial clustering, we refine our initial segmentation boundaries using a Viterbi re-segmentation and Baum-Welch soft speaker clustering algorithm detailed in [3]. At the acoustic feature level, this stage initializes a 32-mixture GMM for each of the clusters (Speaker A, Speaker B, and non-speech N) defined by the First Pass Clustering. Posterior probabilities for each cluster are then calculated given each feature vector x_t (i.e. $P(A|x_t), P(B|x_t), P(N|x_t)$) and pooled across the entire conversation, providing a set of Baum-Welch statistics from which we can re-estimate each respective speaker's GMM. In order to prevent this unsupervised procedure from going out of control, the non-speech GMM is never retrained. In the Viterbi stage, each frame is assigned to the speaker/non-speech model with the highest posterior probability. This algorithm runs until convergence but is capped at 20 Viterbi iterations, each of which involves 5 iterations of Baum-Welch re-estimation [3].

4.5. Second Pass Refinements

We further refine the diarization results of the Re-segmentation stage by extracting a single i-vector for each respective speaker using the (newly-defined) re-segmentation assignments. Each segment i-vector (also newly extracted) is then reassigned to the speaker whose i-vector is closer in cosine similarity. We iterate this procedure until convergence - when the segment assignments no longer change. This can be seen as another pass of K-means clustering, where the "means" are computed according to the process of i-vector estimation detailed in [2].

5. Experiments

We used a gender-independent UBM of 1024 Gaussians built solely on 20-dimensional MFCC feature vectors without derivatives to train a gender-independent Total Variability matrix of rank 400. This configuration was chosen to be somewhat consistent with that of the Variational Bayesian (VB) system described in [3], though we will also report later on the results of using Total Variability matrices of different rank.

5.1. Evaluation Protocol

Set up by NIST, the Diarization Error Rate (DER) is the primary performance measure for the evaluation of diarization systems and is given as the time-weighted sum of the following three error types: *Miss* (M) - classifying speech as non-speech, *False Alarm* (FA) - classifying non-speech as speech, and *Confusion* (C) - confusing one speaker's speech as from another [9]. In evaluating DER's, we first obtain a reference by applying a speech activity detector to each separate channel of the telephone conversation. Then the evaluation code ignores intervals containing overlapped speech as well as errors of less than 250ms in the locations of segment boundaries. Although overlapped speech intervals do not count in evaluating DER's, the diarization systems do have to contend with overlapped speech in performing the speaker segmentation and clustering.

It is clear that the Miss and False Alarm errors are solely caused by a mismatch between the reference speech activity detector and the diarization system's VAD and Re-segmentation output. A more straightforward metric for the effectiveness of our speaker modeling and clustering methods is in the measurement of Confusion error. In order to focus solely on this type of error, the results reported in [3] were based on the use of reference boundaries as the initial speech/non-speech segmentation, thus driving both miss and false alarm error rates to zero. On our end, we will first report on the detailed results achieved using our own VAD to provide an initial segmentation. Then, for proper comparison, we will also report on a final experiment done using the reference boundaries as the initial speech/non-speech segmentation.

5.2. Results

Following the work in [3], we evaluate the performance of our diarization system on the summed-channel telephone data from the NIST 2008 SRE. This consists of 2215 two-speaker telephone conversations, each approximately five minutes in length ($\approx$ 200 total hours). Table 2 shows the results obtained from our system at each stage described in Section 4.

		Error Breakdown			
	DER (%)	M	FA	C	σ (%)
First Pass	13.8	7.7	2.0	4.0	9.6
Re-segmentation	5.6	0.3	2.3	2.9	8.6
Second Pass	4.2	0.3	2.3	1.5	7.0

Table 2: *Results obtained after each stage of the diarization procedure described so far. The configuration for the First Pass Clustering uses 400-dimensional i-vectors as input to a PCA-projection involving 50% of the eigenvalue mass.*

The helpfulness of the Re-segmentation step is readily apparent, both for correcting the mismatch between the initial and reference VAD's as well as for improving on Speaker Confusion error. Because it does not change the speech/nonspeech boundaries, the Second Pass Refinement stage does not affect the Miss/False-Alarm errors, but is rather effective in driving down Speaker Confusion error. We can also see in the breakdown that the reason for a seemingly high DER in the First Pass

Clustering is primarily due to missed speech in the initial segmentation itself.

These results can be further improved by optimizing over different initial ranks of the Total Variability (TV) matrix. Table 3 shows the statistics obtained from the various i-vector dimensions attempted. Note that PCA (50% eigenvalue mass) is still applied to the set of i-vectors corresponding to each individual test file.

	TV40	**TV100**	TV200	TV400	TV600
Avg Dim	7	**14**	20	26	28
DER (%)	3.9	**3.7**	3.8	4.2	4.0
σ (%)	6.6	**6.4**	6.4	7.0	6.9

Table 3: *Overall diarization performance of Total Variability matrices of varying rank. The second row lists the average number of dimenisions that resulted after the PCA projection (50%) was estimated.*

We settled on the TV100 configuration, which gave the best results despite a relatively low dimensionality, for our final experiment. Table 4 compares our final results to those of the systems described in [3]. The BIC-based system served as a baseline for the FA/VB-based work. Both of those systems were initialized using the reference speech detection boundaries; thus, they incurred no Miss (M) or False Alarm (FA) error, and all of their error is attributed to Speaker Confusion (C). For a valid comparison, we report the results of our system (TV100, 50% PCA) using the reference boundaries as an initial segmentation, denoted "Ref VAD." And finally, we also report the results obtained using our "Own VAD" as described in 4.1.

	Speaker Confusion (%)	σ_C (%)
BIC-based Baseline	3.5	8.0
VB-based FA	1.0	3.5
Ref VAD + TV100	**0.9**	**3.2**
Own VAD + TV100	1.1	3.3

Table 4: *Comparison of diarization results on the NIST SRE 2008 Summed-Channel Telephone Data.* (BIC - Bayesian Information Criterion; FA - Factor Analysis; VB - Variational Bayes; VAD - Voice Activity Detector; TV - Total Variability)

We can see that our "Ref VAD" system - which follows the exact same evaluation protocol as the BIC and VB systems - slightly outperforms the VB system, while the performance of our "Own VAD" system degrades slightly as a result of a mismatched initial segmentation. At the end of the day, however, the difference in performance between these three systems (VB, "Ref VAD", "Own VAD") is minimal. Nevertheless, what is clear is that these approaches are both very successful in the two-speaker telephone diarization task at hand.

6. Conclusions

Inspired by the success of factor analysis and Total Variability for the speaker modeling, we have developed a system that achieves state-of-the-art results on the two-speaker telephone diarization task. Our previous benchmark, the VB system described in [3], elegantly integrates the factor analysis paradigm with the prior work on Variational Bayesian methods for speaker diarization described in [10]. In a search for added simplicity, we utilized the effectiveness of the cosine similarity metric in the Total Variability subspace.

There are still many ways in which we can improve and refine this initial approach. For one, there is a need to address the problem of overlapped speech detection. Finding a good way to robustly detect and remove corrupted segments would be helpful for our PCA initialization and subsequent clustering [11]. Additionally, our reported results have been restricted to two-speaker telephone conversations; we have not yet addressed the issue of applying our system to a conversation setting involving an unknown number of speakers. To that end, we see potential in extending our approach to diarization by applying Variational Bayesian methods for model selection (i.e. determining the number of speakers) and clustering in the Total Variability space [12].

7. Acknowledgements

Thanks to Brno University of Technology for providing the reference segmentation used for our experiments.

The MIT Lincoln Laboratory portion of this work is sponsored by the National Security Agency under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

8. References

[1] S. Tranter and D. Reynolds, "An overview of automatic speaker diarisation systems," *IEEE Transactions on Audio, Speech, and Language Processing*, Sept. 2006.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, July 2010.

[3] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, Dec. 2010.

[4] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *Proceedings of ICASSP*, 2008.

[5] S. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," in *Proceedings of IEEE Odyssey*, 2010.

[6] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Dig. Sig. Proc.*, 2000.

[7] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, May 2005.

[8] E. Chuangsuwanich, S. Cyphers, J. Glass, and S. Teller, "Spoken command of large mobile robots in outdoor environments," in *Proceedings of IEEE SLT Workshop*, 2010.

[9] NIST, "Diarization error rate (der) scoring code," 2006, www.nist.gov/speech/tests/rt/2006-spring/code/md-eval-v21.pl.

[10] F. Valente, "Variational bayesian methods for audio indexing," Ph.D. dissertation, Univ. De Nice-Sophia Antipolis, Sept. 2005.

[11] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech," in *Proceedings of Interspeech*, 2008.

[12] M. Beal, "Variational algorithms for approximate bayesian inference," Ph.D. dissertation, University College London, May 2003.