SPEAKER DIARIZATION WITH PLDA I-VECTOR SCORING AND UNSUPERVISED CALIBRATION

Gregory Sell and Daniel Garcia-Romero

Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD, USA

ABSTRACT

Speaker diarization via unsupervised i-vector clustering has gained popularity in recent years. In this approach, i-vectors are extracted from short clips of speech segmented from a larger multi-speaker conversation and organized into speaker clusters, typically according to their cosine score. In this paper, we propose a system that incorporates probabilistic linear discriminant analysis (PLDA) for i-vector scoring, a method already frequently utilized in speaker recognition tasks, and uses unsupervised calibration of the PLDA scores to determine the clustering stopping criterion. We also demonstrate that denser sampling in the i-vector space with overlapping temporal segments provides a gain in the diarization task. We test our system on the CALLHOME conversational telephone speech corpus, which includes multiple languages and a varying number of speakers, and we show that PLDA scoring outperforms the same system with cosine scoring, and that overlapping segments reduce diarization error rate (DER) as well.

1. INTRODUCTION

Most speech processing algorithms are designed under the assumption that only one speaker is present in a given signal. In automatic speech recognition (ASR), for example, speaker adaptation requires a single speaker for an entire utterance. Speaker recognition also typically assumes only one speaker in an audio cut for i-vector extraction. The presence of multiple speakers can potentially disrupt these algorithms.

Outside of controlled evaluations, it is difficult to ensure that a spoken document will have only one speaker. In these situations, speaker diarization is required, which is the process of segmenting speech from the same speaker in a larger conversation. The goal of speaker diarization is not specifically to determine who is talking (speaker recognition), but rather to determine when someone is speaking.

One way to consider speaker diarization is as a series of speaker recognition tasks with short utterances. The approach, in this thinking, is to take two sections of a given conversation and determine if they were spoken by the same individual, a process that can be repeated for all spoken segments of the conversation. In this context, it is no surprise that i-vectors have become a popular tool for speaker diarization, given their success in speaker recognition.

In this paper, we will present an i-vector speaker diarization system that uses agglomerative hierarchical clustering (AHC) with a metric defined by a probabilistic linear discriminant analysis (PLDA) model of i-vectors. Moreover, this metric is adapted to each conversation by means of a principal component analysis (PCA) subspace projection. We also present an AHC stopping criterion based on unsupervised calibration of the PLDA scores using unlabeled in-domain data. Finally, we also demonstrate an improvement in performance with denser sampling of the i-vector speaker distributions with overlapping temporal segmentation.

2. BACKGROUND

Using i-vectors to represent speakers has found great success in speaker recognition [1], and so the approach has broken into speaker diarization as well. However, unlike in speaker recognition tasks like NIST SRE [2], there is no guarantee that there is only one speaker in the recording (and, in fact, the opposite is essentially assured). The typical solution to this challenge is to perform an unsupervised segmentation of the audio into short (1-2 second) segments, and then extract i-vectors for these short segments.

One system that pre-dated i-vectors but should still be included here is presented in [3]. In this case, speaker factors (or eigenvoices), a precursor to i-vectors, were used as features. This system looked at a sliding window for a given conversation, diarizing one minute at a time, under the assumption that any minute of a conversation will have either 1, 2, or 3 speakers. The speakers are segmented with Gaussian models, and the number of speakers within this subset was determined by comparing log-likelihood ratios of the speaker models.

Based on the success of the system in [3], a similar system [4, 5] using total-variability features, or i-vectors, was developed. This system also further reduced the i-vector dimensions using a conversation-dependent PCA. Clustering of these reduced representations was performed with K-means or spectral clustering, both based on cosine score.

In [6], a GMM using Variational Bayes to apply priors to hyper parameter distributions (VB-GMM) is used for cluster-

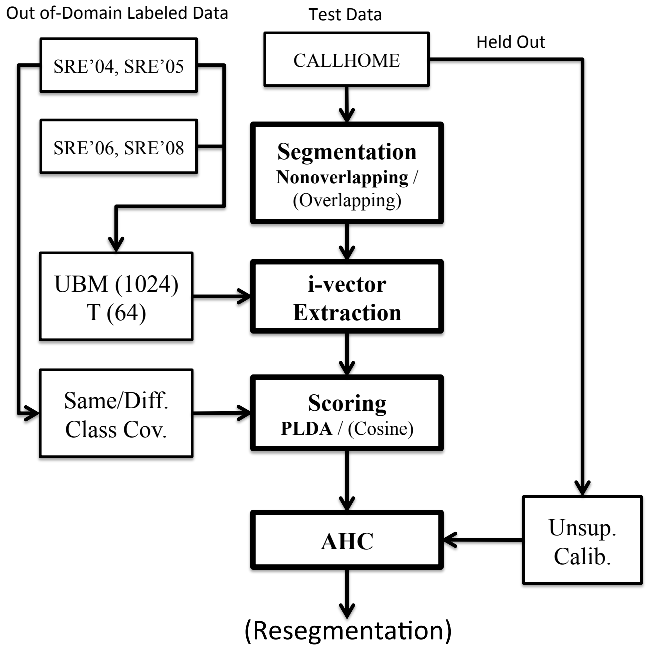

Fig. 1. System diagram for the speaker clustering system tested here. Tests show PLDA scoring with overlapping segments to give the best clusters for diarization, but cosine scoring and nonoverlapping segments are considered as well for comparison.

ing i-vectors. The system performs well for a large number of speakers, but is strangely less effective for the typically easier two-speaker conversations.

Another i-vector clustering algorithm that has worked well in diarization is mean shift [7], a method that finds the nearest mode for any given data point by iteratively finding the center-of-mass within a neighborhood and then redefining the neighborhood around the center-of-mass. However, this method also requires defining the size of a neighborhood (via the "bandwidth"), which is essential for determining the required size of a mode (or speaker, for diarization). The authors define this parameter for each conversation based on the duration. This method provides the best performance-to-date on the CALLHOME corpus (described in Section 4.1).

All the discussed methods are intended for diarization of conversations in which there is only one recording, such as those in conversational telephone speech (CTS) evaluations. It is important to note that other corpora, such as the NIST Rich Transcription data, approach the diarization task from a different angle, in which multiple recordings of the same conversation are available from different locations in the physical space. This version of the task has also seen a great deal of research (see [8] for an overview on a selection of these methods), but these algorithms typically include methods that are not available for single recording conversations, such as beamforming. As a result, we are only considering our work in the context of the methods discussed above.

3. DIARIZATION CLUSTERING SYSTEM

Our approach begins with a temporal segmentation into 1-2 second clips, and then i-vectors are extracted for each of these segments. The i-vectors are further reduced with a conversation-dependent PCA and then scored with PLDA using parameters estimated on separate labeled data. The i-vectors are then clustered with these scores using AHC, using a threshold learned on unlabeled data as the stopping criterion. A system diagram laying out this process is shown in Fig. 1, and each of these modules will be discussed in detail below.

It is important to note that the methods discussed in Section 2 also include resegmentation after clustering (or sometimes iteratively, as in [4]). In this work, we are focusing on improvements to the clustering stage of diarization, and, as a result, will not be considering the resegmentation. However, in a full system, such as those discussed above, resegmentation would be included after the i-vector clustering is completed.

3.1. Temporal Segmentation

We begin our diarization process by breaking the conversation into 1-2 second segments based on the speech activity detection (SAD) marks, as has been common in previous methods as well. However, this process provides a somewhat problematic foundation for i-vectors, because i-vectors are most effective when provided a longer speech section (on the order of a minute rather than a second). On the other hand, increasing the duration of the segments will reduce their number, which hinders the clustering.

We propose a denser sampling of the i-vectors by using overlapping segmentation. In our case, we employ 1-2 second segments with 500ms of overlap with its preceding and following segment (leading to a total of 1 second of overlap). This denser sampling allows for maintaining up to 2 second segments while providing the same number of samples as segmentation at half that length. We will show in Section 4.3 that this denser sampling improves our clustering.

It is important to note that a similar approach to overlapping segments was used in [3]. However, more recent systems have not included this technique, and overlapping and nonoverlapping segmentation have not been previously compared within the same clustering system.

3.2. i-vector Extraction

We use an i-vector extraction system trained on NIST SRE data ('04,'05,'06,'08). Our universal background model

Method	DER
Castaldo et al [3]	13.7
*Shum et al [6]	14.5
Senoussaoui et al [7]	12.1

Table 1. Results for several systems on CALLHOME. The (*) reflects that the results for Shum et al were estimated from plots displaying results per speaker.

(UBM) includes 1024 Gaussians trained on 20 cepstral coefficients for each audio frame (with no deltas). The total variability (T) matrix reduces to 64 dimensions. The i-vectors are also length-normalized [9]. Within each conversation, we also compute a conversation-dependent PCA, which further reduces the dimensionality to 3 dimensions.

3.3. AHC with PLDA metric

We cluster the i-vectors using AHC. Starting with each i-vector as a separate cluster, at every step, we merge the two clusters that are closest based on a predefined metric. This merging schedule defines a path over the space of partitions and a final clustering is obtained based on a stopping criterion (discussed in Section 3.4).

We use a PLDA system [9] to define the metric by computing a pairwise similarity matrix between all i-vectors [10] from a conversation. Then, the similarity between two clusters (i.e. linkage criterion) is defined as the average of the pairwise similarities between the elements of each cluster. Note that this approach only requires precomputing the pairwise similarity matrix once, and then simple averaging of scores during clustering, which is much more computationally efficient than recomputing the similarity matrix after every merge.

The PLDA system uses a 32 dimensional speaker space and was trained from SRE'04 and SRE'05 conversation sides that were chopped into 3 second segments. All the i-vectors from a conversation side were treated as a unique class and no PCA was applied to them, thus defining class as a speaker in a particular channel. Finally, within each conversation in the test data, the parameters of the PLDA system were projected using the conversation-dependent PCA matrix. This effectively resulted in a conversation-dependent metric.

3.4. Determination of number of clusters

To estimate the number of clusters, we define a threshold and stop the merging process when the similarity between the clusters to be merged goes below the threshold. A principled way of doing this is to calibrate the scores of the PLDA system so that we can use Bayesian decision theory to set a threshold analytically.

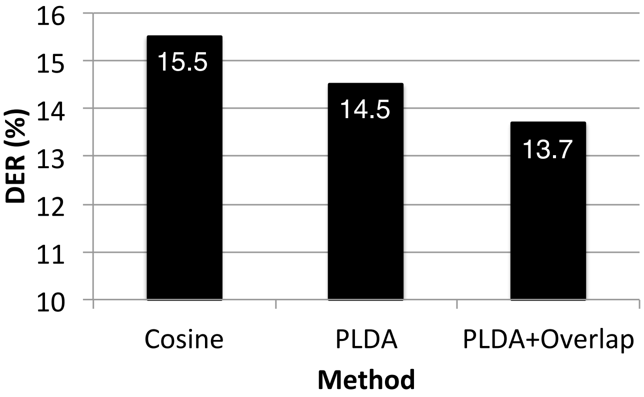

Fig. 2. DERs for each of the three clustering methods tested (no resegmentation): Cosine (baseline), PLDA scoring, and PLDA scoring with overlapping segments.

We use an unsupervised calibration approach [11] where only unlabeled in-domain scores are required. This approach uses a generative model of scores [12] and fits a 2 component Gaussian mixture model (GMM) to a collection of unlabeled in-domain scores. However, unlike in [11], where the proportion of expected scores within the same speaker is very small, we expect a more balanced prior, and so the covariances of the GMM do not need to be tied. As a result, the calibration mapping is not affine (but quadratic).

Once we learn a calibration mapping, we stop the AHC when the calibrated similarity between the clusters to be merged goes below 0. That is, when the evidence in favor of the different-speaker hypothesis, $\mathcal{H}_d$, exceeds the evidence in favor of the same-speaker hypothesis, $\mathcal{H}_s$.

4. EXPERIMENTS

4.1. Data

We evaluated our system using the CALLHOME corpus, which is a CTS collection between familiar speakers. Within each conversation, all speakers are recorded in a single channel. There are anywhere between 2 and 7 speakers (with the majority of conversations involving between 2 and 4), and the corpus also is distributed across six languages: Arabic, English, German, Japanese, Mandarin, and Spanish.

The CALLHOME corpus has been used to evaluate several of the systems discussed in Section 2. Their results are shown in Table 1. Note that, unlike our results, these results also include resegmentation.

4.2. Performance Metrics

We evaluated our methods with Diarization Error Rate (DER), a common metric for diarization. In its purest form, DER

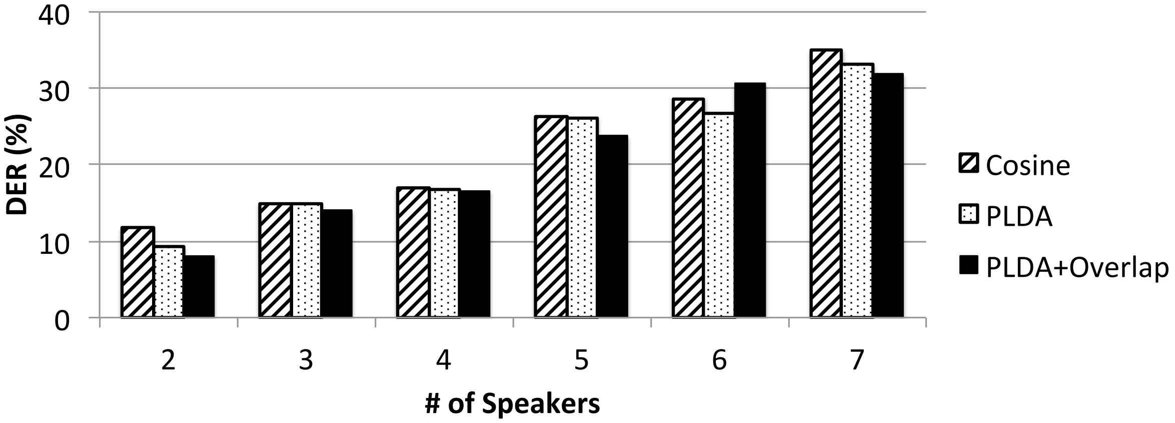

Fig. 3. DERs for each of the three methods organized by the number of speakers in a conversation. Note that the most significant gains for PLDA come with two speakers in a conversation, while improvements from overlapping segments are generally more consistent.

combines all types of error (missed speech, mislabeled non-speech, incorrect speaker cluster), but, as is currently the practice, we used oracle SAD marks. As a result, only incorrect speaker labeling factors into the DER.

Also, as is typical, our DER tolerated errors within 250ms of a speaker transition and ignored overlapping segments in scoring.

4.3. Results

All methods we will discuss used a conversation-dependent PCA and AHC for clustering. Unsupervised global calibration with untied covariances was also used in all cases. We will briefly discuss conversation-dependent calibration at the end of this section as well.

The first system we considered established our baseline with cosine scoring and non-overlapping segmentation of 1-2 seconds, and achieved a DER of 15.5%.

We improved this baseline by using PLDA scoring instead of cosine scoring. The PLDA, which was learned on SRE data from 2006 and 2008, found a 32 dimensional subspace within the original 64 i-vector dimensions. This subspace was then reduced to 3 dimensions based on each conversation's PCA dimensional reduction. This change in scoring improved the DER by a full percent, reducing it to 14.5%.

Utilizing denser sampling at segmentation further improved the DER to 13.7%.

These results can all be seen in Fig. 4.3. Alternatively, Fig. 4 shows the performance of each method by number of speakers. Looking at the results in this light indicate that the gains from both PLDA scoring and overlapping segmentation are being primarility seen for conversations with 2 speakers. This is especially the case for PLDA scoring, which has nearly identical performance to cosine for 3, 4, and 5 speakers. Note that there are very few conversations with 6 or 7 speakers, so these results are subject to greater variance.

The lack of improvement for PLDA for conversations with greater than 2 speakers is an interesting result, and indicates there is room for greater improvement with PLDA scoring if the improvements at 2 speakers can be seen at a greater number of speakers as well.

4.3.1. Conversation-dependent Calibration

We would ideally prefer to use conversation-dependent calibration, similarly to the conversation-dependent bandwidth used in [7]. However, the small number of samples for a given conversation (even with overlapping segmentation) lead to unstable calibration. An experiment with this approach yielded only 16.7% DER. However, it is our belief that conversation-dependent calibration could lead to large gains in the future, as an experiment using oracle calibration for each conversation resulted in 9.7% DER, a drop in 4 percentage points from global unsupervised calibration.

5. CONCLUSION

In this paper, we introduced the use of PLDA scoring for speaker clustering in diarization algorithms. Combined with the conversation-dependent PCA, this creates a scoring metric that is unique to each conversation. The new scoring improved clustering performance over the same system with cosine scoring, the traditionally more popular technique. Performance was further improved with denser sampling of the i-vector subspace with overlapping segmentations.

We also demonstrated that global unsupervised calibration provides a stable stopping criterion for clustering, but also that conversation-dependent calibration could lead to significant improvements if a viable approach is found.

This clustering algorithm on its own achieves near state-of-the-art performance on the CALLHOME corpus, and, when combined with resegmentation, should reduce error rates even further.

6. REFERENCES

[1] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[2] "The NIST Year 2010 Speaker Recognition Evaluation Plan," (Available at `http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf`), 2010.

[3] Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, and Claudio Vair, "Stream-Based Speaker Segmentation Using Speaker Factors and Eigenvoices," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[4] Stephen Shum, Najim Dehak, Ekapol Chuangsuwanich, Douglas Reynolds, and Jim Glass, "Exploiting Intra-Conversation Variability for Speaker Diarization," in *Proceedings of Interspeech*, 2011.

[5] Stephen Shum, Najim Dehak, and Jim Glass, "On the Use of Spectral and Iteratvie Methods for Speaker Diarization," in *Proceedings of Interspeech*, 2012.

[6] Stephen H. Shum, Najim Dehak, Réda Dehak, and James R. Glass, "Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–28, October 2013.

[7] Mohammed Senoussaoui, Patrick Kenny, Themos Stafylakis, and Pierre Dumouchel, "A Study of the Cosine Distance-Based Mean Shift for Telephone Speech Diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 217–27, January 2014.

[8] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–70, February 2012.

[9] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of Interspeech*, 2011.

[10] Daniel Garcia-Romero, Alan McCree, Stephen Shum, Niko Brümmer, and Carlos Vaquero, "Unsupervised Domain Adaptation for I-Vector Based Speaker Recognition," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2014.

[11] Niko Brümmer and Daniel Garcia-Romero, "Generative Modelling for Unsupervised Score Calibration," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2014.

[12] D. van Leeuwen and N. Brümmer, "The distribution of Calibrated Likelihood Ratios," in *Interspeech*, 2013.