DIARIZATION RESEGMENTATION IN THE FACTOR ANALYSIS SUBSPACE

Gregory Sell and Daniel Garcia-Romero

Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD, USA
{gsell,dgromero}@jhu.edu

ABSTRACT

Resegmentation is an important post-processing step to refine the rough boundaries of diarization systems that rely on segment clustering of an initial uniform segmentation. Past work has primarily used a Viterbi resegmentation with MFCC features for this purpose. In this paper, we examine an algorithm for resegmentation that operates instead in factor analysis subspace. By combining this system with a speaker clustering front-end, we yield a diarization error rate of 11.5% on the CALLHOME conversational telephone speech corpus.

***Index Terms*—** Speaker diarization, factor analysis, variational Bayes

1. INTRODUCTION

Speaker diarization is an important front-end process for any analysis of spoken audio. Most downstream processes, such as automatic speech recognition (ASR) or i-vector extractors for speaker/language recognition, assume the presence of only a single speaker, but, outside of controlled evaluations, this is a difficult condition to guarantee. So, for real-world audio, it is wise to run speaker diarization to segment the regions associated with each speaker prior to subsequent analyses.

However, the task of speaker diarization involves several challenges. The process is typically completely unsupervised, in that no information is known a priori regarding the identity of the speakers or even the number of speakers in many cases. As a result, most diarization algorithms (more specifics below in Section 2) utilize an unsupervised clustering algorithm to estimate the number of speakers and an approximate set of boundaries around each speaker's regions of speech. These rough boundaries are then typically refined with a resegmentation stage, usually pairing a model of each speaker's features with temporal constraints on speaker transitions (such as with a Hidden Markov Model (HMM)).

In this paper, we examine the resegmentation stage of speaker diarization. Standard practice in past work has been to use a Viterbi resegmentation based on acoustic models built with mel-frequency cepstral coefficient (MFCC) features [1]. Here, we will instead utilize an HMM-based system that operates in the same factor analysis subspace as i-vectors, and show that this resegmentation outperforms the baseline system on the CALLHOME conversational telephone speech (CTS) corpus. The combined system also yields the lowest published error rates for the corpus at the time of this writing (11.5%).

2. BACKGROUND

Resegmentation for speaker diarization is most commonly seen after some form of segment clustering, which is performed by segmenting the audio based on speech activity detection (SAD), extracting some set of features for each segment, and then clustering those extracted features. In this section, we will provide a brief background of each of these stages.

Segmentation is typically the first stage of cluster-based speaker diarization algorithms, and is intended to divide the speech into short sections that are assumed to have a single or dominant speaker. The common practice is to divide the signal into utterance segments based on SAD marks. Any long speech blocks are further subdivided to 1-2 seconds.

Features are then extracted from these blocks, though the specific features have varied over time, generally matching the progression of the speaker identification community. Initial systems used MFCCs [1], followed by speaker factors [2], and then eventually i-vectors [3, 4, 5, 6, 7].

The segments are subsequently clustered according to these extracted features. Agglomerative Hierarchical Clustering (AHC) is one popular method [1, 7] because the clustering can be dictated by distance-based stopping criteria instead of assuming some number of speakers. Gaussian Mixture Models (GMMs) are also popular for clustering, though it is necessary to either constrain the number of possible speakers to a small subset [2] or apply a distribution to the model parameters [5]. Mean-shift provided state-of-the-art clustering results for speaker diarization in [6], while k-means and spectral clustering have also been explored [3, 4].

Resegmentation is the final stage of the process, in which the rough boundaries that were naively drawn for the initial segmentation are refined based on a frame-level, temporally-constrained process. The most common approach is a Viterbi

resegmentation with MFCC features [1]. In the next section, we will describe an algorithm for resegmentation in the factor analysis subspace, followed by results comparing this system to the more common Viterbi approach.

3. RESEGMENTATION IN THE FACTOR ANALYSIS SUBSPACE

Frame-level diarization in a factor analysis subspace was first proposed in [8]. For this system, the frame level statistics cannot be directly optimized, and so inference is required, in this case with Variational Bayes (VB). The subspace was originally defined with speaker factors, but i-vectors can be easily substituted without altering the updates in any other way. However, as it was proposed, the temporal continuity was not considered in the original framework, potentially ignoring a highly informative characteristic that speaker turns are typically much longer than a single frame and that speaker transitions are relatively rare.

At the 2013 Center for Language and Speech Processing (CLSP) Summer Workshop at Johns Hopkins University, the diarization system in [8] was extended to include an HMM to constrain the speaker transitions. In this version, the updates in [8] are identical except in estimating the speaker posteriors. In that case, the speaker log likelihoods for the HMM are computed (via eq. (29) in [8], excluding the prior term), and the HMM then defines the speaker posterior probabilities. Code for this extension (called VB diarization) is available online [9].

During our experimentation, we further extended this code by modifying the HMM transition matrix to account for non-speech sections. There is not a clear definition of non-speech in the factor analysis subspace, and so these sections are removed prior to running the algorithm. Our extension includes the presence of these regions by backing off to speaker prior estimates after a break in speech (resulting in a time-varying transition matrix). This avoids the potential for highly preferring one speaker after a long break because that speaker was active before the break. We found this extension to improve the diarization in many cases, and at worst has no effect.

Though the algorithm is intended to function as a standalone diarization system, some sort of initialization is required to begin the iterative process, with or without the HMM. In the past, this initialization has been assigned randomly, either at a frame level or by assigning random labels to entire blocks of segmented speech.

In this paper, we will instead consider initializing the VB diarization system with the labels estimated by segment clustering. Another way to consider this orientation is that the VB diarization system will serve as the resegmentation for the clustering algorithm.

For the clustering, we will use a recently developed clustering process utilizing AHC for i-vector scores computed with probabilistic linear discriminant analysis (PLDA) after a cut-dependent PCA. Stopping criteria for the clustering is determined with unsupervised calibration, and the segments themselves also overlap 50% with neighboring segments. Greater detail can be found in [7].

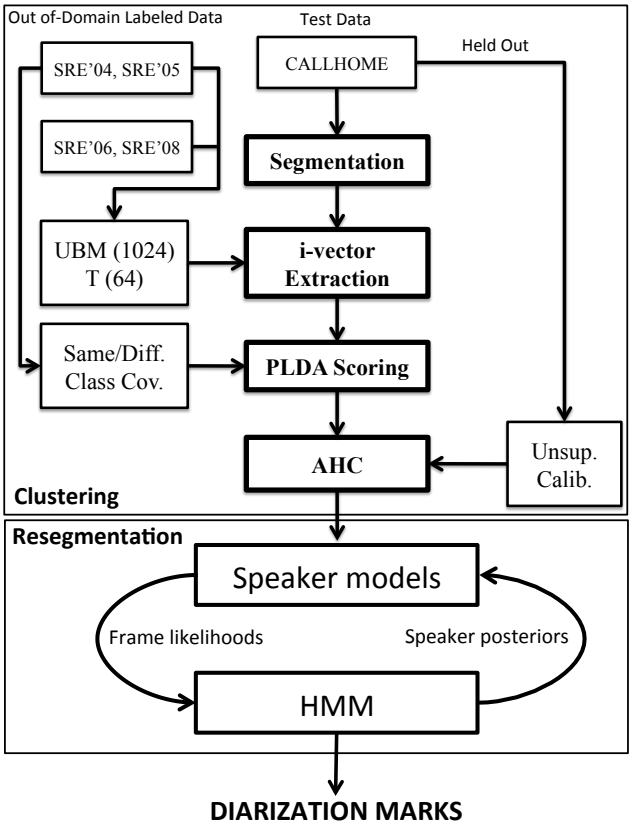

Fig. 1. System diagram for the speaker clustering system and VB resegmentation.

A system diagram of the full combination and best overall performing system is shown in Fig. 1.

4. EXPERIMENTS

4.1. Data

We evaluated the system combinations 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.

Method	DER
Castaldo et al [2]	13.7
*Shum et al [5]	14.5
Senoussaoui et al [6]	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.

4.2. Performance Metrics

We evaluated our methods with Diarization Error Rate (DER), a common metric for diarization. In its purest form, DER 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[1].

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

4.3. Results

Our experimental results can be considered in two categories. First, we examined the performance of our speaker clustering output from [7] when paired with each resegmentation system. Second, we examined the value of the HMM in the VB diarization extension.

4.3.1. Resegmentation

Results for cluster/resegmentation combinations are shown in Fig. 2. Viterbi resegmentation does not have any noteworthy effect on the DER, increasing it marginally from 13.7% to 13.8%. Alternatively, VB diarization (or VB resegmentation, in this case) improves the performance by over 2%, resulting in an overall DER of 11.5%.

These results can also be broken down by the number of speakers, as shown in Fig. 3. It appears from this view that the VB resegmentation provides a balanced improvement across speakers, rather than preferring a small or large number of speakers. The exception to this is with 7 speakers, where there is no improvement, but there are very few examples in CALLHOME with 7 speakers and so this difference could simply be a product of the small sample size. It is also interesting that Viterbi resegmentation makes modest improvements for 3, 4, and 5 speakers, but this is offset in the overall score by the damage the resegmentation does in the case of only 2 speakers, leading to the overall slightly worse DER.

[1]The Viterbi algorithm includes a non-speech class, and so the oracle SAD marks are sometimes adjusted in this resegmentation. In these cases, we only considered error within the oracle SAD speech regions, which are speaker error and missed speech, and ignored mislabeled non-speech.

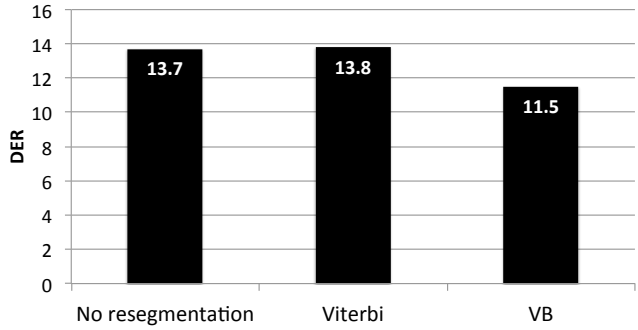

Fig. 2. Diarization error rates for several resegmentations after the clustering in [7]. Viterbi resegmentation has essentially no effect while VB resegmentation reduces the DER by over 2% to 11.5%.

4.3.2. HMM in VB Diarization

We also examined the effect of the HMM in the VB diarization algorithm, with several results shown in Fig. 4. First, we examine the DER for VB diarization with random block initialization (with blocks selected identically to cluster segments). Without the HMM, the diarization system is not particularly competitive, yielding a high DER of 27.5%. However, simply including the HMM, even with random block initialization, reduces the DER all the way to a reasonable 16.4%.

The difference between the HMM and no HMM system is less pronounced when initialized with speaker cluster labels, with DERs differing by only 4%. However, note that, without the temporal smoothing of the HMM, the resegmentation actually damages performance (since speaker clustering labels without any resegmentation yield 13.7% in Fig. 2, 1.8% better than the 15.5% DER after resegmentation with the HMM). With the inclusion of the HMM, the system combination performs at 11.5%, the lowest published DER for CALLHOME at the time of this writing (other published results are shown in Table 1).

5. CONCLUSION

In this paper, we considered the value of resegmentation after speaker clustering, examining both the typically used Viterbi resegmentation as well as a VB diarization system that operates in factor analysis subspace. We find that, for the particular speaker clustering utilized, Viterbi resegmentation contributes no additional improvement, while VB diarization improves the DER by over 2% absolute.

We also found that the HMM extension to VB diarization developed at the 2013 CLSP Summer Workshop improves performance for both random and speaker cluster label initialization, and, while the improvement is much greater in the

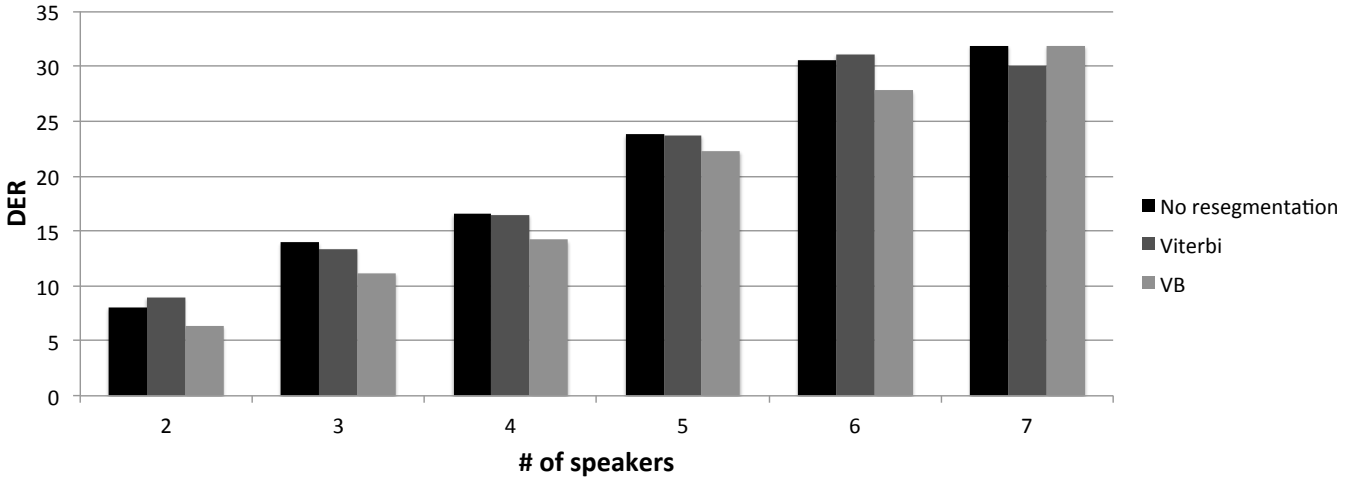

Fig. 3. Breakdown of the overall rates in Fig. 2 by number of speakers. VB resegmentation consistently improves DERs for all cases except 7 speakers.

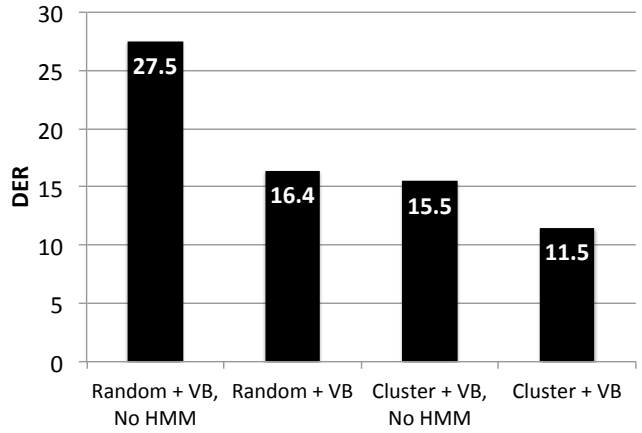

Fig. 4. Diarization error rates for several experiments examining the effect of the HMM in VB diarization. With random initialization, the inclusion of the HMM greatly improves the system. With initialization from cluster labels, the HMM provides comparably less but still significant improvement.

case of the former, the improvement in the latter case pushes the overall system to state-of-the-art performance.

The low DER yielded in this work is not surprising, in a sense, because it is the result of combining two already competitive systems. On their own, each performs at DERs competitive with other standalone systems (13.7% for the clustering, 16.4% for the VB diarization). Further gains at system combination are seen because these algorithms compliment each other very well. Speaker clustering is able to effectively extract overall patterns with no prior knowledge, but the boundaries of the speaker turns are necessarily rough due to naive segmentation. VB diarization has no such constraints about speaker transitions, but it requires initialization, and bad assumptions resulting from bad initialization are difficult to repair. By connecting the two systems, each compensates for the weaknesses of the other, resulting in a highly effective combination.

6. ACKNOWLEDGEMENTS

We would like to thank Lukáš Burget of Brno University of Technology for providing the code for VB diarization.

7. REFERENCES

[1] Patrick Kenny, Douglas Reynolds, and Fabio Castaldo, "Diarization of Telephone Conversations using Factor Analysis," *IEEE Journal of Special Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–70, December 2010.

[2] 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.

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

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

[5] 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.

[6] 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.

[7] Gregory Sell and Daniel Garcia-Romero, "Speaker Diarization with PLDA I-Vector Scoring and Unsupervised Calibration," in *Proceedings of the IEEE Spoken Language Technology Workshop*, 2014.

[8] Patrick Kenny, "Bayesian Analysis of Speaker Diarization with Eigenvoice Priors," Tech. Rep., Centre de Recherche Informatique de Montréal, 2008.

[9] Lukáš Burget, "Vb diarization with eigenvoice and hmm priors," availabe online at http://http://speech.fit.vutbr.cz/software/vb-diarization-eigenvoice-and-hmm-priors.