COMBINING SGMM SPEAKER VECTORS AND KL-HMM APPROACH FOR SPEAKER DIARIZATION

Srikanth Madikeri[1]*, Petr Motlicek*[1] *and Hervé Bourlard*[1,2]

[1] Idiap Research Institute, CH-1920 Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
srikanth.madikeri@idiap.ch, petr.motlicek@idiap.ch, herve.bourlard@idiap.ch

ABSTRACT

In this paper, a method to use SGMM speaker vectors for speaker diarization is introduced. The architecture of the Information Bottleneck (IB) based speaker diarization is utilized for this purpose. The audio for speaker diarization is split into short uniform segments. Speaker vectors are obtained from a Subspace Gaussian Mixture Model (SGMM) system trained on meeting data. The speaker vectors are clustered using the K-means algorithm. Two types of distance measures are explored in the clustering step: cosine distance of the speaker vectors and that of the vectors in a space projected by Probabilistic Linear Discriminant Analysis (PLDA). The clustering output is used as an initialization step for the Kullback Leibler-Hidden Markov Model (KL-HMM) based speech segmentation approach commonly used in the IB system for diarization. The proposed method is compared to clustering the segments using the IB based approach. A relative improvement of approximately 14% is obtained on the diarization performance for the proposed approach using SGMM speaker vectors with PLDA on the NIST RT 09 dataset.

Index Terms— SGMM, speaker diarization, speaker vectors, K-means

1. INTRODUCTION

Speaker diarization addresses the problem of identifying *who spoke when* in a speech recording [1]. Techniques such as the Hidden Markov Model/Gaussian Mixture Model (HMM/GMM) [2, 3] and the Information Bottleneck (IB) method [4] have been successfully applied to speaker diarization on meeting data. Approaches using the Bayesian Information Criterion (BIC) [5, 6, 7] and i-vector based approaches have been shown to be useful [8, 9] on broadcast news recordings and telephone conversational recordings.

Diarizing speech involves unsupervised segmentation and clustering of speakers. A common approach to obtain initial segmentation is to uniformly divide the entire speech into segments of equal length. In the HMM/GMM approach the segments are obtained by splitting the entire speech audio into a fixed number of segments (typically 16). In the IB approach however, the length of the segments are shorter (around 2.5s) compared to the initial segments in the HMM/GMM approach. Each segment is modelled by a Gaussian distribution. The distribution parameters are estimated from these short segments assuming the segments belong to only one speaker. However, the estimates depend on the recording conditions (eg: accuracy of beamforming for meeting recordings, recording types, etc.). Moreover, due to the short length of the segments, the speech information could dominate speaker identity during model estimation. Using prior information, such as a Universal Background Model (UBM), has been observed to provide little or no improvement over estimating segment-level Gaussians as the latter preserve time information in the audio. This motivates investigating alternative approaches that can use prior data and adapt to the observed features in the input audio.

In this paper, the SGMM approach is exploited to estimate speaker models for every segment of audio as it provides explicit factorization of speech and speaker information in its models [10, 11]. We assume that the short segments contain only one speaker to estimate speaker models. The speaker parameters from the SGMM approach have already been used in the context of language identification in [12]. Every segment obtained in the segmentation phase of the IB system is used to estimate one speaker vector from the SGMM system, which is trained on a development set with transcribed audio. The speaker parameters that represent each segment are clustered as opposed to clustering segment-level posteriors in the IB approach. These vectors are clustered to provide an initialization for the KL-HMM segmentation algorithm used as the final step in the IB system. In this paper, K-means algorithm is used for clustering [13]. K-means algorithm has been used for speaker diarization on telephone conversations in [8]. The proposed system is tested on the NIST RT 09 benchmark dataset. Its performance is compared with that of the IB approach. Additionally, the proposed approach is compared to an alternative i-vector approach by replacing i-vectors instead of SGMM speaker vectors in the proposed framework. It should be noted that i-vector-based approaches have not been

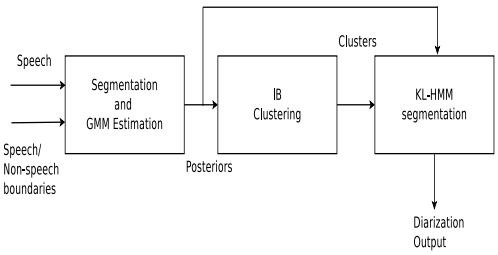

Fig. 1. Block diagram of the IB based diarization system.

extensively tested in meeting environments. It is hypothesized that the SGMM approach will perform better compared to the i-vector approach as it is exploiting the phonetic contents of the audio explicitly.

The rest of the paper is organized as follows: Section 2 describes the IB method. Section 3 details the speaker vector model in the SGMM approach. Section 4 describes the architecture of the proposed system. The results of the experiments on the NIST RT datasets are discussed in Section 5.

2. INFORMATION BOTTLENECK METHOD

The architecture of the IB approach is given in Figure 1. Uniform segments from speech data are modelled by a GMM with shared covariance parameters and a mean for every segment. The segments are clustered using the IB criterion ([14]) with the posteriors for speech features obtained from the GMM.

The clustering output is used as an intialization step to the KL-HMM segmentation algorithm. The KL-HMM segmentation algorithm reuses the posteriors and initalizes the HMM states with the mean of the posteriors in the cluster. Then, Viterbi decoding is applied to speech with respect to the state models and the posteriors. Kullback Leibler (KL) divergence between the speech posteriors and the state models are computed and the overall KL-HMM measure is minimized in the decoding process. KL-divergence is computed between the frames $\mathbf{y}_t = [y_{t,1} \dots y_{t,D}]^{\mathrm{T}}$ and state model $\mathbf{m}_i = [m_{i,1} \dots m_{i,D}]^{\mathrm{T}}$ of state i, where the posterior is D-dimensional. The KL divergence measure is given by:

$$v_{t,i} = -\sum_{d=1}^{D} y_{t,d} \log \left(y_{t,d} / m_{i,d} \right). \tag{1}$$

To compute KL divergence, the mean posterior vector is used as a reference while the speech frame posteriors is used as the test vector. In this paper, the IB architecture is modified to provide clustering output of speaker vectors obtained from the SGMM system to the KL-HMM segmentation algorithm.

3. SGMM SPEAKER VECTOR

The SGMM method is an acoustic modeling approach in which a common GMM structure is shared across all the phonetic states. Each state is represented by a state vector that defines a mapping to the means and weights of the state's GMM. Let $\mathbf{x}$ be a F-dimensional feature, j represent a speech state, $\mathbf{v}_j$ the S-dimensional state vector. The model of a state is defined by the following equations:

$$p\left(\mathbf{x}|j\right) = \sum_{i=1}^{I} w_{ji} \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_i\right), \tag{2}$$

$$\boldsymbol{\mu}_{ji} = \mathbf{M}_i \mathbf{v}_j, \tag{3}$$

$$w_{ji} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_j}{\sum_i^I \exp \mathbf{w}_i^T \mathbf{v}_j}, \tag{4}$$

where I is the number of Gaussians in the state. $\mathbf{M}_i$ and $\mathbf{w}_i$ are globally shared parameters. Typically, S is much less than $I(F+1)$ and hence the model is called "subspace" GMM. Each state j has M_j substates as S is less than the total number of globally shared parameters. The substates have their own mixture weights c_{jm} and vector $\mathbf{v}_{jm}$. The above three equations now become:

$$p\left(\mathbf{x}|j\right) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_{jmi}, \boldsymbol{\Sigma}_i\right), \tag{5}$$

$$\boldsymbol{\mu}_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}, \tag{6}$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_i^I \exp \mathbf{w}_i^T \mathbf{v}_{jm}}, \tag{7}$$

3.1. Speaker vector extraction

Speaker specific parameters in the SGMM system can be obtained by decomposing $\boldsymbol{\mu}_{jmi}$ into speech specific and speaker specific parameters. That is,

$$\boldsymbol{\mu}^s_{jmi} = \mathbf{M}_i \mathbf{v}_{jm} + \mathbf{N}_i \mathbf{v}^{(s)}. \tag{8}$$

The $\mathbf{N}_i$ matrices define the speaker subspace and $\mathbf{v}^{(s)}$ is the speaker vector for $\boldsymbol{\mu}^s_{jmi}$. The above equation can be compared to Joint Factor Analysis of speaker and channel subspace in speaker recognition. In this case, the speech and speaker subspaces are being separated.

In the context of speaker diarization, the estimation of $\mathbf{v}^{(s)}$ can be interesting. The SGMM approach provides a method to estimate speaker specific parameters from the speech. This is useful for several reasons. The SGMM system itself requires fewer parameters than a GMM system. Importantly, we obtain a small fixed-dimensional representation of a speaker that is direct mapping to a GMM. In this work we use a 39-dimensional speaker representation. This dimension is related to the feature dimension used to train the SGMM based ASR system.

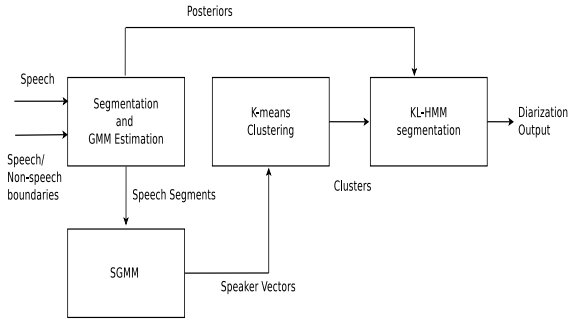

Fig. 2. Architecture of the proposed system that uses SGMM speaker vectors.

3.2. Speaker vector whitening and PLDA

The speaker vectors obtained are proposed to be clustered for speaker diarization. In this work, the speaker vectors are used in two ways: (i) the speaker vectors obtained from the SGMM system are whitened and (ii) the whitened speaker vectors used in (i) are projected in PLDA (Probabilistic Linear Discriminant Analysis) space. Whitening the speaker vectors Gaussianizes the vectors for K-means clustering. The PLDA parameters are trained on a development dataset. The G-PLDA model (Gaussian-PLDA) is a commonly used technique in speaker recognition [15]. Given a speaker vector $\mathbf{v}_r^s$ for speaker s, the G-PLDA model is given by

$$\mathbf{v}_r^s = \boldsymbol{\mu}_v + \boldsymbol{\Phi}\mathbf{y} + \epsilon_r^s, \tag{9}$$

where $\boldsymbol{\mu}_v$ is the mean of the speaker vectors, $\boldsymbol{\Phi}$ defines the speaker space and ϵ_r^s is the channel noise. The PLDA hyperparameters are Φ and the covariance of the residue $\boldsymbol{\Sigma}_P$. The projection ($\hat{\mathbf{v}}_r^s$) of $\mathbf{v}_r^s$ onto $\boldsymbol{\Phi}$ is computed using the single model assumption for the PLDA system:

$$\hat{\mathbf{v}}_r^s = \left(\mathbf{I} + \boldsymbol{\Phi}^t\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^t\boldsymbol{\Sigma}_P^{-1}\mathbf{v}_r^s, \tag{10}$$

where $\mathbf{I}$ is an identity matrix. The architecture of the proposed system is given in the next section.

4. SGMM FOR DIARIZATION

The IB system's architecture is modified to use the speaker vectors obtained from the SGMM system. Instead of using the posterior representation of segments for clustering, the speaker vectors are clustered using K-means algorithm with Euclidean distance as distance measure between vectors [13]. The value of K is empirically decided.

The overall architecture of the proposed system is shown in Figure 2. Similar to the IB approach, the segmented speech is modelled by a GMM, where each segment forms the mixture of the GMM. The segment boundaries are passed to the SGMM system. The SGMM system estimates a speaker vector for every segment generated. The speaker vectors are clustered using the K-means algorithm. The clusters generated are used to initialize the KL-HMM segmentation algorithm (described in Section 2).

In the proposed method, multiple iterations of the KL-HMM is required (as opposed to only one in the IB approach). The modification is required as it is observed that the clustering output produced by the K-means algorithm has worse speaker error rate before resegmentation compared to the clustering output from the IB approach. However, only few reiterations of segmentation and modelling are necessary (typically 10 compared to only 1 in the IB approach). Also, in the IB approach reiteration of segmentation and modelling does not improve the performance of the system. The KL-HMM system uses the posteriors produced in the initial step of the process along with the clustering output from the speaker clustering algorithm. The modified boundaries are given as the diarization output.

In the architecture described above, the SGMM method can be replaced by other methods that can provide speaker representations. In our experiments the performance of SGMM vectors and i-vectors, which is the state-of-the-art speaker modelling technique in speaker recognition, are compared. We refer the reader to [16] for details on the i-vector approach and to [17] for details on i-vector system implementation .

5. EXPERIMENTS

Speaker diarization experiments are performed on the NIST RT 05, 06 and 2009 benchmark datasets. The NIST RT05 and RT06 are used as a development dataset while RT09 forms the test set. The development set is used to tune number of clusters (K value) and train PLDA parameters. Multiple Distant Microphone (MDM) recordings are used for the experiments after their enhancement using *Beamformit* [18]. The proposed diarization system is compared with two other diarization systems: the IB based diarization system and a modification of the proposed system in which i-vectors of the segments are used instead of SGMM speaker vectors for clustering.

5.1. System parameters

MFCC (Mel Frequency Cepstral Co-efficients) features are extracted from the audio at 10ms frame rate with a window size of 25ms. A Gaussian is modelled for every 250 frames and the covariance matrix is shared across the Gaussians. The posteriors are estimated for every frame with respect to all these Gaussians. The speech/non-speech segmentation is common for all systems used in this work and is derived from ground truth.

The SGMM system is trained as follows: the shared parameters are trained on the AMI corpus [19] with 39 dimensional MFCC features (including delta and delta-delta) on the SDM+IHM meetings. The SDM+IHM meetings are used for SGMM training because it is observed to generalize better (in terms of performance in mismatch conditions) than systems

Table 1. Comparison of performance of 3 clustering algorithms: IB clustering with posteriors, K-means clustering with SGMM speaker vectors and K-means clustering with i-vectors. The experiments are performed on NIST RT 06 dataset. SER: Speaker Error Rate, +PLDA: vectors projected in the PLDA space.

Clustering algorithm	SER
IB	20.5
i-vector	55.7
i-vector+PLDA	54.8
SGMM	62.4
SGMM+PLDA	61.8

Table 2. Results of experiments conducted on the NIST RT 06 and 09 datasets comparing the IB clustering and speaker vector (before and after PLDA) clustering methods SER: Speaker Error Rate, +PLDA: vectors projected in the PLDA space.

System/Dataset	RT06 (SER)	RT09 (SER)
Baseline (IB)	18.5	22.9
i-vector	28.4	24.2
i-vector + PLDA	25.9	**21.3**
SGMM	24.8	**19.9**
SGMM + PLDA	**18.4**	**19.7**

trained on individual conditions. The system is trained with 4000 states and 120 substates. The WER (Word Error Rate) on the test corpus in the AMI set is 63.4% on SDM recordings and 41.9% on IHM recordings.

The i-vector system is also trained on the AMI corpus. The Universal Background Model (UBM) is trained with 19-dimensional MFCCs. The T-matrix is estimated with the convetional EM algorithm ([20]) for 10 iterations. The i-vector dimension is set to 60. The PLDA parameters for the SGMM speaker vectors and i-vectors are trained on the NIST RT05 data set. The data set has 50 speakers. For PLDA, 40 and 20 dimensions are retained for i-vector and SGMM systems, respectively. The optimal value of K in K-means clustering is set to 10 using the development set.

5.2. Results

The system parameters (PLDA dimension and stopping criterion) are optimized on RT06 and the systems are tested on RT 06 (development set) and RT09 datasets (test set). NIST RT 07 is used as a validation set during the development of the SGMM system and hence is not used to test the speaker diarization performance. The results of speaker clustering are presented in Table 1. The results suggest that output of speaker clustering with both i-vectors and SGMM speaker vectors are noisy compared to the IB clustering output and hence requiring more iterations for the resegmentation step. The performance improvement obtained after applying PLDA suggests that the technique is useful. However, the gains obtained are not as much as that observed in speaker recognition experiments where the amount of data available for training is much higher (typically thousands of speakers as opposed to only 50 used here). Particularly, the gains are beneficial for RT06 than RT09 as the former has more speakers.

The results of experiments on the RT datasets on the complete systems are presented in Table 2. The IB system is compared with the methods that use i-vector and SGMM speaker vector. In general, the SGMM speaker vector based approach is better than the i-vector and IB based approaches. The proposed system can be seen to provide an absolute improvement of 0.1% in terms of Speaker Error Rate (SER) on the development set (RT06). There is no improvement for the whitened speaker vectors but minor improvement is observed after applying PLDA. However, applying PLDA gives an improvement of 6.4% in absolute terms as the performance is optimized on the development set. The i-vector system however performs consistently poor compared to the IB system as well. The performance of the i-vector system is expected as the length of the segments used to estimate i-vector is short, while the i-vector system is trained on long segments (which is also the case for the SGMM system). However, in both cases PLDA projected vectors provide improvements. In the best case, 2.9% improvement is observed in absolute terms for the i-vector PLDA system.

In the test set (RT09), performance improvements of 3.0% and 3.2% in absolute terms are obtained before and after applying PLDA compensation, respectively. The SGMM method is therefore shown to provide benefits compared to both the baseline IB approach for speaker diarization and using i-vectors instead of the SGMM speaker vectors.

6. SUMMARY

A speaker diarization system using SGMM speaker vectors and KL-HMM segmentation is presented. The speaker vectors are estimated on short segments of speech in the input audio. The vectors are clustered and the output is used to initialize the states of the KL-HMM. The KL-HMM adjusts the boundaries with posteriors computed from the the Gaussians representing the short speech segments in the audio. The approach is compared with the IB approach that shares the same architecture. The SGMM speaker vectors clustered using K-means clustering is shown to perform better than the IB system. A relative performance improvement of up to 14% is observed.

7. ACKNOWLEDGEMENT

This work was supported by project Diarizing Massive Amounts of Heterogeneous Audio (DIMHA) and EU FP7 project Speaker Identification Integrated Project (SIIP). The authors would like to thank Mathew Magimai-Doss for his valuable comments on the paper.

8. REFERENCES

[1] Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] Jitendra Ajmera and Chuck Wooters, "A robust speaker clustering algorithm," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2003*. IEEE, 2003, pp. 411–416.

[3] Chuck Wooters and Marijn Huijbregts, "The icsi rt07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*. 2008, pp. 509–519, Springer.

[4] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1382–1393, 2009.

[5] Scott Chen and Ponani Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*. Virginia, USA, 1998, p. 8.

[6] Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *INTERSPEECH*, 2013.

[7] Sylvain Meignier and Teva Merlin, "LIUM SpkDiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, 2010, vol. 2010.

[8] Stephen Shum, Najim Dehak, Ekapol Chuangsuwanich, Douglas A Reynolds, and James R Glass, "Exploiting intra-conversation variability for speaker diarization.," in *INTERSPEECH*, 2011, pp. 945–948.

[9] Stephen Shum, Najim Dehak, and Jim Glass, "On the use of spectral and iterative methods for speaker diarization," in *INTERSPEECH, Portland, Oregon*, 2012.

[10] Daniel Povey, Lukas Burget, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondrej Glembek, Nagendra K Goel, Martin Karafiát, Ariya Rastrow, et al., "Subspace gaussian mixture models for speech recognition," in *IEEE Intl. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 4330–4333.

[11] Daniel Povey, Lukáš Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ondřej Glembek, Nagendra Goel, Martin Karafiát, Ariya Rastrow, et al., "The subspace gaussian mixture modela structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[12] Oldřich Plchot, Martin Karafiát, Niko Brümmer, Ondřej Glembek, Pavel Matejka, and E de Villiers J Cernockỳ, "Speaker vectors from subspace gaussian mixture model as complementary features for language identification," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[13] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, Wiley India, 2007.

[14] Noam Slonim, *The information bottleneck: Theory and applications*, Ph.D. thesis, Hebrew University of Jerusalem, 2002.

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

[16] Najim Dehak, Patrick Kenny, Rda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," 2011, vol. 19(4), pp. 788–798, IEEE Tran. on Audio, Speech and Language Processing.

[17] Srikanth Madikeri, "A hybrid factor analysis and probabilistic pca-based system for dictionary learning and encoding for robust speaker recognition," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[18] Xavier Anguera, "Beamformit (the fast and robust acoustic beamformer)," http://www.xavieranguera.com/beamformit/.

[19] Thomas Hain, Lukas Burget, John Dines, Philip N Garner, Frantisek Grezl, Asmaa El Hannani, Marijn Huijbregts, Martin Karafiát, Mike Lincoln, and Vincent Wan, "Transcribing meetings with the amida systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[20] Ondrej Glembek, Lukas Burget, Pavel Matejka, Martin Karafiát, and Patrick Kenny, "Simplification and optimization of i-vector extraction," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4516–4519.