IDIAP RESEARCH REPORT

KL-HMM BASED SPEAKER DIARIZATION SYSTEM FOR MEETINGS

Srikanth Madikeri Hervé Bourlard

Idiap-RR-19-2015

JUNE 2015

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

KL-HMM BASED SPEAKER DIARIZATION SYSTEM FOR MEETINGS

Srikanth Madikeri[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, herve.bourlard@idiap.ch

ABSTRACT

In this paper, the Kullback-Leibler Hidden Markov Model (KL-HMMs) is applied for unsupervised diarization of speech. A general approach to speaker diarization is to split the audio into uniform segments followed by one or more iterations of clustering of the segments and resegmentation of the audio. In the Information Bottlneck (IB) approach to diarization, short uniform segments are clustered using the IB criterion followed by resegmentation with KL-HMM. The KL-HMM approach has been shown to be an effective resegmentation procedure in this respect. Thus, the potential of KL-HMM as an independent diarization system is explored where the uniform segments are clustered and segmented using a sequence of posteriors obtained from the audio with respect to a Gaussian Mixture Model (GMM). The segmentation is performed using KL divergence, while the Jensen Shanon (JS) divergence is used for clustering. The diarization procedure is stopped by applying a Normalized Mutual Information (NMI) based criterion between two consecutive clustering outputs. The proposed method is tested on the NIST RT datasets. A best case relative improvement of 30% is observed in terms of Speaker Error Rate (SER) on the NIST RT 09 dataset when compared with the IB approach.

Index Terms— Kullback Leibler divergence, Hidden Markov Models, speaker diarization

1. INTRODUCTION

A speaker diarization system segments an audio containing speech based on *who spoke when* [1]. The problem is commonly approached in an unsupervised fashion where the knowledge of the speakers is unavailable. Commonly used approaches on recordings of meetings include the Hidden Markov Model/Gaussian Model Mixture (HMM/GMM) approach [2, 3] and the Information Bottleneck (IB) approach [4]. On broadcast news data Bayesian Information Criterion-based (BIC) approaches have been applied [5, 6, 7]. On data such as dyadic telephone conversations that are typically used for speaker recognition, i-vector based approaches are shown to be useful [8, 9].

In the HMM/GMM approach to speaker diarization, the audio is split into long segments as an initialization step. The segments are modelled with a GMM. Multiple iterations of re-segmentation and re-estimation steps follow. After a fixed number of iterations, two clusters are merged based on the BIC [5] or other suitable criteria [10, 11, 12].

In the IB approach [4], short segments of the audio are clustered using the IB criterion. The clustering output is used as an input to a resegmentation algorithm. Two approaches can be used for resegmentation: HMM/GMM (Hidden Markov Model/Gaussian Mixture Model) segmentation and KL-HMM (Kullback-Leibler Hidden Markov Model) segmentation. The decoding step is crucial in determining the accuracy of the system as it corrects segment boundaries and acts as a final reclustering step. The HMM/GMM approach and KL-HMM approach are shown to give comparable results [13]. In cases when multiple features are available the KL-HMM approach is shown to perform better than the HMM/GMM approach for resegmentation [4]. In this paper, we investigate the role of KL-HMM based segmentation as an independent diarization system. The KL-HMM method is modified to perform multiple iterations of re-estimation and resegmentation followed by merging of the states of the HMM using Jensen Shannon divergence. The stopping criterion for the process is set using a Normalized Mutual Information-based measure. The systems are tested on the benchmark NIST RT datasets for speaker diarization. The rest of the paper is organized as follows: Section 2 discusses the role of KL-HMM based segmentation in the IB system. Section 3 introduces KL-HMM as an independent diarization system. The results of the experiments on the NIST RT data sets are presented in Section 4.

2. KL-HMM BASED SEGMENTATION

In this section, the details of the KL-HMM based decoding procedure are given. Apart from the IB-based speaker diarization systems, KL-HMMs have been applied to ASR tasks [14]. In particular, KL-HMMs have been shown to be useful when the resources for training are limited [15, 16]. That KL divergence is a non-linear distance measure (between two probability distributions) makes it particularly useful for com-

Table 1. Comparison of performance in terms of Speaker Error Rate (SER) at two stages of the IB system: before and after resegmentation with KL-HMM. Performance of RT05, 06, 07 and 09 are combined together.

System	SER (%)
Before KL-HMM	20.2
After KL-HMM	18.1

plex distributions. It is also useful that the method generalizes to different distributions on data as it models only the multinomial distribution on the posteriors.

The IB approach to speaker diarization has been shown to be successfull and comparable to the HMM/GMM approach [4]. In this approach, the audio file is split into small uniform segments. These segments are modelled using a Guassian distribution. The Gaussians are used to estimate posteriors for every frame of the speech signal. The posteriors are used as relevance variables in the IB clustering method [17]. The clusters obtained are passed through a segmentation algorithm, typically the KL-HMM approach, to smooth the boundaries. Using the KL-HMM approach has an advantage of reusing the posteriors available and the approach has a simpler modelling process.

In KL-HMM, the states representation are estimated by averaging posterior vectors in the state. Viterbi decoding is applied based on the current model estimates using KL-divergence between the speech frame posteriors and state models. One iteration of KL-HMM segmentation is shown to be sufficient to obtain optimal diarization performance [4]. The performance of the system before and after resegmentation with KL-HMM are compared in Table 1 to emphasize the importance of resegmentation. The diarization performance increases by 2.1% (absolute) after resegmentation. The benefit of the KL-HMM system for segmentation is evident. This motivates the investigation of the KL-HMM as an independent speaker diarization module similar to the HMM/GMM speaker diarization system. This requires the use of posteriors instead of speech features (as in the case of HMM/GMM systems). Therefore, the KL-HMM approach to speaker diarization is proposed by modifying the IB approach: the IB clustering is entirely avoided and the are clusters initialized in a flat start method (as often done in the HMM/GMM approach). Then, multiple iterations of resegmentation and state model estimation are performed. Two states are merged after a fixed number of iterations of resegmentation and estimation. This process of resegmentation-estimation-merging is continued until a stopping criterion is met. A Normalized Mutual Information-based (NMI) stopping criterion is defined for this purpose.

3. KL-HMM BASED DIARIZATION

In this section, the KL-HMM based diarization system is introduced. First, the overall architecture is presented followed by the definition of the stopping criterion.

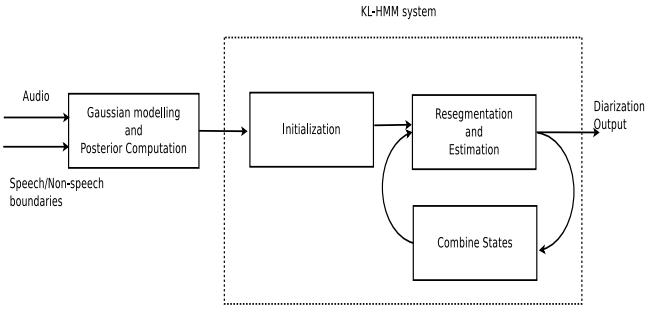

Fig. 1. Block diagram representing KL-HMM based diarization system. Multiple iterations of resegmentation and estimation are performed before merging two states

3.1. Proposed system

The architecture of the proposed system is presented in Figure 1. There are 4 steps involved in the diarization procedure:

1. *Initialization*: Initialize the states of the KL-HMM by computing the means of the posteriors of the states. The initialization is uniform as in the HMM/GMM systems. The approach to obtain posteriors is the same as that used in the IB system.

2. *Resegmentation*: Based on the current models of the HMM states, KL-divergence is computed between the speech frame posterior $\mathbf{y}_t = [y_{t,1} \dots y_{t,D}]^{\mathrm{T}}$ and state model $\mathbf{m}_i = [m_{i,1} \dots m_{i,D}]^{\mathrm{T}}$ of state i for all speech frames, where the posterior is D-dimensional. The KL-divergence measure is given by:

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

   The reference distribution is that of the state and the frame posterior forms the test distribution. The divergence uses a negative sign as the global cost is being maximized. The diarization procedure stops here when the stopping criterion (discussed later) is met.

3. *Re-estimation*: Based on the segmentation derived from the previous step, the models of the states are re-estimated. The model of a state i is given by

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{t \in i} \mathbf{v}_t, \quad (2)$$

   where n_i is the number of speech frames in state i. Steps 2 and 3 are repeated and allowed to converge. In practise, 5 iterations are observed to be sufficient.

4. *Clustering*: After multiple iterations of the Steps 2 and 3, two states are merged using the Jensen-Shannon (JS) divergence. The merging criterion is similar to the one

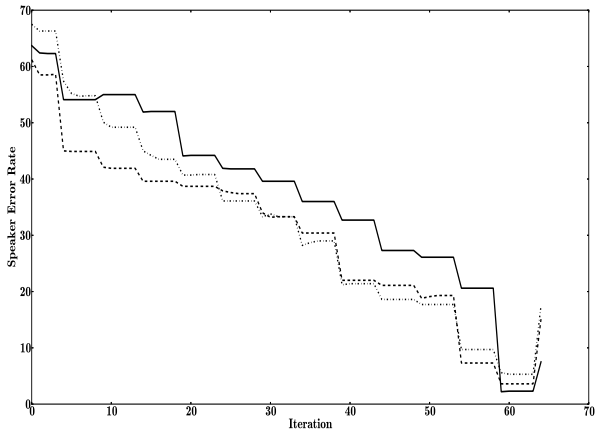

Fig. 2. The plot shows the reduction in Speaker Error Rate with each iteration of the KL-HMM system. Initially, there are 16 states and at each state, there are 5 iterations

used in the IB method. If two states are modelled by $\mathbf{m}_i$ and $\mathbf{m}_j$, the JS divergence is given by

$$JS(\mathbf{m}_i, \mathbf{m}_j) = \sum_{d=1}^{D} \pi_i KL(\mathbf{m}_i, \mathbf{m}_{ij}) + \pi_j KL(\mathbf{m}_i, \mathbf{m}_{ij}), \tag{3}$$

where $\pi_i = \frac{n_i}{n_i+n_j}$, n_i is the number of feature vectors assigned to cluster i and $\mathbf{m}_{ij} = \pi_i \mathbf{m}_i + \pi_j \mathbf{m}_j$ is the combined mean of the two states. The KL divergence between two posterior vectors is given by

$$KL\left(\mathbf{m}_i, \mathbf{m}_j\right) = \sum_{d=1}^{D} m_{i,d} \log\left(m_{i,d}/m_{j,d}\right) \tag{4}$$

Equation 3 is computed between all states and the pair with minimum JS divergence is merged.

3.2. NMI-based criterion

As the KL-HMM system is applied in an unsupervised environment, the choice of stopping criterion is critical in determining its performance. In this paper, the Normalized Mutual Information (NMI) [18] based criterion is used as the stopping criterion. This is similar to the criterion used the IB system.

Figure 2 helps illustrate that a stopping point exists. It shows that the error rate desirably decreases after every iteration. The existence of minima suggests that applying appropriate threshold to an appropriate stopping criterion can provide automatic procedure to stop the algorithm.

The NMI is calculated between two consecutive clustering outputs obtained from two consecutive iterations even if they have different number of clusters. Given two clustering outputs $C_1 = \{c_{1,1}, c_{1,2}, \ldots, c_{1,N_1}\}$ and $C_2 = \{c_{2,1}, c_{2,2}, \ldots, c_{2,N_2}\}$, the criterion is computed as

$$N\hat{M}I(C_1, C_2) = \sum_{a \in C_1, b \in C_2} n_{a,b} * \log \frac{N n_{a,b}}{n_a n_b} \tag{5}$$

where n_a and n_b are the number of data points in cluster i and j and $n_{i,j}$ is the set intersection of the two clusters and N is the total number of points. The hat is used on NMI to suggest that the criterion is a simplified version of NMI. As the number of iterations increase, the NMI decreases. Thus, applying a threshold on the minimum NMI value provides a stopping criterion.

4. EXPERIMENTS

Speaker diarization experiments are performed on the NIST RT 05, 06, 07 and 2009 benchmark datasets. The NIST RT05 is used as a development dataset while others form the test set. The development set is used to tune the NMI threshold and optimum number initial clusters. Multiple Distant Microphone (MDM) recordings are used for the experiments after their enhancement using *Beamformit* [19].

4.1. System parameters

MFCC features are extracted from the audio at 10ms frame rate with a window size of 25ms. A Gaussian is modelled for every 250 frames. The covariance matrix is shared across the Gaussians. The posteriors are estimated for every frame with respect to the Gaussians. The initial number of states is set to 16 after optimization on RT05 dataset for best performance. The minimum state duration for KL-HMM is set to 350 frames.

Two variants of the KL-HMM are developed based on the posteriors used for resegmentation: (i) in *smooth* posterior approach the posteriors obtained from the Gaussians are retained as it is (same as that used in the IB system) (ii) in *hard* posterior approach the top scoring Gaussian is set to 1.0 in the posterior vector for every frame while the rest are set to 0.0. However, for the estimation of the means of the states in both cases, only smooth posteriors are used as retaining uncertainty is observed to be useful for modelling.

4.2. Results

The results of experiments on the RT datasets are reported in Table 2. The KL-HMM system performance is compared with the HMM/GMM system and IB system. The comparison with IB is important as KL-HMM has already been applied in this context. Two types of results are compared: (i) oracle results, in which the best achieved diarization result is presented and (ii) NMI results that are obtained by using NMI as stopping criterion. The oracle results are presented to show the effectiveness of the approach when the optimal number of clusters for the procedure is known. The results show the best

Table 2. Results of experiments conducted on the NIST RT datasets are presented. The relative improvements are given with respect to IB (non-Oracle) system. Smooth posteriors: posteriors computed from all Gaussians; Hard posteriors: highest scoring Gaussians set to 1.0; Oracle: performance with optimal number of clusters; NMI: system uses NMI; SER: Speaker Error Rate, R.I. : Relative Improvement

	Dev. set		Test set					
System/Dataset	RT05	R.I.	RT06 (SER)	R.I. (%)	RT07 (SER)	R.I. (%)	RT09 (SER)	R.I. (%)
HMM	11.9	-	15.4	-	6.4	-	14.5	-
IB (Oracle)	14.8	-	16.4	-	9.4	-	22.0	-
IB	18.7	-	18.5	-	13.6	-	22.9	-
Smooth posteriors								
KL-HMM (Oracle)	9.5	49.1	15.6	15.7	10.1	18.3	15.8	30.0
KL-HMM (NMI)	16.7	10.6	20.5	-	14.6	-	19.3	15.7
Hard posteriors								
KL-HMM (Oracle)	8.9	52.4	14.5	21.6	9.0	33.8	15.9	30.0
KL-HMM (NMI)	14.4	23.0	18.4	0.5	14.0	-	21.2	7.4

possible performance that can be achieved with the KL-HMM system given the initialization procedure and input posteriors.

The posteriors used for the IB system and KL-HMM are the same. From Table 2, it is clear that in the majority of the cases, the KL-HMM system performs better than the IB system. The KL-HMM system with hard posteriors performs better than the system with smooth posteriors. This is perhaps due to the noisy posterior values in the posterior vector. However, thresholding the posterior values as an intermediate approach between smooth and hard posterior approaches did not give appreciable improvement over smooth posterior approach. Moreover, the binary approach to hard posteriors avoids sorting issues with the posterior values for every frame of speech. Finally, it also emphasizes the need for better methods to estimate posteriors.

The oracle systems show the potential of the KL-HMM as an independent diarization module. A best case relative improvement of 23.0% on the development set and 30.0% on the test set are observed compared to the IB system. The systems that use the NMI criterion perform worse than the oracle systems. In best case (RT09), a relative improvement of 15.7% is observed. This is useful as RT09 has more speakers per meeting than the other datasets. Thus, the methods scales well with the number of speakers and the audio length. Both smooth and hard posterior methods perform poorly on the RT07 dataset.

On the development set, the NMI gets the accurate number of speakers 40% of the time. The IB strategy, however, fixes the number of clusters to 10. The HMM/GMM clustering's accuracy is only 20% with an average estimation error by 1.1 speakers.

4.3. Runtime

We now compare the speed of the three algorithms: HMM/GMM, IB and KL-HMM. The Real Time Factors (RTF), defined as the ratio between the run time and length of the audio, are given in Table 3. The IB system is the fastest among the three systems. A straightforward implementation of the KL-HMM system is only slower than the IB system by a factor of 1.6, but is faster than the HMM/GMM system by factor of **4.5**. Thus, the KL-HMM system performs better than the IB system with a small trade-off in speed.

Table 3. Comparison of runtimes of the algorithms used in the paper.

System	RTF
HMM/GMM	0.45
IB	0.08
KL-HMM (Soft posteriors)	0.16
KL-HMM (Hard posteriors)	0.13

5. CONCLUSION AND FUTURE WORK

The KL-HMM system for speaker diarization is presented. The system is tested as an independent diarization module on the NIST RT datasets and its performance is compared with that of the IB system. The KL-HMM system's performance is shown to be better than that of IB system with best case relative improvement of 30%. However, the added advantage of the speed of the KL-HMM, in which the it is better than HMM/GMM system by a factor of 4.5, presents a reasonable trade-off. The NMI based stopping criterion provides an automatic way to chose the optimal number of clusters. This is observed to be relatively better than the BIC based criterion used in the HMM/GMM by 50%.

The framework presented in this paper opens up ways to use information theoretic measures while decoding for speaker diarization. Measures such as symmetric KL divergence, Jensen-Shannon divergence can be used with suitable modelling strategies for the HMM's states.

6. 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.

7. 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, pp. 1477–1481.

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

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

[9] 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.

[10] Herbert Gish, M-H Siu, and Robin Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*. IEEE, 1991, pp. 873–876.

[11] Perrine Delacourt and Christian J Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech communication*, vol. 32, no. 1, pp. 111–126, 2000.

[12] Matthew A Siegler, Uday Jain, Bhiksha Raj, and Richard M Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Broadcast News Workshop*, 1997, p. 11.

[13] Deepu Vijayasenan and Fabio Valente, "Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings.," in *INTERSPEECH*, 2012.

[14] Guillermo Aradilla, Hervé Bourlard, and Mathew Magimai-Doss, "Using KL-based acoustic models in a large vocabulary recognition task," in *INTERSPEECH*, 2008, pp. 928–931.

[15] David Imseng, Hervé Bourlard, and Philip N Garner, "Using kl-divergence and multilingual information to improve asr for under-resourced languages," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4869–4872.

[16] David Imseng, Ramya Rasipuram, and Mathew Magimai-Doss, "Fast and flexible kullback-leibler divergence based acoustic modeling for non-native speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 348–353.

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

[18] Christopher D Manning and Hinrich Schütze, *Foundations of statistical natural language processing*, MIT press, 1999.

[19] Xavier Anguera, "Beamformit (the fast and robust acoustic beamformer)," .