INTERSPEECH
2009 BRIGHTON

KL Realignment for Speaker Diarization with Multiple Feature Streams

Deepu Vijayasenan[1,2], *Fabio Valente*[1], *Hervé Bourlard*[1,2]

[1]Idiap Research Institute, CP 592, CH-1920 Martigny,Switzerland
[2] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
deepu.vijayasenan@idiap.ch,fabio.valente@idiap.ch, herve.bourlard@idiap.ch

10.21437/Interspeech.2009-325

Abstract

This paper aims at investigating the use of Kullback-Leibler (KL) divergence based realignment with application to speaker diarization. The use of KL divergence based realignment operates directly on the speaker posterior distribution estimates and is compared with traditional realignment performed using HMM/GMM system. We hypothesize that using posterior estimates to re-align speaker boundaries is more robust than gaussian mixture models in case of multiple feature streams with different statistical properties. Experiments are run on the NIST RT06 data. These experiments reveal that in case of conventional MFCC features the two approaches yields the same performance while the KL based system outperforms the HMM/GMM re-alignment in case of combination of multiple feature streams (MFCC and TDOA).

Index Terms: speaker diarization, information bottleneck, feature combination

1. Introduction

Speaker diarization systems address the problem of *"who spoke when"* in a given audio recording. This involves determining the number of speakers and identifying the speech corresponding to each speaker in an unsupervised manner. Conventional speaker diarization systems use short term spectral features like mel frequency cepstral coefficients (MFCC) and are based on ergodic HMMs [1, 2]. Each speaker is modeled with an HMM state with a minimum duration constraint. The state emission probabilities are modeled with Gaussian Mixture Models(GMM). The diarization algorithm follows an agglomerative clustering of initial speech segments followed by the realignment over the estimated speaker models. GMMs have been proved being very effective for clustering and re-alignment when a single feature stream is used.

Recently speaker diarization systems are converging towards combining multiple feature streams. Alternate features such as features obtained from long-time windows, Time Delay of Arrivals (TDOA) features (in case of MDM data) have been explored in the context of speaker diarization [3, 4]. Combination of such complementary features with the conventional MFCC features improves the diarization performance considerably [3, 4]. Conventional HMM/GMM systems construct different models for each feature stream. The feature combination is performed by a linear combination of the log likelihoods. However, different features possess very diverse statistical properties. This could lead to two different problems. On one side different features may need GMMs with different complexity (i.e. different number of gaussians). On the other hand GMMs may have totally different dynamic ranges of log likelihoods for each feature stream. For example, in [3] the number of Gaussian components in the initial model is fixed as five for MFCC features and one for TDOA features. In addition, variabilities across different recording conditions could influence the feature statistics. The dimension of TDOA features varies depending on the number of distant microphones. Using a global linear combination to combine log likelihoods may not be appropriate in such scenarios.

In our previous work [5], we partially addressed the problem using a non parametric approach to multiple-streams speaker diarization based on the Information Bottleneck principle. The clustering is based on a set of relevance variables which are represented as posteriors of a background GMM model. Whenever multiple features are used, the combination happens at the posterior distribution level rather than at the log-likelihood level.

In this paper, we propose a method to perform re-alignment using solely the posterior distribution values and investigate its application into multi stream diarization. The approach is based on the use of Kullback-Leibler divergence between distributions. The problem of minimizing the KL divergence between a reference posteriors and a learned set of models has been studied in the context of Automatic Speech Recognition (ASR) and can be solved by an EM algorithm [6]. In case of multi-stream diarization, a posterior based combination is employed, thus avoiding the problem of different feature dynamic ranges. The posterior space have the same dimension for all features, thus making the system more robust to variations in feature dimension and scale. In addition, the complexity of the realignment algorithm stays the same. In the present paper Section 2 reviews the Information Bottleneck(IB) principle and speaker diarization using agglomerative IB. Section 3 then describes the proposed algorithm for realignment. Experiments and results are presented in Section 4, and finally section 5 concludes the paper.

2. IB based Diarization

Let us consider a set of speech segments $X = \{x_1, \ldots, x_T\}$ obtained from uniform linear segmentation of the speech data in the audio recording. The speaker diarization task aims at clustering the elements of X that are uttered by the same speaker. In [7] we proposed an approach based on the Information Bottleneck principle inspired from rate distortion theory. In contrast to conventional minimum distortion based clustering techniques, it is based on preserving the relevant information specific to a given problem. The IB principle states that the best clustering is the one that compresses the input variables with minimum loss of mutual information with respect to set of relevance variables referred as Y. Relevance variables are variables that are considered important or carry the relevant information for a given clustering problem. We had proposed to use the gaussian components of a background GMM as relevance variable set Y[7].

This is motivated by the wide success of GMMs for speaker recognition. The clustering operates using probabilities $p(y|x)$ obtained in trivial way using Bayes' rule.

Thus, let us consider a set of input variables X (i.e. speech segments) to be clustered into clusters $C = \{c_i, \ldots, c_K\}$, and let Y denote the set of relevance variables which contain useful information about the problem (i.e. the components of a background GMM). The IB principle states that the best clustering representation C must preserve as much information about Y as possible i.e. the clustering representation should maximize the mutual information $I(Y, C)$ under a constraint of minimum mutual information $I(X, C)$ (See [8] for details). This corresponds to the maximization of:

$$\mathcal{F} = I(C, Y) - \frac{1}{\beta} I(X, C) \tag{1}$$

Where β is a Lagrange multiplier (the notation is consistent with [8]). This criterion should be optimized with respect to the stochastic mapping $p(c|x)$. This leads to a consistent system of equations which can be solved using iterative optimization techniques [8].

The optimization of the objective function (1) can be done in greedy fashion using the agglomerative Information Bottleneck method [8]. The algorithm is initialized with the trivial clustering of each point considered as a separate cluster ($|X|$ clusters). At each step of the algorithm a cluster merge is performed such that the information loss with respect to the relevance variables is minimum. The loss of mutual information at each step is given by a Jensen-Shannon divergence which is straightforward to compute from the posterior distribution $p(y|x)$. This method is described in detail in [9]. The information preserved $I(C, Y)$ monotonically decreases at each merge. The optimal number of clusters are selected based on a threshold on the Normalized Mutual Information (NMI) $\frac{I(C,Y)}{I(X,Y)}$. The complete algorithm is summarized as follows.

1 Acoustic feature extraction from the beamformed audio.

2 Speech/non-speech segmentation and rejection of non-speech frames.

3 Uniform segmentation of speech in chunks of fixed size D = 250ms i.e., set X.

4 Estimation of a Gaussian component with shared diagonal covariance matrix for each segment i.e., set Y.

5 Estimation of conditional distribution $p(y|x)$.

6 aIB clustering and model selection

7 Clustering refinement using Viterbi re-alignment.

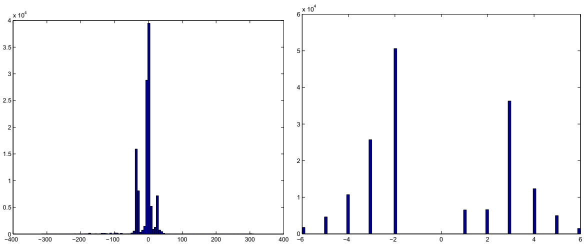
Further details of the algorithm can be found in [7] where it is shown that this approach yields state of the art results with a significant speecd up factor. The algorithm produces a partition of the data (i.e. a clustering) $p(C|X)$ as well as posterior distribution for each speaker (i.e. for each cluster) c i.e $p(Y|C)$. In the following we will discuss how to re-align speaker segmentation using directly the distribution $p(Y|C)$ without any GMM.

2.1. Multiple Features

Whenever multiple feature streams, $\{F_i\}$ are available the combination can directly happen in the space of the relevance variables i.e. using the posterior probabilities $p(y|x)$. For each feature stream F_i we estimate a background GMM M_{F_i}. The combined posterior distribution is then calculated as

$$p(y|x) = \sum_i p(y|x, M_{F_i}) P_F^i \tag{2}$$

where P_F^i is the prior probability corresponding to feature stream F_i. These combination scheme does not suffer from the different dimensionality or the different statistics of the features because it make use of posterior estimates $p(y|x)$ rather than log-likelihoods.



(a) CMU 20050914-0900 (b) EDI 20050216-1051

Figure 1: Histogram of TDOA features of RT06 eval meetings

3. Realignment

Speaker diarization systems make extensive use of Viterbi realignment. The realignment is supposed to improve the speaker boundaries obtained after the agglomerative clustering. HMM/GMM are used for this purpose [1, 2]. Generally multiple realignment, re-estimation iterations are performed. In case of multiple feature streams, a weighted combination of log likelihoods is used for the realignment [3].

However, the statistics of each feature stream are usually different. Consider for example the MFCC features and TDOA features. Figure 1 shows the histogram of the TDOA features of two meetings. It can be seen that the distribution is impulsive in case of TDOA feature stream, while MFCC features follow approximately Gaussian distribution. Also the dimension of the TDOA features can vary from meeting to meeting depending on the number of microphones used in recording. Figure 2 plots the negative log-likelihood obtained using a GMM for MFCC and TDOA features: while MFCC log-likelihood is approximately constant across meetings, TDOA log-likelihood change considerably according to the number of microphones thus the dimensions of delay features.

We investigate here a new realignment algorithm based on the posterior distribution values $p(y|x)$ (i.e. the posterior value of a guassian component given the feature vector x) aiming at being more robust against such variations in statistics. The algorithm is motivated by the IB principle and aims at using Viterbi realignment in posterior space as defined in Section 2. Let us start with the following proposition:

Proposition 1. *The IB maximization of Equation (1) is equivalent to the following minimization:*

$$\min[I(X, C) + \beta\, E(d(X, C))] \tag{3}$$

where $d(X, C) = KL(p(Y|X)||p(Y|C))$, is the KL divergence between the posterior distributions given by the cluster and the input (proof in [10]).

Consider a feature stream $(x_1, x_2, \ldots, x_T)$ partitioned into a set of clusters (speakers) $c_1, \ldots, c_K$ by the aIB algorithm. In case of hard clustering, $\beta \to \infty$ and the IB optimization of (3) reduces to the minimization of second term:

$$\begin{aligned} E(d(X, C)) &= E(KL(p(Y|X)||p(Y|C))) \\ &= \sum_t p(x_t) \sum_i p(c_i|x_t) KL\left(p(Y|x_t)||p(Y|c_i)\right) \end{aligned}$$

Given the cluster assignment $p(c_i|x_t) \in \{0, 1\}$ (hard clustering), and assuming the input clustering elements have uniform prior, the optimization turns out to be the minimization of:

$$\arg\min_{\mathbf{c}} \sum_t KL\left(p(Y|x_t)||p(Y|c_t)\right) \tag{4}$$

Where c_t is such that $p(c_t|x_t) = 1$.

Let us first consider the classical HMM/GMM realignment. The system has a set of speaker models (GMM). These GMM models are used as the state emission probabilities of an ergodic HMM. The optimal Viterbi path (speaker sequence) $\mathbf{c} = (c_1, c_2, ..., c_T)$ is determined as the best sequence of speakers that gives the maximum likelihood for the feature stream:

$$\mathbf{c}^{opt} = \arg\max_{\mathbf{c}} \sum_t \log(b_{c_t}(x_t)) + \log(a_{c_t c_{t+1}}) \tag{5}$$

Where c_t is the speaker at time index t, $b_{c_t}(.)$ is the emission probability distribution (GMM) corresponding to speaker c_t and $a_{c_i c_j}$ is the transition probability of transition from speaker c_i to speaker c_j. In case the speaker is represented with a single feature stream GMM, we have:

$$\log(b_{c_t}(x_t)) = \log \sum_r w^r_{ct} \mathcal{N}(x_t, \mu^r_{ct}, \Sigma^r_{ct}) \tag{6}$$

where $\mathcal{N}(.)$ is the Gaussian pdf; $w^r_{ct}, \mu^r_{ct}, \Sigma^r_{ct}$ are weights, means and covariance matrix corresponding to speaker c_t. In case of multiple feature streams GMM with features $x_t = \{x^1_t, x^2_t\}$, the log linear combination becomes:

$$\begin{aligned} \log(b_{c_t}(x_t)) = P^1_F \log \sum_{r1} w^{r1}_{ct} N(x_{t1}, \mu^{r1}_{ct}, \Sigma^{r1}_{ct}) \\ +(1 - P^1_F)\, log \sum_{r2} w^{r2}_{ct} N(x_{t2}, \mu^{r2}_{ct}, \Sigma^{r2}_{ct}) \end{aligned} \tag{7}$$

where P^1_F is the log-linear combination weight and means, variance and covariance matrices are to be considered relative to each feature stream. The weight is static across different meetings. However, note that likelihood values have large variations according to number of channels (Figure 2).

In a similar manner, we propose to extend the objective function from equation (4) as follows:

$$\mathbf{c}^{opt} = \arg\min_{\mathbf{c}} \sum_t KL\left(p(Y|x_t)||p(Y|c_t)\right) - \log(a_{c_t c_{t+1}}) \tag{8}$$

Thus the KL divergence between each feature vector and the posterior distribution of the speaker model is minimized. The problem of minimizing the KL divergence between a reference posteriors (in this case the $p(y|x)$) and the learned set of models ($p(y|c)$) can be solved by an EM algorithm [6]. The re-estimation formula for $p(y|c)$ is simply given by

$$p(y|c_i) = \sum_{x_t : x_t \in c_i} p(y|x_t) \tag{9}$$

i.e. the new speaker model is obtained by averaging posterior probabilities $p(y|x_t)$ for x_t that belongs to c_i. In both HMM/GMM and HMM/KL systems, a minimum duration constrain on the speaker states is imposed as in [1].

Whenever multiple feature streams are used the re-alignment can be performed using the combined posterior probabilities as defined in Equation (2). These features, being estimates of probability values, are normalized.

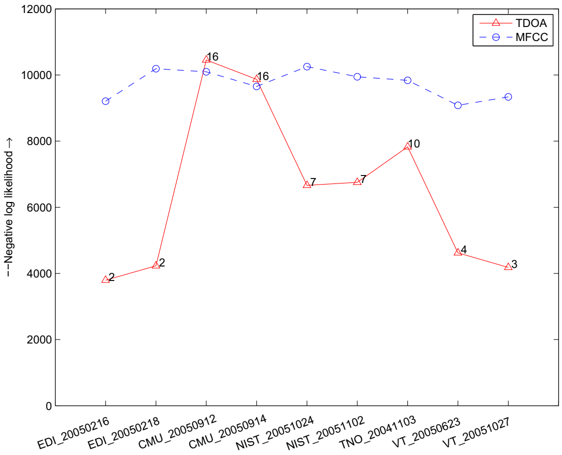


Figure 2: Variation of average negative log likelihoods of the background GMM across meetings for MFCC and TDOA features together with number of microphones. The negative log likelihoods of TDOA features is dependent on the number of microphones.

4. Experiments and Results

Although the proposed framework is general, we explore here the combination of TDOA features with conventional MFCC features. We perform the experiments on NIST RT06 evaluation data for "Meeting Recognition Diarization" task recorded via Multiple Distance Microphones(MDM). The data was preprocessed and beamformed with *BeamformIt* [11] toolkit. The bug fixed version of *BeamformIt 2.2* is used for this purpose which provides different features compared to those used in [5]. We verified an improvement with the new beamforming in the MFCC based system as compared to what reported in [5]. 19 MFCC features and TDOA features were extracted from the beamformed signal. TDOA feature dimension depends on the number of microphones used. The variation of average negative log likelihood of the background GMM for these features is illustrated in Figure 2. It can be seen that the statistics of the TDOA features vary considerably across meetings and depends on the number of microphones used. The log likelihood of MFCC features however, seems to be stable across different meetings.

Diarization systems are evaluated using Diarization Error (DER) as the measure. DER is the sum of speech/non-speech error and speaker errors. Speech/non-speech error consists of missed speech and false alarm errors. Speech/no-speech segmentation is obtained using a forced alignment of the reference transcripts using the AMI RT06 first pass ASR models [12]. Since the same speech/non-speech segmentation is used across all the experiments, only speaker error will be reported henceforth.

Experiments aims at comparing re-alignment performed using the HMM/GMM and the HMM/KL systems in case of single and multiple feature streams. The agglomerative clustering framework is described in details in [7] for the single feature stream and in [5] for the multiple feature streams case. In case of multiple feature streams, the weights are empirically determined from a development dataset. The MFCC weight is fixed to 0.9 in case of HMM/GMM system (log-likelihood combination) and to 0.7 in case of aIB clustering (posterior distribution

Table 1: Speaker error comparison of proposed system and baseline – Individual features and combination

	MFCC features			TDOA features			Feature Combination		
		Realignment			Realignment			Realignment	
Meeting	Without Realign	HMM/ GMM	KL based	Without Realign	HMM/ GMM	KL based	Without Realign	HMM/ GMM	KL based
CMU_20050912-0900	12.2	9.0	8.4	25.40	23.5	22.5	7.6	3.8	5.7
CMU_20050914-0900	15.5	11.6	11.4	24.60	21.5	21.9	4.8	3.0	3.1
EDI_20050216-1051	35.5	31.0	30.7	36.30	40.4	38.7	7.1	4.3	5.1
EDI_20050218-0900	26.8	23.2	24.3	30.00	29.4	31.1	18.6	16.2	15.7
NIST_20051024-0930	14.5	10.1	10.2	10.90	9.2	10.8	5.5	3.4	3.9
NIST_20051102-1323	14.4	10.1	10.3	11.30	8.2	8.7	2.5	1.2	1.6
TNO_20041103-1130	19.9	18.6	16.0	47.90	48.5	48.7	28.3	31.3	26.5
VT_20050623-1400	11.4	5.5	6.6	22.90	21.6	22.2	22.0	22.3	20.4
VT_20051027-1400	26.3	25.3	27.0	11.60	28.0	13.4	12.1	16.6	11.0
ALL	19.3	15.7	15.7	24.40	25.0	23.9	11.6	10.7	9.9

combination).

Table 1 provides the meeting-wise speaker error rate for the agglomerative clustering without realignment as well as with HMM/GMM and KL based realignments. The case of MFCC features, TDOA features and MFCC+TDOA features are considered. Let us consider separately all the different cases.

In case of MFCC features both HMM/GMM and KL based system have the same overall performance showing that in such a case there is no reason for preferring a scheme over the other. In case of TDOA features (where the number of features and their statistical properties change from meeting to meeting) the KL based system outperforms the HMM/GMM by 1.1% absolute. In case of combination of MFCC and TDOA the improvement of the KL based re-alignment is 0.8% absolute i.e. from 10.7% to 9.9%.

5. Conclusions

In this work we have proposed a KL divergence based realignment scheme that operates on the speaker posterior estimates. This extends our previous work on Information theoretic clustering. The system only depends on posterior probabilities of a set of relevance variables defined as the components of a background GMM model. When tested on single feature stream (e.g. MFCC coefficients), the proposed re-alignment produce the same performance as the conventional HMM/GMM re-alignment. On the other hand when the diarization uses multiple feature streams i.e. MFCC and TDOA features with different statistics and different dimensions, the KL divergence based re-alignment outperforms the HMM/GMM by 0.8% absolute reducing the speaker error from 10.7% to 9.9%.

Although in this study, we investigated the combination of MFCC and TDOA the proposed multiple stream diarization system is completely general and can be extended to other features (acoustic or visual) with very different statistical properties . Given that combination and re-alignment is performed with posterior distribution estimates, the proposed approach is supposed to be more robust than conventional HMM/GMM. Experiments with other feature sets are currently investigated and will be addressed in future works.

6. Acknowledgements

This work was supported by the European Union under the integrated projects AMIDA, Augmented Multi-party Interaction with Distance Access, contract number IST-033812, as well as KERSEQ project under the Indo Swiss Joint Research Program (ISJRP) financed by the Swiss National Science Foundation. This project is pursued in collaboration with EPFL under contract number IT02. The authors would like to thank Dr. Xavi Anguera and Dr.Chuck Wooters for their help with the HMM/GMM system and Beamforming as well as Dr. John Dines for his help with speech/no speech segmentation.

7. References

[1] J. Ajmera, “Robust audio segmentation,” Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne (EPFL), 2004.

[2] X. Anguera, “Robust speaker diarization for meetings,” Ph.D. dissertation, Universitat Politecnica de Catalunya, 2006.

[3] J.M. Pardo , X. Anguera, C. Wooters, “Speaker Diarization For Multiple-Distant-Microphone Meetings Using Several Sources of Information,” *IEEE Transactions on Computers*, vol. 56, no. 9, p. 1189, 2007.

[4] O. Vinyals, G. Friedland, “Modulation spectrogram features for speaker diarization,” in *Proceedings of Interspeech*, 2008.

[5] D. Vijayasenan, F. Valente, and H. Bourlard, “Integration of tdoa features in information bottleneck framework for fast speaker diarization,” in *Interspeech 2008*, 2008.

[6] G. Aradilla, “Acoustic models for posterior features in speech recognition,” Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, Lausanne , Switzerland, 2008.

[7] D. Vijayasenan, F. Valente, and H. Bourlard, “Agglomerative information bottleneck for speaker diarization of meetings data,” in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2007, pp. 250–255.

[8] N. Tishby, F. Pereira, and W. Bialek, “The information bottleneck method,” in *NEC Research Institute TR*, 1998.

[9] N. Slonim, N. Friedman, and N. Tishby, “Agglomerative information bottleneck,” in *Proceedings of Advances in Neural Information Processing Systems*. MIT Press, 1999, pp. 617–623.

[10] P. Harremoës and N. Tishby, “The Information bottleneck revisited or how to choose a good distortion measure,” in *IEEE International Symposium on Information Theory, 2007. ISIT 2007*, 2007, pp. 566–570.

[11] X. Anguera, “Beamformit, the fast and robust acoustic beamformer,” in *http://www.icsi.berkeley.edu/x̃anguera/BeamformIt*, 2006.

[12] Hain T. et. al., “The AMI meeting transcription system: Progress and performance,” in *Proceedings of NIST RT’O6 Workshop*, 2006.