IDIAP RESEARCH REPORT

AN INFORMATION THEORETIC COMBINATION OF MFCC AND TDOA FEATURES FOR SPEAKER DIARIZATION

Deepu Vijayasenan Fabio Valente Hervé Bourlard

Idiap-RR-22-2010

JULY 2010

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

‘

An Information Theoretic Combination of MFCC and TDOA Features for Speaker Diarization

Deepu Vijayasenan, *Student Member, IEEE,* Fabio Valente, *Member, IEEE,*
and Hervé Bourlard, *Fellow, IEEE*

Abstract

This correspondence describes a novel system for speaker diarization of meetings recordings based on the combination of acoustic features (MFCC) and Time Delay of Arrivals (TDOA). The first part of the paper analyzes differences between MFCC and TDOA features which possess completely different statistical properties. When Gaussian Mixture Models are used, experiments reveal that the diarization system is sensitive to the different recording scenarios (i.e. meeting rooms with varying number of microphones). In the second part, a new multistream diarization system is proposed extending previous work on Information Theoretic diarization. Both speaker clustering and speaker realignment steps are discussed; in contrary to current systems, the proposed method avoids to perform the feature combination averaging log-likelihood scores. Experiments on meetings data reveal that the proposed approach outperforms the GMM based system when the recording is done with varying number of microphones.

Index Terms

Speaker Diarization, Information Bottleneck, Feature Combination, Meeting data

I. Introduction

Speaker diarization is the task of determining "who spoke when" in an audio stream. It is an unsupervised learning paradigm, where the system learns the number of speakers as well as identifies the speech segments corresponding to each speaker.

Conventional speaker diarization systems use an ergodic Hidden Markov Model (HMM) where each speaker is represented as an HMM state with a minimum duration [1]. The state emission probabilities are modeled with Gaussian Mixture Models. The diarization follows multiple steps of agglomerative clustering and realignment. The system is initialized with an over determined number of speakers by means of uniform segmentation or by speaker change detection methods. At each iteration the two most similar clusters (according to some distance measure) are merged. After that, the time boundaries of segments are realigned using a Viterbi algorithm. This merging/realigning proceeds iteratively until a stopping criterion is met. Common measures used as both distance measure as well as the stopping criterion are the Bayesian Information Criterion [2] and modified versions [1], [3], [4], [5]. To determine

the distance measure between two speaker clusters, those methods first estimate a single speaker model from the data belonging to both clusters. The distance measure then depends on the ratio of individual cluster likelihoods to the likelihood of the single model estimated with data from both clusters.

Typical acoustic features consist of short term spectral features such as Mel frequency cepstral coefficients (MFCC). In the meeting scenario, data recordings are commonly carried out in a non-intrusive way with multiple distant microphones. The spatial redundancy of the different signals can be used for speaker diarization. For instance, whenever the geometry of the microphone array is known, the speaker locations can be estimated and used as complementary features to conventional MFCC [6]. Otherwise if the array geometry is unknown, the estimated time difference of arrival (TDOA) between different channels of a microphone array can be used as features. Experiments have shown that as stand alone features [7], TDOA performs poorly respect to MFCC but significant performance improvements are obtained when TDOA are used in combination with MFCC [8], [9].

MFCC and TDOA are modeled separately with different GMMs and they are combined by linearly weighting the individual log-likelihoods [8]. The log-likelihood combination is used to calculate the BIC distance measure and to refine the speaker boundaries using the Viterbi realignment. The weights of the linear combination are estimated from an independent development data set. This approach has been proven very effective in several evaluations and is implemented in large number of diarization systems [10], [11], [12].

Speaker diarization is applied to recordings performed with varying the number of microphones across different meeting rooms (from 2 microphones to 16 for conference room meetings [13], [14] and up to 64 microphones in case of lecture recordings). If the recording is done with an array of C microphones, the number of TDOA features is equal to $C-1$. As a consequence, the dimension of the TDOA feature vector will vary according to the number of microphones resulting in different ranges of log-likelihoods.

This paper investigates the combination of MFCC and TDOA features for speaker diarization in case of recordings acquired with arrays composed of different number of microphones. In section II this combination is studied in case of HMM/GMM systems. The investigation studies the sensitivity of the system with respect to the feature weights.

Sections III and IV introduce a novel multistream diarization system that extends our previous related work on Information Bottleneck (IB) based clustering [15]. The IB diarization [15] performs the diarization in a space of relevance variables and realigns speaker boundaries with an HMM/GMM system. This paper introduces two contributions to the original system:

- The clustering in the space of relevance variables is extended to handle multiple feature streams (Section IV). In contrary to the HMM/GMM, it avoids the combination of log-likelihoods.
- The HMM/GMM realignment is replaced with a Kullback-Leibler based realignment as it arises from the IB principle (Section V). The realignment scheme operates on the same relevance variable space and again avoids the combination of log-likelihoods.

The rationale behind performing clustering and realignment using the IB framework rather than log-likelihood combination is that the system should gain robustness to the statistics of the different features (MFCC and TDOA). The proposed approaches are validated in experiments using a dataset with a number of microphones between 2

and 16.

II. Baseline System

Let us consider a diarization system based on the HMM/GMM framework. Each speaker is modeled with an HMM state with minimum duration of three seconds. The system is initialized with an over determined number of speakers by means of uniform segmentation or by speaker change detection methods. Multiple iterations of clustering and realignment are then performed.

The clustering follows an agglomerative framework in which at each step two most similar clusters are merged. The similarity between two clusters is based on a modified BIC criterion [3]. The clustering stops when no BIC value is greater than zero i.e. when none of the possible merges between cluster pairs satisfy the criterion.

The emission probability distribution b_{c_i} corresponding to speaker cluster c_i is modeled as a GMM:

$$\log b_{c_i}(s_t) = \log \sum_r w_{c_i}^r \mathcal{N}(s_t, \mu_{c_i}^r, \Sigma_{c_i}^r) \tag{1}$$

where s_t is the input feature, $\mathcal{N}(.)$ is the Gaussian pdf and $w_{c_i}^r$, $\mu_{c_i}^r$, $\Sigma_{c_i}^r$ are the weights, means and covariance matrices corresponding to r^{th} mixture Gaussian of cluster c_i.

Each cluster merge is followed by a Viterbi re-alignment that smooths the speaker boundaries and improves the diarization performance. The entire meeting is then realigned with the estimated speaker models after the merge. The optimal path (speaker sequence) $\mathbf{c} = (c_1, \ldots, c_T)$ is determined as the best sequence of states that maximizes the data likelihood. This can be represented as the following optimization:

$$\mathbf{c}^{opt} = \arg\min_c \sum_t [-\log b_{c_t}(s_t) - \log(a_{c_t c_{t+1}})] \tag{2}$$

where c_t is the speaker cluster at time t. The term $a_{c_i c_j}$ represents the transition probability from speaker state c_i to c_j. The transition probabilities incorporate the minimum duration constraint.

A. Multiple Feature Streams

Whenever multiple feature streams are available, the HMM/GMM system uses a linear combination of log likelihoods. This approach models the two feature streams with separate GMMs and the combination is then performed by linearly weighting their log-likelihoods [8]. GMMs are estimated separately with observations assigned to the same speaker cluster.

The weights are estimated minimizing the diarization error on a independent development data set.

Let s_t^{mfcc} and s_t^{tdoa} represent the feature values at time t. GMM models $b_{c_i}^{mfcc}(.)$ and $b_{c_i}^{tdoa}(.)$ are estimated separately from MFCC and TDOA features assigned to the same cluster. A linear combination of the log likelihoods is computed as:

$$P_{mfcc} \log b_{c_i}^{mfcc}(s_t^{mfcc}) + P_{tdoa} \log b_{c_i}^{tdoa}(s_t^{tdoa}) \tag{3}$$

P_{mfcc} and P_{tdoa} denote the weights corresponding to MFCC and TDOA features respectively such that $P_{mfcc} + P_{tdoa} = 1$. The diarization system then performs both agglomerative clustering and Viterbi realignment using the combination of the log-likelihoods as in Equation (3).

TABLE I

LIST OF MEETING USED FOR EVALUATION IN THE PAPER WITH ASSOCIATED NUMBER OF MICROPHONES

sl.no.	meeting id	#microphones
1	CMU_20050912-0900	2
2	CMU_20050914-0900	2
3	EDI_20050216-1051	16
4	EDI_20050218-0900	16
5	NIST_20051024-0930	7
6	NIST_20051102-1323	7
7	TNO_20041103-1130	9
8	VT_20050623-1400	4
9	VT_20051027-1400	3

B. Experiments and Baseline results

In this section the impact of variations in feature statistics in the baseline system is investigated. A dataset of nine meetings recorded using Multiple Distance Microphones (MDM) across five different meeting rooms is used. The set of meetings with associated number of microphones is listed in Table I.

A delay and sum beamforming [16] is performed on the MDM data to obtain a single enhanced channel. The beamforming is performed with the *BeamformIt* [17] toolkit. A Bug-fixed version of *BeamformIt 2.2* is used for this purpose. The beamforming first selects a reference channel based on maximum average cross correlation with other channels. Then the Time Delay of Arrival (TDOA) of each channel with respect to the reference channel is computed. TDOA features are computed using a window of $500ms$ shifted every $10ms$ for each of the individual signals. The delay value that yields the maximum correlation is estimated with a generalized cross correlation phase transform (GCC-PHAT). Hence the number of delay features is always one less than the number of microphones. Following the TDOA estimation, a weighted delay and sum combination of all channels results in a single enhanced channel. 19 MFCC coefficients are estimated from this enhanced output using a $30ms$ window shifted every $10ms$. Both MFCC and TDOA values have the same frame rate.

In order to consider the different statistical properties of the features, MFCC are initially modeled with a five component GMM while TDOA are modeled with a single Gaussian [8]. Fig.1 plots the average negative log likelihood values of two independent GMMs trained on MFCC and TDOA features for the 9 meetings used in this work. It can be seen that their dynamic ranges are quite different. TDOA likelihoods depend on the feature vector dimensions and thus on the number of microphones. Larger feature dimension leads to larger likelihood values. For example meeting 3 and 4 have largest feature dimension (16) among the meetings and posses highest negative log likelihood values. Furthermore TDOA and MFCC statistics are considerably different.

Also there is a two order magnitude difference between the minimum and the maximum values of TDOA log likelihoods across different meetings in Fig.1. Possible reasons of such variations include variable dimension of

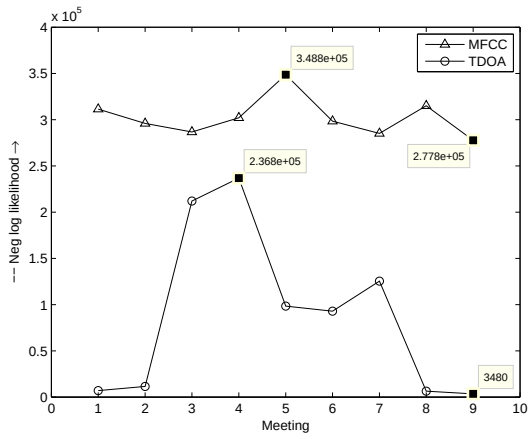

Fig. 1. Average negative log likelihood values of MFCC and TDOA features at the beginning of clustering for a set of nine meetings. The markers denote maximum and minimum values of feature streams.

features, differences in the recording environments etc.

Let us now consider the effect of this in a diarization system. Speaker diarization systems are evaluated using the Diarization Error Rate (DER) measure. DER is the sum of speech/non-speech error and speaker mismatch error. The speech/non-speech error contains missed speech and false alarm errors. Since we use same speech non-speech segmentation (same speech/non-speech error) across all experiments we report speaker error for the purpose of comparison.

The baseline HMM/GMM system is initialized with 16 clusters obtained with uniform linear segmentation and the clustering is performed using modified BIC as the distance measure [11]. At first, the performance of the system is analyzed with a set of optimal weights for each meeting. The optimal weights represent the set of weights that produce the smallest speaker error in each meeting. To determine this, MFCC feature weight P_{mfcc} is varied from zero to one for each meeting while keeping the sum of feature weights as unity. i.e., P_{tdoa} is fixed as $1 - P_{mfcc}$. The weights corresponding to the minimum speaker error are chosen as the optimal weights for that meeting. The corresponding speaker error is compared against the performance of the feature weights estimated from development data ($P_{tdoa} = 0.1$ and $P_{mfcc} = 0.9$ as reported by [8]). The final speaker error as well as the performance just before the last realignment step are reported.

Table II reports the speaker error for the meeting-wise optimal weights, as well as for the estimated weight from development dataset. It can be observed that there is a performance reduction of 6.6% absolute in the latter case, the actual speaker error being almost double of the speaker error with optimal weights.

The individual meeting performances are depicted in Fig.2 and the corresponding optimal weights for the TDOA feature stream are illustrated in Fig.3. It can be seen that:

1 the magnitudes of the weights span a considerably large range (note that the plot is in logarithmic scale). This could happen due to the difference in statistical properties of individual feature streams as discussed before.

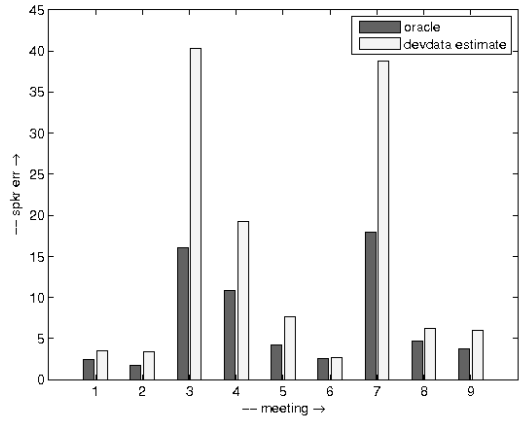

Fig. 2. Meeting-wise speaker error for MFCC+TDOA feature combination of HMM/GMM system: optimal weights for each meeting and the estimate from separate development data ($P_{tdoa} = 0.1$). $P_{mfcc} = 1 - P_{tdoa}$

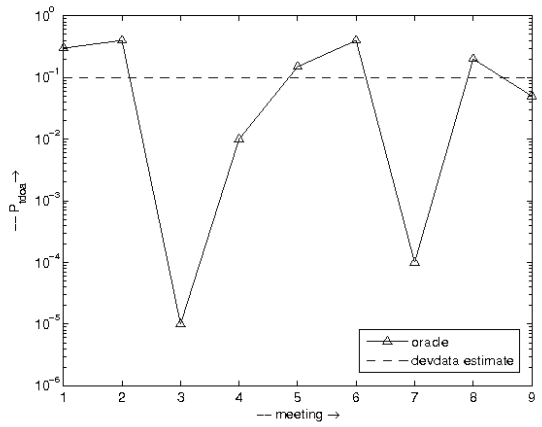

Fig. 3. Variation of optimal weight in the HMM/GMM system for TDOA feature across different meetings $P_{mfcc} = 1 - P_{tdoa}$

TABLE II

OVERALL SPEAKER ERROR FOR MFCC+TDOA COMBINATION OF THE HMM/GMM SYSTEM: OPTIMAL WEIGHTS FOR EACH MEETING, AND THE ESTIMATED WEIGHT FROM SEPARATE DEVELOPMENT DATA ($P_{tdoa} = 0.1$). $P_{mfcc} = 1 - P_{tdoa}$. THE TABLE ALSO REPORTS THE SPEAKER ERROR WITH AND WITHOUT RE-ALIGNMENT AFTER THE LAST CLUSTERING STEP.

	optimal wt.	$P_{tdoa} = 0.1$
no realign	7.9	14.8
with realign	7.0	13.6

2 whenever the estimated weights are notably different from the optimal values (meetings 3,4 and 7), the drop from the optimum performances is large.

In the following an alternative diarization system is investigated for the purpose of reducing the effect of the different statistics between type of features.

III. SINGLE STREAM IB SPEAKER DIARIZATION

In [15], we have proposed a speaker diarization system based on Information Bottleneck (IB) principle. The system is based on clustering and realignment steps. The clustering depends on distribution of a set of relevance variables. The realignment is based on conventional HMM/GMM systems. For completeness of the paper and definition of notation, the IB principle and details of the diarization system are briefly presented. All details can be found in [15].

A. Information Bottleneck Principle

The Information Bottleneck (IB) is a distributional clustering framework based on information theoretic principles [18][19]. The IB principle depends on availability of a set of *relevance variables* Y that carry important information with respect to the problem. For example, the set of words are used as relevance variables for the document retrieval. The method tries to form a clustering representation that preserves maximum mutual information with respect to the relevance variables. IB clustering thus attempts to preserve meaningful information with respect to a given problem.

Consider a set of input elements $X = \{x_1, \ldots, x_T\}$ to be clustered into a set of clusters $C = \{c_1, \ldots, c_K\}$. Let Y be the set of relevance variables that carry useful information pertaining to the problem. According to IB principle the best clustering representation C should be a compact representation of input variables X (minimize mutual information between X and C) and should preserve as much information as possible about the relevance variables Y (maximize mutual information between C and Y). This corresponds to the maximization of:

$$\mathcal{F} = I(Y, C) - \frac{1}{\beta} I(C, X) \tag{4}$$

where β (Notation consistent with [19]) is a Lagrange multiplier. The IB objective function in Equation 4 is optimized with respect to the stochastic mapping $p(C|X)$ that maps each element from input element X to the new cluster representation C. The clustering depends only on the conditional distribution of the relevance variables with respect to the input features $p(y|x)$. Different methods to construct the solution of IB objective function include agglomerative and sequential information bottleneck, iterative optimization, deterministic annealing etc. (See [18] for a detailed review). Here in this work we focus only on agglomerative information Bottleneck that is briefly discussed below.

B. Agglomerative IB

Agglomerative Information Bottleneck (aIB) is a greedy approach towards optimization of the IB objective function [20]. The algorithm is initialized with $|X|$ clusters, i.e., each input element of the set X is considered as a separate cluster. At each step, the algorithm merges two clusters that results in minimum loss in the IB objective function (4). This process is continued until the required number of clusters is reached. It can be shown that this loss could be represented as sum of two Jensen-Shannon divergences. This distance measure arises as the result of maximization of the IB functional and depends only on the distribution $p(y|x)$ (refer [20] for details).

C. IB based Speaker Diarization

Let us define now the input elements X and the relevance variables Y that represent the meaningful information about the diarization problem.

A set of speech segments obtained by uniform linear segmentation of input features is used as the input variables (Set X).

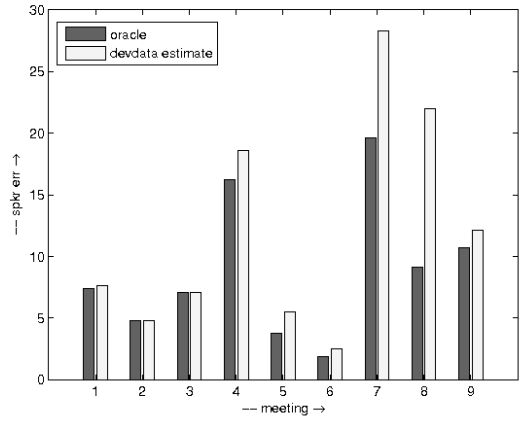

Fig. 4. Meeting-wise speaker error for MFCC+TDOA of IB based feature combination: optimal weights for each meeting and the estimate from separate development data ($P_{tdoa} = 0.3$). $P_{mfcc} = 1 - P_{tdoa}$

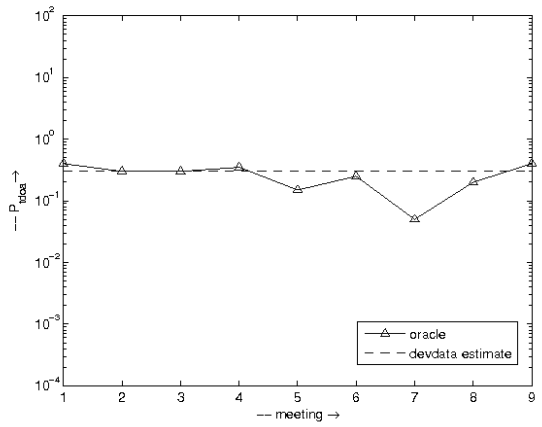

Fig. 5. Variation of optimal weight in the IB system for TDOA features across different meetings $P_{mfcc} = 1 - P_{tdoa}$

Motivated by the success of GMMs on speaker recognition/verification applications, the set of relevance variables Y is defined as the components of a background GMM estimated from the entire input audio recording. Let us consider a GMM trained on the entire recording $f(s) = \sum_j w_j \mathcal{N}(s, \mu_j, \Sigma_j)$. The conditional probability of Gaussian component j (relevance variable) with respect to input feature s_t can be calculated from Bayes' rule:

$$p(y_j|s_t) = \frac{w_j \mathcal{N}(s_t, \mu_j, \Sigma_j)}{\sum_r w_r \mathcal{N}(s_t, \mu_r, \Sigma_r)} \tag{5}$$

For each segment x_i that consists of a set of frames, the distributions are averaged across all frames to determine the relevance variable distribution.

The clustering of speech segments (i.e. variables X) is performed following an agglomerative framework according to their distance in the space of relevance variables Y. The clustering stops based on a Normalized Mutual Information (NMI) criterion defined as:

$$NMI = \frac{I(Y, C)}{I(Y, X)} \tag{6}$$

NMI denotes the fraction of original mutual information $I(Y, X)$ preserved by the clustering representation. This quantity decreases monotonically with cluster merge, and a threshold is used to select the optimal number of clusters. The threshold is determined using a development dataset composed of ten meetings used in previous NIST evaluations. The meetings were around 10 minutes long (all the details can be found in [15]). The complete algorithm is 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 2.5s 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 Agglomerative clustering until the stopping criterion is met.

D. Realignment

The initial uniform segmentation of the audio file into segments X is then refined in order to smooth the speaker boundaries. The realignment is performed using a conventional HMM/GMM system. A separate GMM is estimated for each of the speakers estimated using the IB clustering. The segment boundaries are then re-estimated using a Viterbi realignment with a minimum state duration HMM(as in the baseline system). Multiple iterations of GMM estimation and realignment are performed. Details of the method can be found in [15].

This single feature stream diarization is now extended in order to handle multiple features by adding two contributions:

- the relevance variable distribution are estimated from different feature streams (e.g. MFCC and TDOA; section IV)
- the re-alignment is performed directly in the space of relevance variables Y based on Kullback-Leibler divergence (section V)

IV. Extension to Multiple Features

Let us now consider the case in which MFCC and TDOA features from the same meeting are available. The proposed method can be extended using separate aligned background GMMs for MFCC and TDOA. The background models have the same number of components proportional to the length of the meeting as described in [15], i.e., the number of components is equal to $N = S/250$ where S is the total number of speech frames.

Initially a GMM model is estimated using MFCC features s_t^{mfcc}. Each observation s_t^{mfcc} is then assigned to one of the GMM components. The parameters of the TDOA GMM are estimated using the same mapping between the feature time indices and the GMM components. In other words, suppose the r^{th} component parameters of MFCC GMM were estimated from a set of MFCC features $\{s_{t'}^{mfcc}\}$. The r^{th} component parameters of the TDOA GMM will then be estimated from the set of TDOA features $\{s_{t'}^{tdoa}\}$ that have the same time indices $\{t'\}$.

While in the baseline system [8] MFCC and TDOA GMM for each cluster are estimated separately with observations (MFCC and TDOA) assigned to the same cluster, in the proposed system separate background models are estimated with observations assigned to the same components (from the MFCC background model). Thus the two GMMs have the same number of components and have a strict one-to-one mapping between the components.

The set of these corresponding aligned mixture components represent the relevance variables. The relevance variable distributions $p(y|s_t^{mfcc})$ and $p(y|s_t^{tdoa})$ are estimated as before using Bayes' rule. The estimation of $p(y|s_t^{mfcc}, s_t^{tdoa})$ is obtained as a weighted average of individual distributions as:

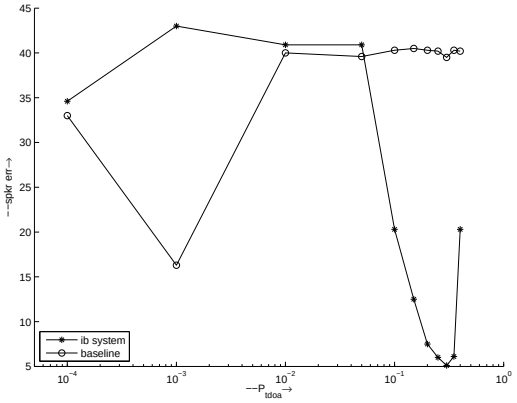

Fig. 6. Speaker error as a function of P_{tdoa} ($P_{mfcc} = 1 - P_{tdoa}$) for a meeting with estimated weight farthest from optimal. The selected weight from development data tuning is $P_{tdoa} = 0.3$ for the IB system and $P_{tdoa} = 0.1$ for the baseline

$$p(y|s_t^{mfcc}, s_t^{tdoa}) = p(y|s_t^{mfcc})P_{mfcc} + p(y|s_t^{tdoa})P_{tdoa} \tag{7}$$

where P_{mfcc} and P_{tdoa} represent the weights such that $P_{mfcc} + P_{tdoa} = 1$. In contrary to GMM log-likelihood combination , here the individual distributions $p(y|s_t^{mfcc})$ and $p(y|s_t^{tdoa})$ are normalized and have the same dynamic range regardless of the dimension of the feature vector. Thus the linear combination does not suffer from dimensionality/statistics problems as in the case of GMM log-likelihoods.

A. Experiments and Results

In order to investigate the effectiveness of the proposed approach, experiments are conducted to study the combination of MFCC and TDOA features on the same set of meetings as described in Section II. The effect of the realignment algorithm will be discussed in the next section.

The feature weights are determined using the same development data set described in [15]. The estimated values of weights are $(P_{mfcc}, P_{tdoa}) = (0.7, 0.3)$. It is interesting to notice that those values are different from those obtained when the tuning is done using log-likelihood combination i.e., $(0.9, 0.1)$.

The evaluation is done on the meeting recordings described in Table I. As before we report the performance of the systems with optimal and estimated weights. The optimal weights correspond to lowest speaker error for each meeting. Table III presents the corresponding speaker error.

Performance with the estimated weights data is only 2.9% worse compared to the optimal weights. The speaker error with aIB clustering is 2% absolute better than the baseline result even before performing the realignment step.

The meeting-wise speaker errors in Fig.4 show that the system performance with the estimated weights are close to the best performance determined by the optimal weights except in case of two meetings 7 and 8.

The optimal weights for each meeting in case of IB system are represented in Fig.5 and as opposed to the HMM/GMM system (Fig.3), they span a smaller range. Fig.6 depicts the variation of speaker error with the variation of P_{tdoa} for the meeting with highest difference between optimal and estimated weight values. The IB system performance with estimated weights is closer to the performance with optimal weights as compared to the baseline system.

In summary, whenever the combination happens at the level of the relevance variables instead of log-likelihoods, the diarization error is less sensitive to the dimension of the TDOA features.

B. Automatic Weighting of Feature Streams

In [4], an automatic weighting schemes based on entropy is proposed to determine the combination weights without development data. Originally the method was proposed in context of ASR in [21] and was later applied to speaker diarization [4]. This section investigates inverse entropy combination for both the baseline system and the proposed system.

Let $c_1, \ldots, c_L$ be the L clusters at any stage of the agglomerative procedure with associated GMM speaker models (Emission probability distribution of HMM states) $b_{c_i}(.)$. A discrete distribution of speakers is first estimated for each feature vector s_t as:

$$P(c_i|s_t) = \frac{b_{c_i}(s_t)}{\sum_{j=1}^{L} b_{c_j}(s_t)}, i = 1 \ldots L \tag{8}$$

The entropy of the speaker distribution is then computed :

$$H_t = -\sum_{i=1}^{L} P(c_i|s_t) \log P(c_i|s_t) \tag{9}$$

The algorithm estimates entropy values for MFCC and TDOA feature streams (H_t^{mfcc} and H_t^{tdoa}) using the respective GMM speaker models. The log-likelihood combination as in Eqn(3) is then performed with weights proportional to inverse entropy values. The MFCC and TDOA weights are given by:

$$P_{mfcc} = \frac{1/H_t^{mfcc}}{1/H_t^{mfcc}+1/H_t^{tdoa}} \tag{10}$$

$$P_{tdoa} = \frac{1/H_t^{tdoa}}{1/H_t^{mfcc}+1/H_t^{tdoa}} \tag{11}$$

The combination operates at frame level i.e. weights are calculated for each frame. In case of IB system the distribution of relevance variables are used instead of speaker distributions.

TABLE III

OVERALL SPEAKER ERROR FOR MFCC+TDOA COMBINATION OF THE IB SYSTEM: OPTIMAL WEIGHTS FOR EACH MEETING AND THE ESTIMATED WEIGHT FROM SEPARATE DEVELOPMENT DATA ($P_{tdoa} = 0.3$). $P_{mfcc} = 1 - P_{tdoa}$

Optimal weight	$P_{tdoa} = 0.3$
8.7	11.6

TABLE IV

OVERALL SPEAKER ERROR FOR MFCC+TDOA COMBINATION: AUTOMATICALLY ESTIMATED WEIGHTS

baseline	IB system
20.8	17.5

Table IV reports the results. Both schemes perform approximatively $6-7\%$ absolute worse compared to the baseline. The proposed system still outperforms the baseline. These results are consistent with findings reported in [4](Section 6.5), i.e., inverse entropy weighting does not outperforms the fixed weighting based on development data. In fact, this scheme assigns higher weights to the TDOA feature stream while the optimal weights is higher for MFCC features.

V. KL BASED REALIGNMENT

The second contribution of this paper consists of introducing a speaker realignment that operates directly in the space of the relevance variables estimated by Equation (5). The rationale behind this is that performing realignment in the space of normalized distributions $p(Y|S)$ would increase the robustness of the system as compared to the log-likelihood domain.

Now let us rewrite the IB objective function according to the following proposition:

Proposition 1: The IB maximization of Equation (4) is equivalent to the following minimization:

$$\min[I(X,C) + \beta\, E(d(X,C))] \tag{12}$$

$$d(X,C) = KL(p(Y|X)||p(Y|C)) \tag{13}$$

where $d(X,C)$, is the KL divergence between distributions given by the cluster and the input [22]. (See Appendix A for a proof)

The re-alignment is performed after the agglomerative clustering to smooth the initial arbitrary boundaries obtained by uniform segmentation. The aIB clustering described in section IV provides an initial partition of features $(s_1,\ldots,s_T)$ (input variables of realignment) into a set of speakers $\{c_1,\ldots,c_K\}$. This corresponds to an hard clustering partition where $p(c|s) \in \{0,1\}$. Hard clustering is obtained by taking the limit $\beta \to \infty$ in the IB optimization criterion (4). Thus the IB criterion reduces to the maximization of $I(C,Y)$ alone [18]. From the above proposition, this is equivalent to minimizing $d(S,C)$. Developing the expression for $d(S,C)$, it is possible to write:

$$\begin{aligned} E(d(S,C)) &= E[KL(p(Y|S)||p(Y|C))] \\ &= \sum_t p(s_t) \sum_i p(c_i|s_t) KL(p(Y|s_t)||p(Y|c_i)) \\ &= \sum_t p(s_t) KL(p(Y|s_t)||p(Y|c_t)) \end{aligned} \tag{14}$$

Where c_t is such that $p(c_t|s_t) = 1$, for other values of C, $p(c_i|s_t) = 0$. Assuming equal priors for s_t, minimization of $E(d(S, C))$ is equivalent to:

$$\min E(d(S, C)) = \min \sum_t KL(p(Y|s_t)||p(Y|c_t)) \tag{15}$$

The term $p(Y|c_t)$ denotes the distribution of relevance variables for each speaker. This can be seen as the "speaker model" estimated using $p(Y|s_t)$. While the GMM realignment selects the speaker that maximizes the log-likelihood sum, the proposed approach selects the speaker that minimize the KL divergence between $p(Y|s_t)$ and $p(Y|c_t)$. The problem of minimizing the KL divergence between a feature stream represented as distributions and a set of learned models has been explored previously in the context of automatic speech recognition [23]. The estimation formula for "speaker models" $p(y|c)$ is given by:

$$p(y|c_i) = \frac{1}{p(c_i)} \sum_{s_t : s_t \in c_i} p(y|s_t)p(s_t) \tag{16}$$

In case of equal priors $p(s_t)$, the estimation formula becomes the arithmetic mean of the distributions $p(y|s_t)$. Thus the speaker model for a cluster c_t is the average of distributions $p(y|s_t)$ assigned to it.

Then, the objective function can be extended to include the minimum duration constraint as in the baseline system:

$$\mathbf{c}^{opt} = \arg\min_c \sum_t [KL(p(Y|s_t)||p(Y|c_t)) - \log(a_{c_t c_{t+1}})] \tag{17}$$

A parallel can be seen between Equations (2) and (17) reported below:

$$\mathbf{c}^{opt} = \arg\min_c \sum_t [-\log b_{c_t}(s_t) - \log(a_{c_t c_{t+1}})]$$

The term $p(Y|c_t)$ represents the speaker model in the relevant variable space and during the Viterbi. The negative log-likelihood ($-\log b_{c_t}(s_t)$) is replaced by the KL divergence $KL(p(Y|s_t)||p(Y|c_t))$ which serves as the distance measure between the speaker model and the input features $p(Y|s_t)$. The realignment depends only on the distribution $p(y|s_t)$ which is normalized. When MFCC and TDOA feature streams are used, this distribution is computed as $p(y|s_t^{mfcc}, s_t^{tdoa}) = p(y|s_t^{mfcc})P_{mfcc} + p(y|s_t^{tdoa})P_{tdoa}$. Performing KL based realignment using $p(y|s_t^{mfcc}, s_t^{tdoa})$ eliminates the combination of log-likelihood scores.

A. Experiments and Results

This section compares the KL based realignment with the HMM/GMM based realignment. Both systems use a minimum duration constraint equal to 2.5 second i.e. 250 frames.

The comparison is done on the same setup of section II-B after aIB clustering. Table V compares the speaker error in case of optimal and estimated weights. Fig. 7 illustrates the meeting-wise speaker error before and after realignment.

Both realignment schemes reduce the overall speaker error in case of optimal weights as well as in case of weights estimated from development data. However the KL realignment outperforms the HMM/GMM realignment

TABLE V

OVERALL SPEAKER ERROR FOR MFCC+TDOA COMBINATION OF THE IB SYSTEM WITH KL REALIGNMENT: USING OPTIMAL WEIGHTS FOR EACH MEETING, AND ESTIMATE FROM DEVELOPMENT DATA ($P_{tdoa} = 0.3$). $P_{mfcc} = 1 - P_{tdoa}$

Realignment	Optimal wt.	$P_{tdoa} = 0.3$
HMM/GMM	7.9	10.7
HMM/KL	7.0	9.9

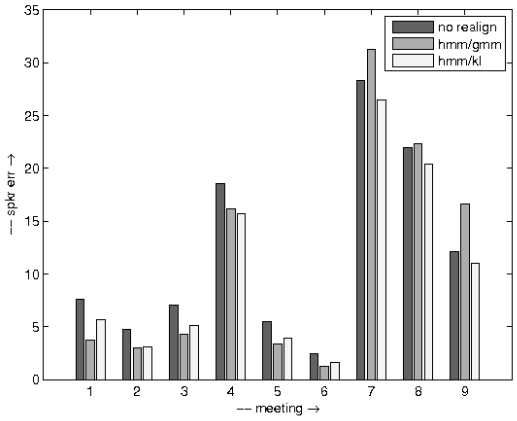

Fig. 7. Speaker error with and without realignment for feature combination with estimated weights

by close to 1% absolute. Fig. 7 shows that the HMM/GMM system improves the diarization output in six out of nine meetings whereas the KL realignment is improving consistently across all meetings of the data set.

All results of the paper are summarized in Table VI. It can be seen that the optimal weights yield the same performance for the baseline system (Row 2) and the IB system with KL realignment (Row 5). However, with the estimated weights, performance of the baseline system degrades considerably (6.6% abosolute worse). The IB system is more robust to weights, and the performance of the system is closer (2.9% absolute worse) to performance of the optimal weights.

VI. SUMMARY AND CONCLUSIONS

This paper discusses the combination of MFCC and TDOA features for speaker diarization introducing two new contributions that extends previous work on information theoretic diarization [15]:

- *Combination Scheme*: State-of-the-art multiple stream diarization uses a linear combination of GMM log-likelihoods trained on MFCC and TDOA features. TDOA features have different statistics compared to MFCC. Furthermore their dimensionality varies according to the number of channels used for recordings. Setting linear

TABLE VI

COMPARISON OF SPEAKER ERRORS OF THE IB SYSTEM WITH THE BASELINE. THE BASELINE RESULTS ARE REPORTED WITH AND WITHOUT THE LAST REALIGNMENT STEP. THE TABLE PRESENTS RESULTS THAT CORRESPONDS TO OPTIMAL WEIGHTS AS WELL AS WEIGHTS ESTIMATED FROM DEVELOPMENT DATA.

system		weights	
		optimal wts.	devdata tuning
baseline	no realign	7.9	14.8
	realigned	7.0	13.6
IB	no realign	8.7	11.6
	HMM/GMM	7.9	10.7
	KL based	7.0	**9.9**

combination weights according to log-likelihoods present robustness problems across different meeting rooms. A combination scheme performed in a normalized space of relevance variables is proposed and investigated.

- *KL based Realignment*: Instead of re-aligning boundaries with an HMM/GMM system, a KL based realignment scheme is proposed. This method uses only the frame level relevance variable distributions.

The experiments are performed on a dataset with number of TDOA features of variable dimension from 2 to 16. Both optimal weights as well as weights estimated from tuning on a development data set are investigated. The proposed combination performs 2% absolute better compared to the baseline even before realignment. Both realignments (HMM/GMM and KL) reduce the speaker error, the KL outperforming the HMM/GMM by 1% absolute.

The performance of the overall system (IB clustering + KL realignment) is 4% absolute (28% relative) better than the baseline system. It is important to notice that the two systems hold the same optimal performance meaning that when meeting-wise optimal weights are selected, the speaker error is similar. On the other hand, whenever weights are fixed, the IB system is more robust to variations across data. The individual weights variations is much larger when the combination happens at the log-likelihood level.

Although the feature combination of only two features (MFCC and TDOA) is investigated in this work, the algorithms proposed are general and could be extended to other features (acoustic or visual). The framework only uses the distribution $p(y|s_t)$ that is normalized and is hence more robust to features with diverse statistics compared to the conventional HMM/GMM system. Experiments with more than two feature sets would be addressed in future works.

APPENDIX A
PROOF OF PROPOSITION 1

Proof: Consider $I(X,Y) - I(C,Y)$

$$\begin{aligned}
&= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} - \sum_{y,c} p(y,c) \log \frac{p(y,c)}{p(y)p(c)} \\
&= \sum_{x,y,c} p(x,y,c) \log \frac{p(x,y)p(c)}{p(y,c)p(x)} \\
&= \sum_{x,y,c} p(y|x,c)p(c|x)p(x) \log \frac{p(x,y)p(c)}{p(y,c)p(x)} \\
&= \sum_{x,y,c} p(y|x)p(c|x)p(x) \log \frac{p(y|x)}{p(y|c)} \\
&= \sum_{x} p(x) \sum_{c} p(c|x) \sum_{y} p(y|x) \log \frac{p(y|x)}{p(y|c)} \\
&= \sum_{x} p(x) \sum_{c} p(c|x) KL\left(p(Y|x)||p(Y|c)\right) \\
&= \sum_{x,c} p(x,c) KL\left(p(Y|x)||p(Y|c)\right) \qquad (18)
\end{aligned}$$

Consider the IB criterion in Equation (4); i.e, the maximization of $I(C,Y) - \frac{1}{\beta} I(X,C)$. This can be rewritten as a minimization in the following form: $\min[I(X,C) - \beta I(C,Y)]$ which is equivalent to $\min[I(X,C) + \beta.\left(I(X,Y) - I(C,Y)\right)]$ since $I(X,Y)$ is a constant for the minimization. Using the result of Equation 18 the optimization becomes $\min\left[I(X,C) + \beta \sum_{x,c} p(x,c) KL\left(p(Y|x)||p(Y|c)\right)\right]$ ■

ACKNOWLEDGMENTS

This work is supported by the Swiss National Science Foundation (under the NCCR on Interactive Multimodal Information Management and the MULTI grants), and by the European Community's Seventh Framework Programme (FP7/ 2007 – 2013), under grant agreement no. 231287 (SSPNet).

The Authors thank all the reviewers for their valuable comments and suggestions.

REFERENCES

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

[2] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings of DARPA speech recognition workshop*, 1998, pp. 127–138.

[3] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *Signal Processing Letters, IEEE*, vol. 11, no. 8, pp. 649–651, 2004.

[4] X. Anguera, "Robust speaker diarization for meetings," Ph.D. dissertation, Universitat Politecnica de Catalunya, 2006.

[5] D. van Leeuwen and M. Huijbregts, "The AMI speaker diarization system for NIST RT06s meeting data," *Lecture Notes in Computer Science*, vol. 4299, p. 371, 2006.

[6] G. Lathoud and I. McCowan, "Location based speaker segmentation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003.

[7] J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multi-microphone meetings using only between-channel differences," in *MLMI*, 2006.

[8] ——, "Speaker diarization for multi-microphone meetings: Mixing acoustic features and inter-channel time differences," in *International Conference on Speech and Language Processing*, 2006.

[9] 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.

[10] X. Anguera, C. Wooters, B. Peskin, and M. Aguiló, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," *Lecture Notes in Computer Science*, vol. 3869, p. 402, 2006.

[11] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," *Lecture Notes in Computer Science*, vol. 4625, pp. 509–519, 2008.

[12] D. van Leeuwen and M. Konecnỳ, "Progress in the AMIDA speaker diarization system for meeting data," *Multimodal Technologies for Perception of Humans: International Evaluation Workshops Clear 2007 and Rt 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, p. 475, 2008.

[13] J. Fiscus, J. Ajot, M. Michel, and J. Garofolo, "The Rich Transcription 2006 Spring Meeting Recognition Evaluation," *Lecture Notes in Computer Science*, vol. 4299, p. 309, 2006.

[14] J. Fiscus, J. Ajot, and J. Garofolo, "The rich transcription 2007 meeting recognition evaluation," *Multimodal Technologies for Perception of Humans, Lecture Notes in Computer Science, Berlin*, 2008.

[15] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 7, pp. 1382 – 1393, 2009.

[16] X. Anguera, C. Wooters, and J. H. Hernando, "Speaker diarization for multi-party meetings using acoustic fusion," in *Proceedings of Automatic Speech Recognition and Understanding*, 2006, pp. 426–431.

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

[18] N. Slonim, "The information bottleneck: Theory and applications," Ph.D. dissertation, The Hebrew University of Jerusalem, 2002.

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

[20] 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.

[21] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *Proceedings ICASSP*, vol. 3, 2003, pp. 1–5.

[22] 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.

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