Integration of TDOA Features in Information Bottleneck Framework for Fast Speaker Diarization

Deepu Vijayasenan [a] [b] Fabio Valente [a]
Hervé Bourlard [a] [b]

IDIAP–RR 08-26

June 2008

submitted for publication

[a] IDIAP Research Institute, Martigny, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

IDIAP Research Institute **www.idiap.ch**
Av. des Prés−Beudin 20 P.O. Box 592 1920 Martigny − Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 08-26

Integration of TDOA Features in Information Bottleneck Framework for Fast Speaker Diarization

Deepu Vijayasenan Fabio Valente Hervé Bourlard

June 2008

submitted for publication

Abstract. In this paper we address the combination of multiple feature streams in a fast speaker diarization system for meeting recordings. Whenever Multiple Distant Microphones (MDM) are used, it is possible to estimate the Time Delay of Arrival (TDOA) for different channels. In [9], it is shown that TDOA can be used as additional features together with conventional spectral features for improving speaker diarization. We investigate here the combination of TDOA and spectral features in a fast diarization system based on the Information Bottleneck principle. We evaluate the algorithm on the NIST RT06 diarization task. Adding TDOA features to spectral features reduces the speaker error by 7% absolute. Results are comparable to those of conventional HMM/GMM based systems with consistent reduction in computational complexity.

1 Introduction

Speaker diarization determines "*who spoke when*" in a given audio recording. This involves finding the number of speakers and the identification of speech segments of each speaker in an unsupervised manner.

In meeting case scenario, data acquisition is done in a non-intrusive manner using a microphone array often referred as Multiple Distant Microphones (MDM). Several speaker diarization systems for meetings have been developed and recently computational complexity issues have been addressed (e.g. see [18],[15]). Fast speaker diarization with low computational complexity is meant to be performed on a common machine eventually while the meeting itself is taking place.

Conventional speaker diarization systems are based on parametric models like ergodic HMM in which each state is associated with a speaker. The emission probabilities are modeled with Gaussian Mixture Models (GMM). The clustering is performed in agglomerative fashion until a stopping criterion (generally the Bayesian Information Criterion [6] or a modified version [3]) is met. BIC is obtained penalizing the ratio of likelihoods of the individual clusters to the likelihood of the merged cluster thus at each step, the algorithm has to estimate likelihood of individual clusters and of all possible merging. This can be a computational demanding task and assume that enough data are available for estimating a parametric model of each merge.

In our previous work [15], we had proposed a non-parametric clustering approach based on Information Bottleneck (IB) framework for speaker diarization. The clustering is performed based on the distance in a space of relevance variables. The approach is non-parametric and does not require any model re-estimations. This system achieves similar diarization error rates to HMM/GMM systems with comparatively less computational requirements [15].

In this paper we extend the work on fast diarization describing the integration of a set of features that can significantly improve the result: the Time Delay of Arrival (TDOA) of different channels. Whenever the signal acquisition is done using a microphone array, there is a spatio-temporal redundancy in the data that can be used in the diarization. The most common solution consists in obtaining a single audio stream out of the multiple channels using beamforming techniques. If the geometry of the microphone array is known, the location of speakers can be determined and used as complementary information [10]. In the case of unknown microphone array geometry, the estimated difference of arrival of the signal between the different channels (i.e. the TDOA) can be used as features for conventional diarization systems [11], [10]. In [11] it is shown that as stand alone features they performs poorly but used together with spectral features (e.g. MFCC) [9] they can provide large improvements.

When those features are integrated into a HMM/GMM models, the two different streams (MFCC, TDOA) are treated independently. Separate GMM models for both features are estimated. The combined log-likelihood is obtained as weighted sum of the individual log-likelihoods [9]. Given that the HMM/GMM system is parametric, new features will lead to larger number of parameters.

In this work, we investigate an extension of the non-parametric clustering based on IB for including TDOA. This clustering is based on a set of relevance variables with dimension independent on the number of features thus the complexity of the clustering itself does not increase.

The rest of the paper is organized as follows: section 2 reviews the Information Bottleneck framework for speaker diarization. Section 3 describes the full speaker diarization algorithm. TDOA features and combination of TDOA features with MFCC features are described in section 4. Section 5 reports experimental evaluations and section 6 concludes the paper.

2 Information Bottleneck Principle

Let X, be a set of elements to cluster into a set of C clusters. Let Y be a set of variables of interest associated with X such that $\forall x \epsilon X$ and $\forall y \epsilon Y$ the conditional distribution $p(y|x)$ is available. Clusters C can be interpreted as a compression (bottleneck) of initial data set X in which information that X contains about Y is passed through the bottleneck C. The Information Bottleneck (IB) principle states that the clustering C should preserve as much information as possible from the original data set X w.r.t. relevance variables Y.

IB method [14] is inspired from Rate-Distortion theory and states the best representation C of data X minimizes the mutual information $I(X, C)$, i.e. the distortion and preserves as much information as possible about Y (maximizing $I(C, Y)$). Thus the IB objective function can be formulated as minimization of the Lagrangian,

$$I(X, C) - \beta I(C, Y) \tag{1}$$

where β is the trade-off between the amount of information $I(C, Y)$ to be preserved and the compression of the initial representation $I(C, X)$. Function (1) must be optimized w.r.t. the stochastic mapping $p(C|X)$. Expressions for $I(X, C)$ and $I(C, Y)$ can be developed as:

$$I(X, C) = \sum_{x \epsilon X, c \epsilon C} p(x)p(c|x)log\frac{p(c|x)}{p(c)} \tag{2}$$

$$I(Y, C) = \sum_{y \epsilon Y, c \epsilon C} p(c)p(y|c)log\frac{p(y|c)}{p(y)} \tag{3}$$

This leads to a set of self-consistent equations as shown in [12] which can be solved to obtain the cluster representation.

The limit $\beta \rightarrow \infty$ induces a hard partition of the input space i.e. the probabilistic map $p(c|x)$, takes values of 0 and 1 only. This is equivalent to minimizing only the information loss in the clustering i.e. $I(Y, C)$. This is performed by one of the following methods.

2.1 Agglomerative Information Bottleneck

The agglomerative Information Bottleneck (aIB) is a greedy approach to minimize the objective function of equation (1). The initialization consists of the trivial clustering of $|X|$ clusters i.e. each data point is treated as separate cluster. Subsequently the clusters are merged such that after each step the loss of mutual information w.r.t the relevant variables Y is minimum.

The loss of mutual information δI_y obtained by merging x_i and x_j is given by Jensen-Shannon divergence between $p(Y|x_i)$ and $p(Y|x_j)$ (see [12]). In case of discrete probabilities, this divergence is straightforward to compute. The information preserved in each step decreases monotonically. Details about implementation of aIB algorithm can be found in [12] and will not be further discussed here. The optimal number of clusters is selected by thresholding the Normalized Mutual Information $\frac{I(C,Y)}{I(X,Y)}$. Details of this method are described in [15]. However, aIB does not necessarily converge to the global minimum of the objective function due to its greedy nature.

2.2 Sequential Information Bottleneck

Sequential Information Bottleneck [13] targets to find the global maximum of the objective function. The algorithm is initialized with an initial partition of fixed number of clusters W, $\{c_1, ..., c_W\}$. The sIB draws some element x from the initial partition to make a new singleton cluster. x is then merged into a cluster c_{new} such that $c_{new} = argmin_{c \epsilon C}\, d(x, c)$ where $d(., .)$ is the Jensen-Shannon distance between x and c. It can be verified that if $c_{new} \neq c_{old}$ then $F(C_{new}) < F(C_{old})$ i.e., at each step either the objective function (1) improves or stays unchanged. This step is repeated several times until there is no change in the clustering assignment. To avoid local maxima, the procedure can be repeated with several random initializations. However, sIB works only with fixed number of clusters. For this reason in [16], we proposed to first obtain a partition using agglomerative clustering and then refine the partition using the sequential optimization. We refer to this method as aIB+sIB and it will be used in the experiment section.

2.3 IB for Speaker Diarization

For applying Information Bottleneck in speaker diarization we should define the input variables X to be clustered and the relevance variables Y. The input variables $X = \{x_i\}$ are defined as uniformly segmented speech segments from the meeting. In order to compute the relevance variables, a shared covariance matrix GMM is estimated from the data in the audio file. The relevant variables $Y = \{y_j\}$ are chosen as the components of the GMM. The posterior values of each component gives the conditional distribution $p(y_j|x_i)$ which can be estimated in a straightforward manner using the Bayes' rule. In other words each speech segment is projected in the posterior space obtained using a GMM trained on all the audio file (for details see [15]).

3 Speaker Diarization Algorithm

We summarize here the speaker diarization algorithm described in detail in [16]. The steps are the following:

1 Beamforming of the MDM data to obtain a single audio stream.

2 Acoustic feature extraction from the beamformed audio file.

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

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

5 Estimation of GMM model with shared diagonal covariance matrix for each segment i.e., definition of set Y.

6 Estimation of conditional probability $p(y|x)$.

7 Clustering and model selection using one of the methods described in sections 2.1 and 2.2.

8 Viterbi realignment using conventional HMM/GMM system estimated from previous segmentation.

This clustering relies on the purity of initial segments X which are arbitrarily obtained by uniform segmentation. If the length of the segment D is small enough segments may be considered as generated by a single speaker. Although this hypothesis can be true in case of Broadcast News audio data, in case of conversational speech with fast speaker change rate and overlapping speech (like in meetings data), initial segments may contain speech from several speakers. Thus Viterbi re-alignment is performed in order to refine the segment boundaries.

4 Delay Features

In case of meeting data recorded with Multiple Distance Microphones (MDM), the relative time delay of arrival (TDOA) between the different microphones can be estimated. Assuming the speakers are not changing position, those features can be used in speaker diarization [11]. It has been shown that TDOA improves the speaker diarization significantly in combination with conventional spectral features [9].

TDOA features are estimated using the generalized cross correlation phase transform (GCC-PHAT) [4]. All time delays are calculated with respect to a reference channel. This channel is chosen based on the signal to noise ratio or depending on the average cross correlation of the channel with other channels. After choosing a reference channel, signal in each channel is windowed using a $500ms$ window. Given two windowed signals $s_i(n)$ and $s_j(n)$ the GCC-PHAT is defined as

$$G_{PHAT}(f) = \frac{X_i(f)X_j^*(f)}{|X_i(f)||X_j(f)|} \tag{4}$$

where $X_i(f)$ and $X_j(f)$ are the Fourier transforms of the two signals. The TDOA for channels s_i and s_j is estimated as

$$d_{PHAT}(i,j) = \arg\max_{d} R_{PHAT}(d) \tag{5}$$

where $R_{PHAT}(d)$ is the inverse Fourier transform of $G_{PHAT}(f)$.

4.1 Combination of TDOA Spectral features

In this section we describe how to combine information coming from two different feature streams. Given a set of MFCC features and a set of TDOA features, two separate GMM models M_{MFCC} and M_{TDOA} are trained on data from the entire meeting recording. Posteriors distributions $p(y|x, M_{MFCC})$ and $p(y|x, M_{TDOA})$ are separately estimated from the two GMMs and then combined together as follows:

$$p(y|x) = p(y|x, M_{MFCC})P(M_{MFCC}) + p(y|x, M_{TDOA})P(M_{TDOA}) \tag{6}$$

where $P(M_{MFCC})$ and $P(M_{TDOA})$ are prior probabilities assigned to the individual features. Those weights are empirically estimated from a developmentdatabase. The speaker diarization algorithm is the same as described in section 3 but with a set of posteriors estimated from two different feature streams. In summary, the cardinality of X and Y does not change and the clustering keep the same complexity as with a single feature stream. The combination of different feature streams is done at the conditional distribution level as in equation 6. The only extra cost comes from the estimation of the two GMMs for the two feature sets.

5 Experiments and Results

In this section we describe experiments in order to compare IB system versus conventional HMM/GMM in terms of performances and in terms of computational complexity. We performed all the experiments on the NIST RT06 evaluation data for "Meeting Recognition Diarization" task based on data from Multiple Distant Microphones (MDM) [2] and results are provided in terms of Diarization Error Rates (DER). DER is the sum of missed speech error, false alarm speech error and speaker error (for details on DER see [1]). Speech/non-speech (spnsp) is the sum of missed speech and false alarm speech. System parameters are tuned on the development data. We used the NIST RT05 evaluation data as the development data.

Pre-processing of the data consists of a Wiener filter denoising for individual channels followed by a beam-forming algorithm (delay and sum) as described in [4],[5]. This was performed using the *BeamformIt* toolkit [17]. 19 MFCC features are then extracted from the beam-formed signal as well as the TDOA features.

Speech/non-speech segmentation is obtained using a forced alignment of the reference transcripts on close talking microphone data using the AMI RT06s first pass ASR models [8]. Results are scored against manual references force aligned by an ASR system. The same speech/non-speech segmentation is used across all experiments.

The baseline system is based on agglomerative clustering using HMM/GMM framework [9]. It uses a modified version of the BIC criterion in which the model complexity is kept constant while merging to avoid fine tuning the BIC penalty term [3]. Feature combination is done weighting the sum of log-likelihoods obtained by MFCC and TDOA models . Results of the baseline system are presented in table 1 with and without TDOA features.

Table 1: Results of the HMM/GMM baseline system in terms of missed speech, false alarm, speech/non-speech and speaker error.

Feature	Miss	FA	spnsp	spkr err	DER
MFCC	6.5	0.1	6.6	17.0	23.6
MFCC + Delay	6.5	0.1	6.6	9.3	15.9

In case of the IB framework, we estimate two posterior streams – one from the spectral features and the other from the delay features. The posteriors are combined as discussed in section 4.1. The effect of varying the weights $P(M_{MFCC})$ and $P(M_{TDOA})$ is shown in Figure 1. The best parameter was obtained based on tuning on the development data. The optimal value is found to be $P(M_{MFCC}) = 0.7$ and $P(M_{TDOA}) = 0.3$ (those values are different from those reported in [9] because this algorithm is combining probabilities and not log-likelihoods). We experimented with both agglomerative (aIB) and sequential clustering (aIB+sIB) as described in sections 2.1 and 2.2. Normalized Mutual Information was used to infer the number of speakers. Table 2 provide the results for IB based speaker diarization with and without TDOA features. For completeness results are reported also with and without Viterbi realignment. Since we also use the same speech/non speech reference as the baseline only speaker error rates are reported. The use of MFCC+TDOA features reduces the speaker error by $\sim 7\%$ absolute w.r.t. only MFCC. The combination of agglomerative and sequential clustering (aIB+sIB) provides the best results. The use of Viterbi realignment further reduces the error.

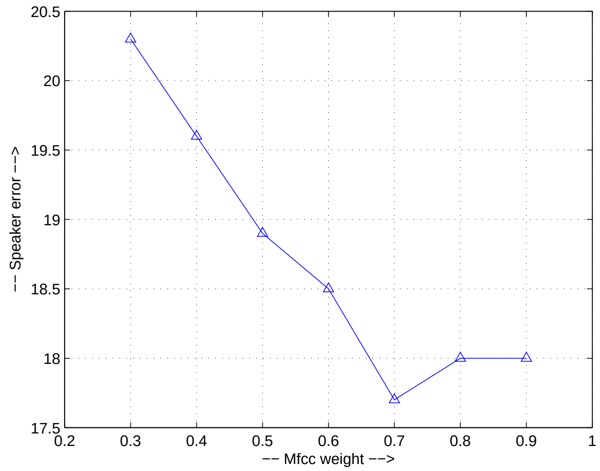

Figure 1: DER function of the mfcc weight

Table 2: Speaker error(%) for MFCC and MFCC+TDOA system for aIB and aIB+sIB clustering.

	aIB		aIB+sIB	
Feature	No Vit	With Vit	No Vit	With Vit
MFCC	22.1	17.1	18.3	16.6
MFCC+TDOA	12.0	11.4	10.3	9.7

Figure 2 illustrates speaker error for each meeting for the MFCC and for the (MFCC+TDOA) system.

The lowest DER is obtained with MFCC+TDOA features using aIB+sIB followed by Viterbi realignment. It can be seen that the proposed algorithm is able to achieve similar results as compared to state-of-the art results (13.9% speaker error versus 12.8% of the HMM/GMM). Moreover, the proposed algorithm has very low complexity as there is no speaker model estimations. The only additional computation for incorporating TDOA features is the estimation of a GMM model and conditional probabilities $p(y|x)$. The complexity of the clustering remains the same, since the dimensionality of the relevance variable Y and input variable X does not change. Table 3 shows the real time factors for the baseline and the proposed algorithms together with the speaker error. Real time factors are computed only on the clustering part i.e. they do not include feature extraction (MFCC,TDOA)

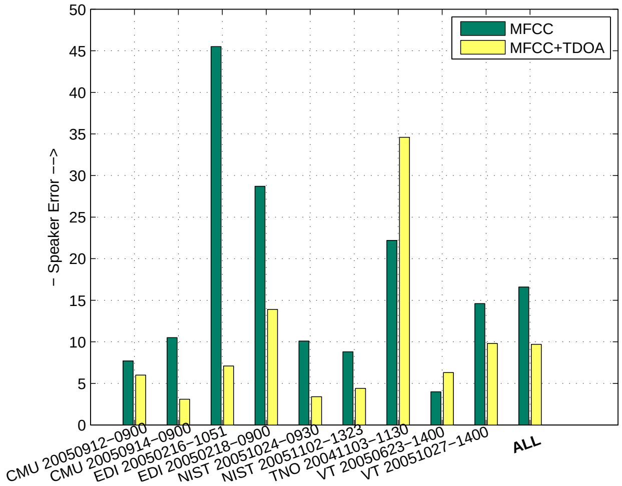

Figure 2: Speaker Error summary for all meetings: Baseline system, aIB+sIB system

and speech/non-speech detection. All experiments are benchmarked on a desktop machine with AMD Athlon 2.4GHz 64 X2 Dual Core Processor and 2GB RAM. The IB clustering runs almost half real time. HMM/GMM system is comparatively much slower (5.58xRT).

Table 3: Real time factor and speaker error for the aIB,aIB+sIB and HMM/GMM baseline system with MFCC+TDOA features.

	aIB	aIB+sIB	Baseline
Spk error	11.4%	9.7%	9.3%
RT factor	0.34xRT	0.41xRT	3.63xRT

6 Conclusions

In this work, we describe the integration of TDOA into a fast speaker diarization system for meeting recordings. The goal is to reduce the diarization error without significantly increasing the overall processing time.

The use of MFCC+TDOA reduce the DER by 7% absolute compared to only MFCC. The IB algorithm is 0.4% inferior to the state-of-the art system (9.7% versus 9.3% in the baseline). However, the system require much less resources compared to the state of the art system (0.63XRT versus 5.58xRT in the baseline). The system is almost twice as fast as real time on a normal desktop system.

Furthermore this framework can be used to integrate several other feature sets that have been shown to be useful in diarization [7] with very limited extra computational complexity.

7 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 gratefully thank the EU and Switzerland for their financial support, and all project partners for a fruitful collaboration.

Authors would like to thank Dr. Chuck Wooters and Dr. Xavier Anguera for their help with baseline system and beamforming toolkit. Authors also would like to thank Dr. John Dines for his help with the speech/non-speech segmentation

References

[1] http://nist.gov/speech/tests/rt/rt2004/fall/.

[2] http://www.nist.gov/speech/tests/rt/rt2006/spring/.

[3] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *IEEE Automatic Speech Recognition Understanding Workshop*, pages 411–416, 2003.

[4] 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.

[5] Xavier Anguera. *Robust Speaker Diarization for Meetings*. PhD thesis, Universitat Politecnica de Catalunya, 2006.

[6] S.S Chen and P.S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proceedings of DARPA speech recognition workshop*, 1998.

[7] V. et al. Gupta. Multiple feature combination to improve speaker diarization of telephone conversations. In *Proceedings of IEEE ASRU*, 2007.

[8] Hain T. et. al. The ami meeting transcription system: Progress and performance. In *Proceedings of NIST RT'O6 Workshop*, 2006.

[9] J.M. Pardo , X. Anguera, C. Wooters. Speaker diarization for multi-microphone meetings: Mixing acoustic features and inter-channel time differences. In *International Conference on Speech and Language Processing*, 2006.

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

[11] J.M. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multi-microphone meetings using only between-channel differences. In *Proceedings of Machine Learning for Multimodal Interaction Workshop, Washington DC, USA*, 2006.

[12] N. Slonim, N. Friedman, and N. Tishby. Agglomerative information bottleneck. In *Proceedings of Advances in Neural Information Processing Systems*, pages 617–623. MIT Press, 1999.

[13] Friedman F. Slonim N. and Tishby N. Unsupervised document classification using sequential information maximization. In *Proceeding of SIGIR'02, 25th ACM international Conference on Research and Development of Information Retireval*, 2002.

[14] N. Tishby, F.C. Pereira, and W. Bialek. The information bottleneck method. In *NEC Research Institute TR*, 1998.

[15] 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)*, pages 250–255, 2007.

[16] D. Vijayasenan, F. Valente, and H. Bourlard. Combination of agglomerative and sequential clustering for speaker diarization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

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

[18] Y. Huang, O. Vinyals, G. Friedland, C. Mller, N. Mirghafori, C. Wooters. A fast-match approach for robust, faster than real-time speaker diarization. In *Proceedings of IEEE ASRU*, pages 693–698, 2007.