IDIAP RESEARCH REPORT

An Information Theoretic Approach to Speaker Diarization of Meeting Data

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

IDIAP–RR 07-58

July 2008

submitted for publication

[a] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Report 07-58

An Information Theoretic Approach to Speaker Diarization of Meeting Data

Deepu Vijayasenan, Fabio Valente, Hervé Bourlard

July 2008

submitted for publication

Abstract. A speaker diarization system based on an information theoretic framework is described. The problem is formulated according to the *Information Bottleneck* (IB) principle. Unlike other approaches where the distance between speaker segments is arbitrarily introduced, IB method seeks the partition that maximizes the mutual information between observations and variables relevant for the problem while minimizing the distortion between observations. This solves the problem of choosing the distance between speech segments, which becomes the Jensen-Shannon divergence as it arises from the IB objective function optimization. We discuss issues related to speaker diarization using this information theoretic framework such as the criteria for inferring the number of speakers, the trade-off between quality and compression achieved by the diarization system, and the algorithms for optimizing the objective function. Furthermore we benchmark the proposed system against a state-of-the-art system on the NIST RT06 (Rich Transcription) data set for speaker diarization of meeting. The IB based system achieves a Diarization Error Rate of 23.2% as compared to 23.6% of the baseline system. This approach being mainly based on non-parametric clustering, it runs significantly faster then the baseline HMM/GMM based system, resulting in faster-then-real-time diarization.

1 Introduction

Speaker Diarization is the task of deciding *who spoke when* in an audio stream and is an essential step for several applications such as speaker adaptation in Large Vocabulary Automatic Speech Recognition (LVCSR) systems, speaker based indexing and retrieval. This task involves determining the number of speakers and identifying the speech segments associated with each speaker.

The number of speakers is not a priori known and must be estimated from data in an unsupervised manner. The most common approach to speaker diarization stays the one proposed in [10] which consists of agglomerative bottom-up clustering of acoustic segments. Speech segments are clustered together according to some similarity measure until a stopping criterion is met. Given that the final number of clusters is unknown and must be estimated from data, the stopping criterion is generally related to the complexity of the estimated model. The use of *Bayesian Information Criterion* [20] as model complexity metric has been proposed in [10] and is currently used in several state-of-the-art diarization systems.

Agglomerative clustering is based on similarity measures between segments. Several similarity measures have been considered in the literature based on BIC [10], modified versions of BIC [4, 3], Generalized Log-Likelihood Ratio [8], Kullback-Leibler divergence [9] or cross-likelihood distance [19]. The choice of this distance measure is somehow arbitrary.

In this paper we investigate the use of a clustering technique motivated from a information theoretic framework known as the *Information Bottleneck* (IB) [25]. The IB method has been applied to clustering of different types of data like documents [22, 24] and images [12]. IB clustering [23, 25] is a distributional clustering inspired from Rate-Distortion theory [11]. In contrary to many other clustering techniques, it is based on preserving the relevant information specific to a given problem instead of arbitrarily assuming a distance function in between given elements. Furthermore, given a data set to be clustered, IB tries to find the trade-off between the most compact representation and the most informative representation of the data. The first contribution of this paper is the investigation of IB based clustering for speaker diarization and its comparison with state-of-the-art systems based on Hidden Markov Models/ Gaussian Mixture Models (HMM/GMM) framework. We discuss differences and similarities of the two approaches and benchmark them in a speaker diarization task for meeting recordings.

Speaker diarization has been applied to several types of data e.g. broadcast news recordings, conversational telephone speech recordings and meetings recordings. The most recent efforts in the NIST Rich Transcription campaigns focus on meetings data acquired in several rooms with different acoustic properties and with a variable number of speakers. The audio data is recorded in a non-intrusive manner using Multiple Distant Microphones (MDM) or microphone array. Given the variety of acoustic environment, the conversational nature of the recordings and the use of distant microphones, those recordings represent a very challenging data set. Progresses along year in the diarization task for meetings data can be found in [2] and in [7].

Recently, attention has shifted on faster-then-real-time diarization systems with low computational complexity (see e.g. [30],[26, 27],[16]). In fact in the meeting case scenario, faster then realtime diarization would allow the use of several applications (meetings browsing, meeting summarization, speaker retrieval) on common desktop machine while the meeting is taking place.

Conventional systems model the audio stream using a fully connected HMM in which each state corresponds to a speaker with emission probabilities represented by GMM probability density functions [4],[15]. Merging two segments means estimating a new GMM model that represent data coming from both segments as well as the similarity measure in between the new GMM and the remaining speakers. This procedure can be computationally demanding.

This paper also investigates the IB clustering for a fast speaker diarization system. IB is a non-parametric framework that does not use any explicit speaker model. Thus, the algorithm does not need to estimate a GMM for each speaker, resulting in a considerably reduced computational complexity with similar performance to conventional systems.

The remainder of the paper is organized as follows. In Section 2, we describe the Information

Bottleneck principle. Sections 2.1 and 2.2, respectively, summarize agglomerative and sequential optimization of the IB objective functions. Section 3 discusses methods for inferring the number of speakers. Section 4 describes the full diarization system, while Sections 5 and 6 present experiments and benchmark tests. Finally, Section 7 discusses results and conclusions.

2 Information Bottleneck Principle

The Information Bottleneck (IB) [23, 25] is a distributional clustering framework based on information theoretic principles. It is inspired from the Rate-Distortion theory [11] in which a set of elements X is organized into a set of clusters C minimizing the distortion between X and C. Unlike the Rate-Distortion theory, the IB principle does not make any assumption on the distance between elements of X. On the other hand, it introduces the use of a set of *relevance variables* Y which provides meaningful information about the problem. For instance, in a document clustering problem, the relevance variables could be represented by the vocabulary of words. Similarly, in a speech recognition, problem the relevance variables could be represented as the target sounds. IB tries to find the clustering representation C that conveys as much information as possible about Y. In this way the IB clustering attempts to keep the meaningful information with respect to a given problem.

Let Y be the 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. Let clusters C be a compressed representation of input data X. Thus, the information that X contains about Y is passed through the compressed representation (bottleneck) C. The Information Bottleneck (IB) principle states that this clustering representation should preserve as much information as possible about the relevance variables Y (i.e., maximize $I(Y, C)$) under a constraint on the mutual information between X and C i.e. $I(C, X)$. Dually, the clustering C should minimize the coding length (or the compression) of X using C i.e. $I(C, X)$ under the constraint on preserving the mutual information $I(Y, C)$. In other words, IB tries to find a trade-off between the most compact and most informative representation w.r.t. variables Y. This corresponds to maximization of the following criterion:

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

where β (notation consistent with [25]) is the Lagrange multiplier representing the trade off between amount of information preserved $I(Y, C)$ and the compression of the initial representation $I(C, X)$.

Let us develop mathematical expressions for $I(C, X)$ and $I(Y, C)$. The compression of the representation C is characterized by the mutual information $I(C, X)$:

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

The amount of information preserved about Y in the representation is given by

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

The objective function $\mathcal{F}$ must be optimized w.r.t the stochastic mapping $p(C|X)$ that maps each element of the dataset X into the new cluster representation C.

This minimization yields the following set of self -consistent equations that defines the conditional distributions required to compute mutual informations (2) and (3) (see [25] for details):

$$\begin{cases} p(c|x) &= \frac{p(c)}{Z(\beta,x)} \exp(-\beta D_{KL}[p(y|x)||p(y|c)]) \\ p(y|c) &= \sum_x p(y|x) p(c|x) \frac{p(x)}{p(c)} \\ p(c) &= \sum_x p(c|x) p(x) \end{cases} \tag{4}$$

where $Z(\beta, x)$ is a normalization function and $D_{KL}[p||q] = \sum_y p(y) log \frac{p(y)}{q(y)}$ represents the Kullback-Liebler divergence.

We can notice from the system of equations (4) that for $\beta \rightarrow \infty$ the stochastic mapping $p(c|x)$ becomes a hard partition of X, i.e. $p(c|x)$ can take values 0 and 1 only.

Various methods to construct solutions of the IB objective function include iterative optimization, deterministic annealing, agglomerative and sequential clustering (for exhaustive review see [23]). Here we focus only on two techniques referred to as agglomerative and sequential information bottleneck, which will be briefly presented in the next sections.

2.1 Agglomerative Information Bottleneck

Agglomerative Information Bottleneck (aIB) [22] is a greedy approach to maximize the objective function (1). The aIB algorithm creates hard partitions of the data. The algorithm is initialized with the trivial clustering of $|X|$ clusters i.e, each data point is considered as a cluster. Subsequently, elements are iteratively merged such that the decrease in the objective function (1) at each step is minimum.

The decrease in the objective function $\Delta\mathcal{F}$ obtained by merging clusters c_i and c_j is given by:

$$\Delta\mathcal{F}(c_i, c_j) = (p(c_i) + p(c_j)) \cdot \bar{d}_{ij} \tag{5}$$

where $\bar{d}_{ij}$ is given as a combination of two Jensen-Shannon divergences:

$$\bar{d}_{ij} = JS[p(y|c_i), p(y|c_j)] - \frac{1}{\beta} JS[p(x|c_i), p(x|c_j)] \tag{6}$$

where JS denotes the Jensen-Shannon (JS) divergence between two distributions and is defined as:

$$\begin{aligned} JS(p(y|c_i), p(y|c_j)) &= \pi_i\, D_{KL}[p(y|c_i)||q_Y(y)] + \\ &\quad +\pi_j\, D_{KL}[p(y|c_j)||q_Y(y)] \qquad (7) \\ JS(p(x|c_i), p(x|c_j)) &= \pi_i\, D_{KL}[p(x|c_i)||q_X(x)] + \\ &\quad +\pi_j\, D_{KL}[p(x|c_j)||q_X(x)] \qquad (8) \end{aligned}$$

with:

$$\begin{aligned} q_Y(y) &= \pi_i\, p(y|c_i) + \pi_j\, p(y|c_j) \qquad (9) \\ q_X(x) &= \pi_i\, p(x|c_i) + \pi_j\, p(x|c_j) \qquad (10) \\ \pi_i &= p(c_i)/(p(c_i) + p(c_j)) \\ \pi_j &= p(c_j)/(p(c_i) + p(c_j)) \end{aligned}$$

$D_{KL}(.)$ denotes the KL divergence between two distributions. JS divergence can be written as sum of two KL divergences as in (7). The objective function (1) decreases monotonically with the number of clusters. The algorithm merges cluster pairs until the desired number of clusters is attained. The new cluster c_r obtained by merging the individual clusters c_i and c_j is characterized by:

$$\begin{aligned} p(c_r) &= p(c_i) + p(c_j) \qquad (11) \\ p(y|c_r) &= \frac{p(y|c_i)p(c_i) + p(y|c_j)p(c_j)}{p(c_r)} \qquad (12) \end{aligned}$$

It is interesting to notice that the JS divergence is not an arbitrarily introduced similarity measure between elements but a measure that naturally arises from the maximization of the objective function. For completeness we report the full procedure described in [22] in Fig 1.

However, at each agglomeration step, the algorithm takes the merge decision based only on local criterion. Thus aIB is a greedy algorithm and produces only an approximation to the optimal solution which may not be the global solution to the objective function.

Input:

Joint Distribution $p(x, y)$

Trade-off parameter β

Output:

C_m: m-partition of X, $1 \leq m \leq |X|$

Initialization:

$C \equiv X$

For $i = 1 \ldots N$

$c_i = \{x_i\}$
$p(c_i) = p(x_i)$
$p(y|c_i) = p(y|x_i) \forall y \in Y$
$p(c_i|x_j) = 1$ if $j = i, 0$ otherwise

For $i, j = 1 \ldots N, i < j$

Find $\Delta\mathcal{F}(c_i, c_j)$

Main Loop:

While $|C| > 1$

$\{i, j\} = argmin_{i',j'} \Delta\mathcal{F}(c_i, c_j)$
Merge $\{c_i, c_j\} \Rightarrow c_r$ in C
$p(c_r) = p(c_i) + p(c_j)$
$p(y|c_r) = \frac{[p(y|c_i)p(c_i) + p(y|c_j)p(c_j)]}{p(c_r)}$
$p(c_r|x) = 1, \forall x \in c_i, c_j$
Calculate $\Delta\mathcal{F}(c_r, c), \forall c \in C$

Figure 1: Agglomerative IB algorithm [23]

2.2 Sequential Information Bottleneck

Sequential Information Bottleneck (sIB) [24] tries to improve the objective function (1) in a given partition. Unlike agglomerative clustering, it works with a fixed number of clusters M. The algorithm starts with an initial partition of the space into M clusters $\{c_1, ..., c_M\}$. Then some element x is drawn out of its cluster c_{old} and represents a new singleton cluster. x is then merged into the cluster c_{new} such that $c_{new} = argmin_{c \epsilon C} \Delta\mathcal{F}(x, c)$ where $\Delta\mathcal{F}(.,.)$ is as defined in (5). It can be verified that if $c_{new} \neq c_{old}$ then $\mathcal{F}(C_{new}) < \mathcal{F}(C_{old})$ i.e., at each step the objective function (1) either improves or stays unchanged. This process is repeated several times until there is no change in the clustering assignment. To avoid local maxima, this procedure can be repeated with several random initializations. The sIB algorithm is summarized for completeness in Fig 2.

3 Model Selection

In typical diarization tasks, the number of speakers in a given audio stream is not a priori available and must be estimated from data. This problem is often casted into a model selection problem. Consider a dataset X, and a set of parametric models $\{m_1, \cdots, m_M\}$ where m_j is a parametric model with n_j parameters trained on the data X. Model selection aims at finding the model $\hat{m}$ such that:

$$\hat{m} = argmax_j\{p(m_j|X)\} = argmax_j \left[\frac{p(X|m_j)p(m_j)}{p(X)}\right] \tag{13}$$

Given that $p(X)$ is constant and assuming uniform prior probabilities $p(m_j)$ on models m_j, maximization of (13) only depends on $p(X|m_j)$. In case of parametric modeling with parameter set θ_j, e.g. HMM/GMM, it is possible to write:

$$p(X|m_j) = \int p(X, \theta_j|m_j)d\theta_j \tag{14}$$

This integral cannot be computed in close form in case of complex parametric models with hidden variables (e.g. HMM/GMM). However several approximations for (14) are possible, the most popular one being the *Bayesian Information Criterion (BIC)* [20]:

$$BIC(m_j) = log(p(X|\hat{\theta_j}, m_j)) - \frac{p_j}{2} log\, N \tag{15}$$

where p_j is the number of free parameters in the model m_j, $\hat{\theta_j}$ is the MAP estimate of the model computed from data X, and N is the number of data samples. Rationale behind (15) is straightforward: models with larger number of parameters will produce higher values of $log(p(X|\theta_j, m_j))$ but will be more penalized by the term $\frac{p_j}{2} log\, N$. Thus the optimal model is the one that achieves the best trade-off between data explanation and complexity in terms of number of parameters. However, BIC is exact only in the asymptotic limit $N \to \infty$. It has been shown [10] that in case of finite sample case, like in speaker clustering problems, the penalty term must be tuned according to a heuristic threshold. In [4, 3, 5], a modified BIC criterion that needs no heuristic tuning has been proposed and will be discussed in more details in Section 6.1.

In the case of IB clustering, there is no parametric model that represents the data and model selection criteria based on a Bayesian framework like BIC cannot be applied. Several alternative solutions have been considered in the literature.

Because of the information theoretic basis, it is straightforward to apply the *Minimum Description Length* (MDL) principle [21]. The MDL principle is a formulation of the model selection problem from the information theory perspective. The optimal model minimizes the following criterion.

$$\mathcal{F}_{MDL}(m) = L(m) + L(X|m) \tag{16}$$

Input:

Joint Distribution $p(x, y)$

Trade-off parameter β

Cardinality value M

Output:

Partition C of X into M clusters

Initialization:

$C \leftarrow$ random partition of X into M clusters

For every $c_i \in C$:

$p(c_i) = \sum_{x_j \in c_i} p(x_j)$

$p(y|c_i) = \frac{1}{p(c_i)} \sum_{x_j \in c_i} p(y|x_j)p(x_j)$

$p(c_i|x_j) = 1$ if $x_j \in c_i$, 0 otherwise

Main Loop:

While not *Done*

$Done \leftarrow TRUE$

For every $x \in X$:

$c_{old} \leftarrow$ cluster x belongs

Split $c_{old} \Rightarrow \{c'_{old}, \{x\}\}$

$p(c'_{old}) = p(c_{old}) - p(x)$

$p(y|c'_{old}) = \frac{p(y|c_{old})p(c_{old}) - p(y,x)}{p(c'_{old})}$

$p(c'_{old}|x_i) = 1; \forall x_i \in c_{old}, x_i \neq x$

$c_{new} = argmin_{c \in C} \Delta\mathcal{F}(\{x\}, c)$

Merge $\{c_{new}, \{x\}\} \Rightarrow c'_{new}$

$p(c'_{new}) = p(c_{new}) + p(x)$

$p(y|c'_{new}) = \frac{p(y|c_{new})p(c_{new}) + p(y,x)}{p(c'_{new})}$

$p(c'_{new}|x_i) = 1; \forall x_i \in c_{new}, x_i = x$

If $c_{new} \neq c_{old}$

$Done \leftarrow FALSE$

Figure 2: Sequential IB algorithm [23]

where $L(m)$ is the code length to encode the model with a fixed length code and $L(X|m)$ is the code length required to encode the data given the model. As model complexity increases, the model explain the data better, resulting in a decrease in number of bits to encode the data given the model (lower $L(X|m)$). However, the number of bits required to encode the model increases (high $L(m)$). Thus, MDL selects a model that has the right balance between the model complexity and data description.

In case of IB clustering, let $N = |X|$ be the number of input samples, and $M = |C|$ the number of clusters. The number of bits required to code the model m and the data X given the model is :

$$L(m) = N \log \frac{N}{M} \quad (17)$$

$$L(X|m) = N[H(Y|C) + H(C)] \quad (18)$$

Since $H(Y|C) = H(Y) - I(Y,C)$ the MDL criterion becomes:

$$\mathcal{F}_{MDL} = N[H(Y) - I(Y,C) + H(C)] + N \log \frac{N}{M} \quad (19)$$

Similar to the BIC criterion, $N \log \frac{N}{M}$ acts like a penalty term that penalizes codes that uses too many clusters.

Another way of inferring the right number of clusters can be based on the *Normalized Mutual Information (NMI)* $\frac{I(Y,C)}{I(X,Y)}$. The Normalized Mutual Information $\frac{I(Y,C)}{I(X,Y)}$ represents the fraction of original mutual information that is captured by the current clustering representation. This quantity decreases monotonically with the number of clusters (see Figure 3). It can also be expected that this quantity will decrease more when dissimilar clusters are merged. Hence, we investigate a simple thresholding of $\frac{I(Y,C)}{I(X,Y)}$ as a possible choice to determine the number of clusters. The threshold is heuristically determined on a separate development data set.

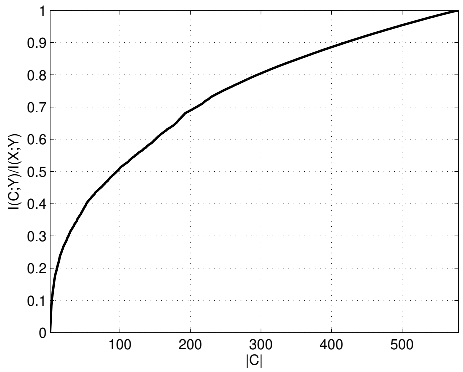


Figure 3: Normalized mutual information decreases monotonically with the number of clusters.

4 Applying IB to Diarization

To apply Information Bottleneck principle to the diarization problem, we need to define input variables X to be clustered and the relevance variables Y representing the meaningful information about the input.

In the initial case of document clustering, documents represent the input variable X. The vocabulary of words is selected as the relevance variable. Associated conditional distributions $\{p(y_i|x_j)\}$ are the probability of each word y_i in document x_j. Documents can be clustered together with IB using the fact that similar documents will have similar probabilities of containing same words.

In this paper, we investigate the use of IB to clustering of speech segments according to speaker similarity. We define in the following the input variables $X = \{x_j\}$, the relevance variables $Y = \{y_i\}$ and the conditional probabilities $p(y_i|x_j)$.

4.1 Input Variables X

The Short Time Fourier Transform (STFT) of the input audio signal is computed using 30ms windows shifted by a step of 10ms. 19 MFCC features are extracted from each windowed frame. Let $\{s_1, s_2, \cdots s_T\}$ be the extracted MFCC features. Subsequently, a uniform linear segmentation is performed on the feature sequence to obtain segments of a fixed length D (typically 2.5 seconds). The input variables X are defined as the set of these segments $\{x_1, x_2, \cdots, x_M\}$. Thus each segment x_j consists of a sequence of MFCC features $\{s_k^j\}_{k=1,\cdots,D}$.

If the length of the segment is small enough, X may be considered as generated by a single speaker. This hypothesis is generally true in case of Broadcast News audio data. However 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.

4.2 Relevance Variables Y

Motivated from the fact that GMMs are widely used in speaker recognition and verification systems (see e.g. [18]), we choose the relevant variables $Y = \{y_j\}$ as components of a GMM estimated from the meeting data. A shared covariance matrix GMM is estimated from the entire audio file.

The computation of conditional probabilities $p(Y = y_i|X = x_j)$ is straightforward. Consider a Gaussian Mixture Model $f(s) = \sum_{j=1}^{M} w_j \mathcal{N}(s, \mu_j, \Sigma_j)$ where M is the number of components, w_j are weights, μ_j means and Σ_j covariance matrices. It is possible to project each speech frame s_k onto the space of Gaussian components of the GMM. Adopting the notation used in previous sections, the space induced by GMM components would represent the relevance variable Y.

Computation of $p(y_i|s_k)$ is then simply given by:

$$p(y_i|s_k) = \frac{w_i \mathcal{N}(s_k, \mu_i, \Sigma_i)}{\sum_{j=1}^{N} w_j \mathcal{N}(s_k, \mu_j, \Sigma_j)}; \; i = 1, \ldots, N \tag{20}$$

The probability $p(y_i|s_k)$ estimates the relevance that the i^{th} component in the GMM has for speech frame s_k. Since segment x_j is composed of several speech frames $\{s_k^j\}$, distributions $\{p(y_i|s_k^j)\}$ can be averaged over the length of the segment to get the conditional distribution $p(Y|X)$.

In other words, a speech segment X is projected into the space of relevance variables Y estimating a set of conditional probabilities $p(Y|X)$.

4.3 Clustering

Given the variables X and Y, the conditional probabilities $p(Y|X)$, and trade-off parameter β, Information Bottleneck clustering can be performed. The diarization system involves two tasks: finding the number of clusters (i.e. speakers) and an assignment for each speech segment to a given cluster.

The procedure we use is based on the agglomerative IB described in Section 2.1. The algorithm is initialized with M clusters with $M = |X|$ and agglomerative clustering is performed, generating a set of possible solutions in between M and 1 clusters.

Out of the $M = |X|$ possible clustering solutions of aIB, we choose the one that maximizes the model selection criteria described in Section 3 i.e. *Minimum Description Length* or *Normalized Mutual Information.*

However, agglomerative clustering does not seek the global optimum of the objective function and can converge to local minima. For this reason, sIB algorithm described in Section 2.2 can be applied to improve the partition. Given that sIB works only on fixed cardinality clustering, we propose to use it to improve the greedy solution obtained with the aIB.

To summarize, we study in the following four different types of clustering/model selection algorithms:

1 agglomerative IB + MDL model selection.

2 agglomerative IB + NMI model selection.

3 agglomerative IB + MDL model selection + sequential IB.

4 agglomerative IB + NMI model selection + sequential IB.

4.4 Diarization algorithm

We can summarize the complete diarization algorithm as follows:

1 Extract acoustic features $\{s_1, s_2, \cdots, s_T\}$ from the audio file.

2 Speech/non-speech segmentation and reject non-speech frames.

3 Uniform segmentation of speech in chunks of fixed size D, i.e. definition of set $X = \{x_1, x_2, \cdots, x_M\}$.

4 Estimation of GMM with shared diagonal covariance matrix i.e. definition of set Y.

5 Estimation of conditional probability $p(Y|X)$.

6 Clustering based on one of the methods described in Section 4.3.

7 Viterbi realignment using conventional GMM system estimated from previous segmentation.

Step 1 and 2 are common to all diarization systems. Speech is segmented into fixed length segment in step 3. This step tries to obtain speech segments that contain speech from only one speaker. We use a uniform segmentation in this work though other solutions could be employed like speaker change detection or K-means algorithm.

Step 4 trains a background GMM model with shared covariance matrix from the entire audio stream. Though we use data from the same meeting, it is possible to train the GMM on a large independent dataset i.e. a Universal Background Model (UBM) can be used.

Step 5 involves conditional probability $p(y|x)$ estimation. In step 6 clustering and model selection are performed on the basis of the Information Bottleneck principle.

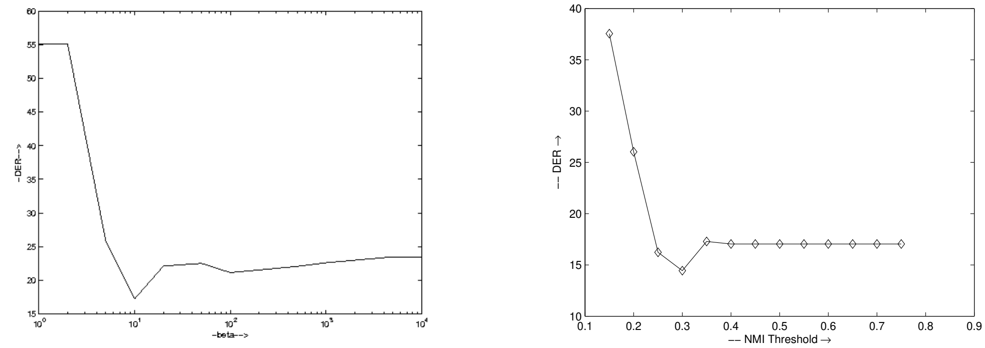
Step 7 refines initial segmentation by performing a set of Viterbi realignments. Given the inferred number of speakers and a mapping from X segments to C clusters, a GMM is trained for each speaker c_j using data x_i Then the whole meeting data is re-aligned using Viterbi algorithm. This step does not change the mapping from X to C but modifies the initial boundaries obtained from uniform segmentation.

5 Experiments and Results

5.1 Data description

The data used for the experiment consists of meeting recordings obtained using an array of far-field microphones also referred as Multiple Distant Microphones (MDM). Those data contains mainly conversational speech with high speaker change rate and represent a very challenging data set.

We study the impact of different system parameters on the development dataset which contains meetings from previous years NIST evaluations for "Meeting Recognition Diarization" task [2]. This development dataset contains 12 meeting recordings each one around 10 minutes. The best set of



(a) (b)

Figure 4: Effect of varying different parameters on the diarization error. (a) DER as a function of NMI threshold. (b) DER as a function of parameter β. The optimal parameters are chosen as NMI threshold $= 0.3$ and $\beta = 10$

parameters is then used for benchmarking the proposed system against a state-of-the-art diarization system. Comparison is performed on the NIST RT06 evaluation data for "Meeting Recognition Diarization" task . The dataset contains nine meeting recordings of approximately 30 minutes each.

Preprocessing consists of the following steps: signal recorded with Multiple Distant Microphones are filtered using a Wiener filter denoising for individual channels followed by a delay-and-sum beamforming [6],[7]. This was performed using the *BeamformIt* toolkit [29]. Such pre-processing produces a single enhanced audio signal out those recorded with the far-field microphones. 19 MFCC features are then extracted from the beam-formed signal.

The system performance is evaluated in terms of Diarization Error Rates (DER). DER is the sum of missed speech errors (speech classified as non-speech), false alarm speech error (non-speech classified as speech) and speaker error[1]. Speech/non-speech (spnsp) error is the sum of missed speech and false alarm speech.

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 [13]. Results are scored against manual references forced aligned by an ASR system. Being interested in comparing the clustering algorithms, the same speech/non-speech segmentation will be used across all experiments.

In the following we study the impact of the trade-off parameter β (Section5.2), the performance of the agglomerative and sequential clustering (Section 5.3),the model selection criterion (Section 5.4) and the effect of the Viterbi re-alignment on development data.

5.2 Trade-off β

The parameter β represents the trade-off between the amount of information preserved and the level of compression. To determine its value, we studied the diarization error of the IB algorithm in the development dataset. The performance of the algorithm is studied by varying β on a log-linear scale and applying aIB clustering. The optimal number of speaker is chosen according to an oracle. Thus, the influence of the parameter can be studied independently of model selection methods or thresholds. The diarization error (DER) for different values of beta is presented in Fig 4. Those results do not include Viterbi re-alignment. The value of $\beta = 10$ produce the lowest Diarization Error Rate.

Figure 5 shows the DER curve w.r.t. number of clusters for two meetings (NIST_20051024-0930 and CMU_20050914-0900). We can notice that the DER is flat for $\beta = 1$ and does not decrease with increase in number of clusters. This low value of β implies more weighting to the regularization term $\frac{1}{\beta} I(C, X)$ of the objective function in Equation 1. Thus the optimization tries to minimize $I(C, X)$.

Since $I(C, X) = H(C)$ for hard partitions, this forces $P(C)$ to be a low entropy distribution. This leads to a highly unbalanced distribution where most of the elements are assigned to one single cluster. Thus the algorithm always converges towards one large cluster followed by several spurious clusters and the DER stays almost constant. On the otherhand, when β is high (eg: $\beta = \infty$), effect of this regularization term vanishes. The optimization criterion focuses only on the relevance variable set $I(Y, C)$ regardless of the data compression. DER curve thus becomes less smooth.

For intermediate values of β, the clustering seeks the most informative *and* compact representation. ($|C|$). For the value of $\beta = 10$, the region of low DER is almost constant for comparatively more values of $|C|$. In this case, the algorithm forms large speaker clusters initially. Most of the remaining clusters are small and merging these clusters does not change the DER considerably. This results in a regularized DER curve as function of number of speakers (see Figure 5).

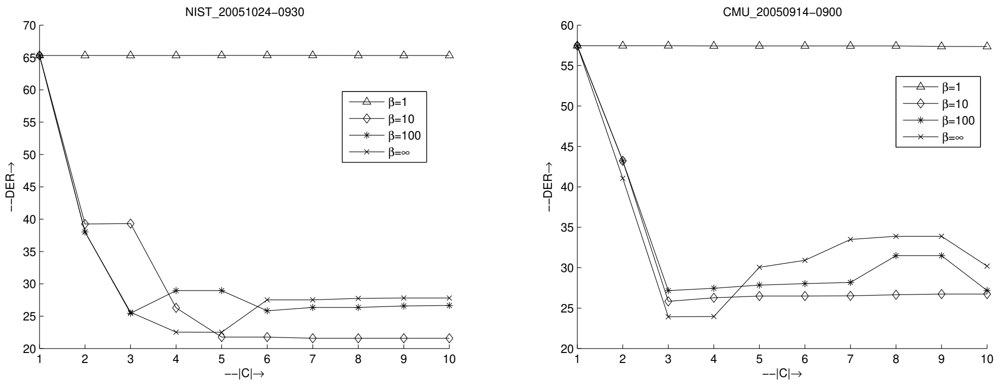


Figure 5: DER as a function of number of clusters ($|C|$) for different values of parameter β

5.3 Agglomerative and Sequential clustering

In this section, we compare the agglomerative and sequential clustering described in Sections 2.1,2.2 on the development data. As before model selection is performed using an oracle and the value of β is fixed to 10 as found in the previous section. Agglomerative clustering achieves a DER of 13.3% while sequential clustering achieves a DER of 12.4% i.e. 1% absolute better. Results are presented in Table 1. Improvements are obtained on 8 of the 12 meetings included in the development data.

5.4 Model selection

In this section, we discuss experimental results with model selection algorithms presented in Section 3. Two different model selection criteria – Normalized Mutual Information (NMI) and Minimum Description Length (MDL) – are investigated to select the number of speakers. They are compared with an oracle model selection which manually choose the clustering with the lowest DER.

The Normalized Mutual Information is a monotonically increasing function with the number of clusters. The value is compared against a threshold to determine the optimal number of speaker in the model. Figure 4 illustrates the change of DER with changing the value of this threshold. The lowest DER is obtained for the value of 0.3. The MDL criterion described in equation(19) is also explored for performing model selection. Speaker error rates corresponding to both the methods are reported in Table 2. NMI criterion outperform MDL model selection by $\sim 2\%$. The NMI criterion is 2.5% worse then the oracle model selection.

After Viterbi re-alignment of the data, the DER is reduced by roughly 3% absolute for all the different methods. The lowest DER is obtained using sequential clustering with NMI model selection.

Table 1: Diarization error rate of development data for individual meetings for aIB and aIB+sIB using oracle model selection and without Viterbi re-alignment.

Meeting	aIB	aIB + sIB
AMI_20041210-1052	4.6	3.7
AMI_20050204-1206	10.0	8.3
CMU_20050228-1615	25.3	25.2
CMU_20050301-1415	9.4	10.1
ICSI_20000807-1000	12.3	13.2
ICSI_20010208-1430	12.9	13.0
LDC_20011116-1400	8.7	7.0
LDC_20011116-1500	18.7	17.5
NIST_20030623-1409	6.0	5.7
NIST_20030925-1517	24.3	23.9
VT_20050304-1300	7.3	5.2
VT_20050318-1430	29.7	25.6
ALL	13.3	12.4

Table 2: Diarization Error Rates for dev dataset with NMI, MDL and oracle model selection.

	aIB		aIB+sIB	
Model selection	without Viterbi	with Viterbi	without Viterbi	with Viterbi
Oracle	13.3	10.3	12.4	10.0
MDL	17.3	14.3	16.2	13.8
NMI	15.4	12.6	14.3	12.5

6 RT06 Meeting Diarization

In this section we compare the IB system with a state-of-the-art diarization system based on HMM/GMM. Results are provided for the NIST RT06 evaluation data. Section 6.1 describe the baseline system while section and Section 6.2 describes results of the IB based system.

6.1 Baseline System

The baseline system is an ergodic HMM as described in [4, 7]. Each HMM state represents a speaker. The state emission probabilities are modeled by Gaussian Mixture Models(GMM) with a minimum duration constrain of $2.5s$. The algorithm follows an agglomerative framework, i.e, it starts with a large number of clusters (hypothesized speakers) and then iteratively merge similar clusters till it reaches the best model.

The initial HMM model is built using uniform linear segmentation and each speaker is modeled with a 5 components GMM. The algorithm then proceeds with bottom-up agglomerative clustering of the initial speaker models [10]. At each step, all possible cluster merges are compared using a modified version of BIC criterion[20, 4] which is described below.

Consider a pair of clusters c_i and c_j with associated data D_i and D_j respectively. Also let the number of parameters for modeling each cluster be p_i and p_j parameterized by the GMM models m_i and m_j. Assume the new cluster c having data D obtained by merging D_i and D_j is modeled with a GMM model m parameterized by p Gaussians. The pair of clusters that results in maximum increase in the BIC Criterion (given by equation 15) are merged.

$$(i', j') = argmax_{i,j} BIC(m) - [BIC(m_j) + BIC(m_i)] \tag{21}$$

In the work [4], the model complexity (i.e. the number of parameters) before and after the merge is made the same. This is achieved by keeping the number of gaussians in the new model m as the sum of number of gaussians in m_j and m_i. i.e., $p = p_i + p_j$. Under this condition equation (21) reduces to

$$(i', j') = argmax_{i,j} \log \frac{p(D|m)}{p(D_i|m_i)p(D_j|m_j)} \tag{22}$$

This eliminates the need of the penalty term from the BIC criterion. Following the merge, all speaker models are updated using an EM algorithm. The merge/re-estimation continues until no merge results in an increase in the BIC criterion. This decides the number of speakers in the final model. This approach yields state-of-the art results [7] in several diarization evaluations. The performance of the baseline system is presented in Table 3. The table lists missed speech, false alarm, speaker error and diarization error. [1]

Table 3: Results of the baseline system

File	Miss	FA	spnsp	spkr err	DER
ALL	6.5	0.1	6.6	17.0	23.6

6.2 Results

In this section we benchmark the IB based diarization system on RT06 data. The same speech/non-speech segmentation is used for all methods. According to results of previous sections the value of β is fixed to 10. The NMI threshold value is fixed to 0.3. Viterbi re-alignment of the data is performed after the clustering with a minimum duration constrain of $2.5s$ to refine speaker boundaries.

Table 4 reports results for aIB and aIB+sIB clustering. Results for both NMI and MDL criteria are reported.

NMI is more effective then MDL by 0.7%. Sequential clustering (aIB+sIB) outperforms agglomerative clustering by 0.5%. As in the development data, the best results is obtained by aIB+sIB clustering with NMI model selection. This system achieves a DER of 23.2% as compared to 23.6% of the baseline system.

Table 5 reports diarization error for individual meetings of the RT06 evaluation data set. We can notice that overall performances are very close to those of the baseline system but results per meeting are quite different.

Table 4: Diarization Error Rate for RT06 evaluation data.

Model selection	aIB+Viterbi	sIB+Viterbi
MDL	24.4	23.8
NMI	23.7	**23.2**

Table 6 denotes the number of speakers estimated by different algorithms for the eval data. The number of speakers is mostly higher than the actual. This happens due to the presence of small spurious clusters with very short duration (typically less then 5 seconds). However those small clusters does not significantly affect the final Diarization Error Rate.

[1] We found that one channel of the meeting in RT06 denoted with VT_20051027-1400 is considerably degraded. This channel was removed before beamforming. This produces better results for both baseline and IB systems compared to those presented in [26].

Table 5: Diarization error rate for individual meetings using NMI model selection.

		Viterbi realign	
Meeting	Baseline	aIB	aIB + sIB
CMU_20050912-0900	17.8	20.1	18.7
CMU_20050914-0900	15.3	21.9	20.8
EDI_20050216-1051	46.0	48.5	50.5
EDI_20050218-0900	23.8	33.3	33.1
NIST_20051024-0930	12.0	16.2	17.3
NIST_20051102-1323	23.7	15.7	15.0
TNO_20041103-1130	31.5	28.7	26.1
VT_20050623-1400	24.4	9.6	9.4
VT_20051027-1400	21.7	20.0	18.4
ALL	23.6	23.7	23.2

Table 6: Estimated number of speakers by different model selection criteria.

		aIB + sIB	
Meeting	#speakers	NMI	MDL
CMU_20050912-0900	4	5	5
CMU_20050914-0900	4	6	6
EDI_20050216-1051	4	7	7
EDI_20050218-0900	4	7	7
NIST_20051024-0930	9	7	7
NIST_20051102-1323	8	7	7
TNO_20041103-1130	4	7	6
VT_20050623-1400	5	8	8
VT_20051027-1400	4	6	4

6.3 Algorithm Complexity

Both the The IB bottleneck algorithm and the baseline HMM/GMM system uses the agglomerative clustering framework. Let the number of clusters at a given step in the agglomeration be K. At each step, the agglomeration algorithm needs to calculate the distance measure between each pair of clusters. i.e., $\frac{1}{2}K(K-1)$ distance calculations.

In the HMM/GMM model, each distance calculation involves computing the BIC criterion as given by equation (22). Thus, the new parametric model m has to be estimated for every possible merge. This require training a GMM model for every pair of clusters. On the otherhand, the distance measure in the IB framework is the combination of two Jenson-Shannon divergences as described by equation (6). The JS divergence calculation is straightforward and computationally very efficient.

All the experiments are benchmarked on a desktop machine with AMD Athlon$^{\mathrm{TM}}$ 2.4GHz 64 X2 Dual Core Processor and 2GB RAM. Table 7 lists the real time factors for the baseline and IB based diarization systems. It can be seen that the IB based systems are significantly faster than HMM/GMM based system. Note that most of the algorithm time for IB systems is consumed for estimating the posterior features. The clustering is very fast and takes only around 30% of the total algorithm time. Overall the proposed diarization system is considerably faster than-real time.

7 Discussions and Conclusions

We presented a speaker diarization systems based on information theoretic framework known as the Information Bottleneck. This system can achieve Diarization Error rates close to those obtained

Table 7: Real time factors for different algorithms on RT06 eval data

method	posterior calculation	clustering	Viterbi realign	Total
aIB	0.09	0.06	0.07	0.22
aIB +sIB	0.09	0.08	0.07	0.24
Baseline	–	–	–	3.5

with conventional HMM/GMM agglomerative clustering. In the following we discuss main differences between this framework and traditional approaches.

- *Distance measure*: in literature, several distance measure have already been proposed for clustering speakers e.g. BIC, generalized log-likelihood ratio, KL divergence and cross-likelihood distances. IB principle states that when the clustering seeks the solution that preserve as much information as possible w.r.t a set of relevance variables, the optimal distance between clusters is represented by the *Jensen-Shannon* divergence (see equation 7). JS divergence can be written as sum of two KL divergences and has many appealing properties related to Bayesian error (see [14] for detailed discussion). This similarity measure in between clusters is not arbitrary introduced but is naturally derived from the IB objective function (see [22]).

- *Regularization*: The trade-off parameter β between amount of mutual information and compression regularize the clustering solution as shown in Section 5.2. We verified that this term can reduce the DER and make the DER curve more smooth with the number of clusters.

- *Parametric Speaker Model*: HMM/GMM based systems build an explicit parametric model for each cluster and for each possible merge. This assume that each speaker provides enough data for estimating such a model. On the other hand, the system presented here is based on the distance between clusters in a space of relevance variables without any explicit speaker model. The set of relevance variables is defined through a GMM estimated on the entire audio stream. Furthermore the obtained clustering techniques is significantly faster then conventional systems given that merges are estimated in a space of discrete probabilities.

- *Sequential clustering*: Conventional systems based on agglomerative clustering (aIB) can produce sub-optimal solutions due to its greedy nature. On the other hand, sequential clustering (sIB) seeks a global optimum of the objective function. In Sections 5.3 and 6.2, it is shown that sequential clustering outperforms agglomerative clustering by $\sim 1\%$ on development and evaluation data set. The sequential clustering can be seen as a "purification" algorithm. In literature methods aiming at obtaining clusters that contain speech from a single speaker are referred as "purification" methods. They refine the agglomerative solution according to smoothed log-likelihood [28] or cross Expectation-Maximization in between models [17] for findings frames that were wrongly assigned. In case of sIB, the purification is done according to the same objective function and the correct assignment of each speech segment is based on the amount of mutual information it conveys on the relevance variables. Furthermore, as reported in Table 7, its computational complexity is only marginally higher then the one obtained using agglomerative clustering.

In conclusion the proposed system based on IB principle can achieve on RT06 evaluation data a DER of 23.2% as compared to 23.6% of HMM/GMM baseline while running 0.3xRT i.e. significantly faster then the baseline system.

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, I. McCowan, and H. Bourlard. Robust speaker change detection. *Signal Processing Letters, IEEE*, 11(8):649–651, 2004.

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

[5] Jitendra Ajmera. *Robust Audio Segmentation*. PhD thesis, Ecole Polytechnique Federale de Lausanne (EPFL), 2004.

[6] 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.

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

[8] C. Barras, X. Zhu, S. Meignier, and J.L. Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1505–1512, 2006.

[9] M. Ben and F. Bimbot. D-MAP: a distance-normalized MAP estimation of speaker models for automatic speaker verification. *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 2, 2003.

[10] 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.

[11] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & sons, 1991.

[12] J. Goldberger, H. Greenspan, and S. Gordon. Unsupervised image clustering using the information bottleneck method. In *Proceedings of the 24th DAGM Symposium on Pattern Recognition*, pages 158–165, 2002.

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

[14] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[15] S. Meignier, J.F. Bonastre, C. Fredouille, and T. Merlin. Evolutive HMM for multi-speaker tracking system. *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, 2, 2000.

[16] H. Ning, M. Liu, H. Tang, and T.S. Huang. A Spectral Clustering Approach to Speaker Diarization. In *Ninth International Conference on Spoken Language Processing*. ISCA, 2006.

[17] H. Ning, W. Xu, Y. Gong, and T. Huang. Improving Speaker Diarization by Cross EM Refinement. In *IEEE International Conference on Multimedia and Expo*, pages 1901–1904, 2006.

[18] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[19] D.A. Reynolds, E. Singer, B.A. Carlson, G.C. O'Leary, J.J. McLaughlin, and M.A. Zissman. Blind Clustering of Speech Utterances Based on Speaker and Language Characteristics. In *Fifth International Conference on Spoken Language Processing*. ISCA, 1998.

[20] G. Schwartz. Estimation of the dimension of a model. *Annals of Statistics*, 6, 1978.

[21] Y. Seldin, N. Slonim, and N. Tishby. Information bottleneck for non co-occurrence data. In *Advances in Neural Information Processing Systems 19*. MIT Press, 2007.

[22] 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.

[23] Noam Slonim. *The Information Bottleneck: Theory and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2002.

[24] 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 Retrieval*, 2002.

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

[26] 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.

[27] 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)*, pages 4361–4364, 2008.

[28] Anguera X. and J. Wooters, C.and Hernando. Purity algorithms for speaker diarization of meetings data. In *Proceedings of ICASSP*, 2006.

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

[30] 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 Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 693–698, 2007.