An Information Theoretic Approach to Speaker Diarization of Meeting Data

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

***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, the 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 tradeoff 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 meetings. The IB-based system achieves a diarization error rate of 23.2% compared to 23.6% for the baseline system. This approach being mainly based on nonparametric clustering, it runs significantly faster than the baseline HMM/GMM based system, resulting in faster-than-real-time diarization.**

***Index Terms*—Information bottleneck (IB), meetings data, speaker diarization.**

I. 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 and 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 remains the one proposed in [1] 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* (BIC) [2] as a model complexity metric has been proposed in [1] 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 [1], modified versions of BIC [3], [4], Generalized Log-Likelihood Ratio [5], Kullback–Leibler divergence [6], or cross-likelihood distance [7]. The choice of this distance measure is somewhat arbitrary.

In this paper, we investigate the use of a clustering technique motivated from an information theoretic framework known as the *Information Bottleneck* (IB) [8]. The IB method has been applied to clustering of different types of data like documents [9], [10] and images [11]. IB clustering [8], [12] is a distributional clustering inspired from Rate-Distortion theory [13]. In contrast 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 between elements. Furthermore, given a data set to be clustered, IB tries to find the tradeoff 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 a hidden Markov model/Gaussian mixture model (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 meeting recordings. The most recent efforts in the NIST Rich Transcription campaigns focus on meeting data acquired in several rooms with different acoustic properties and with a variable number of speakers. The audio data is recorded in a nonintrusive manner using multiple distant microphones (MDM) or a microphone array. Given the variety of acoustic environments, the conversational nature of recordings and the use of distant microphones, those recordings represent a very challenging data set. Progress in the diarization task for meeting data can be found in [14] and in [15].

Recently, attention has shifted onto faster-than-real-time diarization systems with low computational complexity (see, e.g., [16]–[19]). In fact, in the meeting case scenario, faster than real-time diarization would enable several applications (meeting browsing, meeting summarization, speaker retrieval)

Manuscript received July 04, 2008; revised December 23, 2008. Current version published July 17, 2009. This work was supported in part by the European Union under the integrated projects AMIDA, Augmented Multi-Party Interaction with Distance Access under Contract IST-033812 and in part by the 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 IT02. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Timothy J. Hazen.

The authors are with the Idiap Research Institute, CH-1920 Martigny, Switzerland (e-mail: dvijaya@idiap.ch).

Digital Object Identifier 10.1109/TASL.2009.2015698

on a 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 cluster with emission probabilities represented by GMM probability density functions [3], [20]. Merging two segments implies estimating a new GMM model that represents data coming from both segments as well as the similarity measure between the new GMM and the remaining speaker clusters. This procedure can be computationally demanding.

As second contribution, this paper also investigates the IB clustering for a fast speaker diarization system. IB is a nonparametric framework that does not use any explicit modeling of speaker clusters. Thus, the algorithm does not need to estimate a GMM for each cluster, resulting in a considerably reduced computational complexity with similar performance to conventional systems.

The remainder of the paper is organized as follows. In Section II, we describe the Information Bottleneck principle. Sections II-A and II-B, respectively, summarize agglomerative and sequential optimization of the IB objective functions. Section III discusses methods for inferring the number of clusters. Section IV describes the full diarization system, while Sections V and VI present experiments and benchmark tests. Finally, Section VII discusses results and conclusions.

II. Information Bottleneck (IB) Principle

The IB principle [8], [12] is a distributional clustering framework based on information theoretic principles. It is inspired from the Rate-Distortion theory [13] 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 about 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 $\forall x \in X$ and $\forall y \in Y$ 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 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 of preserving the mutual information $I(Y,C)$. In other words, IB tries to find a tradeoff 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 [8]) 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 $I(C,X)$

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

The amount of information preserved about Y in the representation is given by $I(Y,C)$

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

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 [8] for details)

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

where $Z(\beta,x)$ is a normalization function and $D_{KL}[.,.]$ represents the Kullback–Liebler divergence given by

$$D_{KL}\left[p(y|x)||p(y|c)\right] = \sum_{y \in Y} p(y|x) \log \frac{p(y|x)}{p(y|c)}. \tag{5}$$

We can see from the system of (4) that as $\beta \to \infty$ 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 [12]). Here, we focus only on two techniques referred to as agglomerative and sequential information bottleneck, which will be briefly presented in the next sections.

A. Agglomerative Information Bottleneck

Agglomerative Information Bottleneck (aIB) [9] 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

$$\Delta\mathcal{F}(c_i, c_j) = (p(c_i) + p(c_j)) \cdot \bar{d}_{ij} \tag{6}$$

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

$$\bar{d}_{ij} = JS\left[p(y|c_i), p(y|c_j)\right] - \frac{1}{\beta} JS\left[p(x|c_i), p(x|c_j)\right] \tag{7}$$

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

$$\begin{aligned} JS\left(p(y|c_i), p(y|c_j)\right) =& \pi_i D_{KL}\left[p(y|c_i)||q_Y(y)\right] \\ &+ \pi_j D_{KL}\left[p(y|c_j)||q_Y(y)\right] \end{aligned} \tag{8}$$

$$\begin{aligned} JS\left(p(x|c_i), p(x|c_j)\right) =& \pi_i D_{KL}\left[p(x|c_i)||q_X(x)\right] \\ &+ \pi_j D_{KL}\left[p(x|c_j)||q_X(x)\right] \end{aligned} \tag{9}$$

with

$$q_Y(y) = \pi_i p(y|c_i) + \pi_j p(y|c_j) \tag{10}$$

$$\begin{aligned} q_X(x) =& \pi_i p(x|c_i) + \pi_j p(x|c_j) \\ \pi_i =& p(c_i)/\left(p(c_i) + p(c_j)\right) \\ \pi_j =& p(c_j)/\left(p(c_i) + p(c_j)\right). \end{aligned} \tag{11}$$

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

$$p(c_r) = p(c_i) + p(c_j) \tag{12}$$

$$p(y|c_r) = \frac{p(y|c_i)p(c_i) + p(y|c_j)p(c_j)}{p(c_r)}. \tag{13}$$

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 [12] in Fig. 1.

However, at each agglomeration step, the algorithm takes the merge decision based only on a 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.

B. Sequential Information Bottleneck

Sequential Information Bottleneck (sIB) [10] 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 $\{c_1, \ldots, c_M\}$. 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 $c_{\text{new}} = \arg\min_{c \in C} \Delta\mathcal{F}(x, c)$ where $\Delta\mathcal{F}(.,.)$ is as defined in (6). It can be verified that if $c_{\text{new}} \neq c_{\text{old}}$ then $\mathcal{F}(C_{\text{new}}) < \mathcal{F}(C_{\text{old}})$, i.e., at each step the objective function (1) either improves or stays unchanged. This is performed for each $x \in X$. This process is repeated several times until there is no change in the clustering assignment for any input element. To avoid local maxima, this procedure can be repeated with several random initializations. The sIB algorithm is summarized for completeness in Fig. 2.

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 $|C| > 1$
$\{i, j\} = \arg\min_{i',j'} \Delta\mathcal{F}(c_i, c_j)$
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 $\Delta\mathcal{F}(c_r, c)$, $\forall c \in C$

Fig. 1. Agglomerative IB algorithm [12].

III. Model Selection

In typical diarization tasks, the number of speakers in a given audio stream is not a priori known and must be estimated from data. This means that the diarization system has to solve simultaneously two problems: finding the actual number of speakers and clustering together speech from the same speaker. This problem is often cast into a model selection problem. The number of speakers determines the complexity of the model in terms of number of parameters. The model selection criterion chooses the model with the right complexity and thus the number of speakers. Let us consider the theoretical foundation of model selection.

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

$$\hat{m} = \arg\max_j \{p(m_j|X)\} = \arg\max_j \left[\frac{p(X|m_j)p(m_j)}{p(X)}\right]. \tag{14}$$

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 $c_{old} \Rightarrow \{c'_{old}, \{x\}\}$
$p(c'_{old}) = p(c_{old}) - p(x)$
$p(y|c'_{old}) = \frac{p(y|c_{old})p(c_{old}) - p(y,x)}{p(c'_{old})}$
$p(c'_{old}|x_i) = 1; \forall x_i \in c_{old}, x_i \neq x$
$c_{new} = \arg\min_{c \in C} \Delta\mathcal{F}(\{x\}, c)$
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*

Fig. 2. Sequential IB algorithm [12].

Given that $p(X)$ is constant and assuming uniform prior probabilities $p(m_j)$ on models m_j, maximization of (14) 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

$$p(X|m_j) = \int p(X, \theta_j|m_j) d\theta_j. \tag{15}$$

This integral cannot be computed in closed form in the case of complex parametric models with hidden variables (e.g., HMM/GMM). However, several approximations for (15) are possible, the most popular one being the *Bayesian Information Criterion* (BIC) [2]

$$BIC(m_j) = log\left(p(X|\hat{\theta}_j, m_j)\right) - \frac{p_j}{2} logN \tag{16}$$

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. The rationale behind (16) is straightforward: models with larger numbers of parameters will produce higher values of $log(p(X|\theta_j, m_j))$ but will be more penalized by the term $(p_j/2)logN$. Thus, the optimal model is the one that achieves the best tradeoff between data explanation and complexity in terms of number of parameters. However, BIC is exact only in the asymptotic limit $N \rightarrow \infty$. It has been shown [1] that in the finite sample case, like in speaker clustering problems, the penalty term must be tuned according to a heuristic threshold. In [3], [4], [21], a modified BIC criterion that needs no heuristic tuning has been proposed and will be discussed in more details in Section VI-A.

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 [22]. The MDL principle is a formulation of the model selection problem from an information theory perspective. The optimal model minimizes the following criterion:

$$\mathcal{F}_{MDL}(m) = L(m) + L(X|m) \tag{17}$$

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 explains 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

$$L(m) = N \log \frac{N}{M} \tag{18}$$

$$L(X|m) = N\left[H(Y|C) + H(C)\right]. \tag{19}$$

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

$$\mathcal{F}_{MDL} = N\left[H(Y) - I(Y, C) + H(C)\right] + N \log \frac{N}{M}. \tag{20}$$

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

When aIB clustering is applied, expression (20) is evaluated for each stage of the agglomeration that produces $|X|$ different clustering solutions ranging from each input element considered as a singleton cluster ($|C| = |X|$) to all input elements assigned to one cluster ($|C| = 1$). Then, the number of clusters that minimizes (20) is selected as the actual number of speakers.

Another way of inferring the right number of clusters can be based on the *Normalized mutual information* (NMI) $I(Y, C)/I(X, Y)$. The NMI $I(Y, C)/I(X, Y)$ 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 Fig. 3). It can also be expected that this quantity will decrease more when dissimilar clusters are merged. Hence, we investigate a simple thresholding of $I(Y, C)/I(X, Y)$ as a possible choice to determine the number of clusters. The threshold is heuristically determined on a separate development data set.

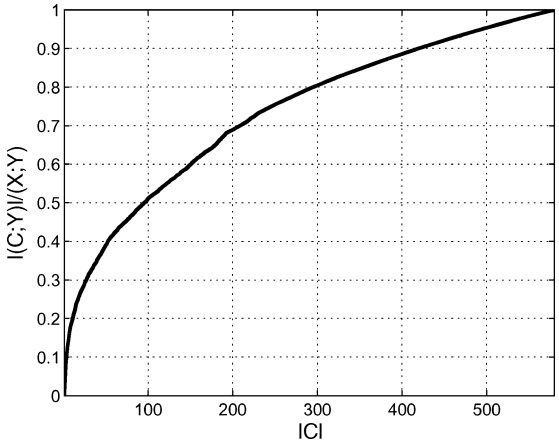

Fig. 3. Normalized mutual information decreases monotonically with the number of clusters.

IV. Applying IB to Diarization

To apply the 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 the same words.

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

A. Input Variables X

The short-time Fourier transform (STFT) of the input audio signal is computed using 30-ms windows shifted by a step of 10 ms. Nineteen Mel frequency cepstral coefficients (MFCCs) 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 s). 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 meeting data), initial segments may contain speech from several speakers.

B. Relevance Variables Y

Motivated by the fact that GMMs are widely used in speaker recognition and verification systems (see, e.g., [23]), 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 number of components of the GMM is fixed proportional to the length of the meeting, i.e., the GMM has P/D components where P is the length of the audio stream (in seconds) and D is length of segments (in seconds) defined in Section IV-A.

The computation of conditional probabilities $p(Y = y_i|X = x_j)$ is straightforward. Consider a GMM $f(s) = \sum_{j=1}^{L} w_j \mathcal{N}(s, \mu_j, \Sigma_j)$, where L 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

$$p(y_i|s_k) = \frac{w_i \mathcal{N}(s_k, \mu_i, \Sigma_i)}{\sum_{j=1}^{L} w_j \mathcal{N}(s_k, \mu_j, \Sigma_j)}; \quad i = 1, \ldots, L. \quad (21)$$

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 $p(Y|X)$.

C. Clustering

Given the variables X and Y, the conditional probabilities $p(Y|X)$, and tradeoff 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 II-A. 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 one according to the model selection criteria described in Section III, i.e., MDL or NMI.

However, agglomerative clustering does not seek the global optimum of the objective function and can converge to local minima. For this reason, the sIB algorithm described in Section II-B 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 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.

D. 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/nonspeech segmentation and reject nonspeech 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 IV-C.
7) Viterbi realignment using conventional GMM system estimated from previous segmentation.

Steps 1 and 2 are common to all diarization systems. Speech is segmented into fixed length segments 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 like speaker change detection or K-means algorithm could be employed.

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 uniform segmentation by performing a set of Viterbi realignments. This step modifies the speaker boundaries and is discussed in the following section.

E. Viterbi Realignment

As described in Section IV-A, the algorithm clusters speech segments of a fixed length D. Hence, the cluster boundaries obtained from the IB are aligned with the endpoints of these segments. Those endpoints are clearly arbitrary and can be improved by realigning the whole meeting using a Viterbi algorithm.

The Viterbi realignment is performed using an ergodic HMM. Each state of the HMM represents a speaker cluster. The state emission probabilities are modeled with GMMs, with a minimum duration constraint. Each GMM is initialized with a fixed number of components.

The IB clustering algorithm infers the number of clusters and the assignment from X segments to C clusters. A separate GMM for each cluster is trained using data assignment produced by the IB clustering. The whole meeting data is then realigned using the ergodic HMM/GMM models. During the realignment, a minimum duration constraint of 2.5 s is used as well.

V. Effect of System Parameters

In this section, we study the impact of the tradeoff parameter β (Section V-B), the performance of the agglomerative and sequential clustering (Section V-C), the model selection criterion (Section V-D) and the effect of the Viterbi realignment (Section V-E) on development data.

A. Data Description

The data used for the experiments consist of meeting recordings obtained using an array of far-field microphones also referred as MDMs. Those data contain 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 [14]. This development dataset contains 12 meeting recordings each one around 10 min. The best set of 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 min each. After evaluation, the TNO_20041103-1130 was found noisy and was not included in the official evaluation. However, results are reported with/without this meeting in the literature [24], [25]. We present results with and without this meeting for the purpose of comparison.

Preprocessing consists of the following steps: signals recorded with MDMs are filtered using a Wiener filter denoising for individual channels followed by a delay-and-sum beamforming [15], [26]. This was performed using the *BeamformIt* toolkit [27]. Such preprocessing produces a single enhanced audio signal from individual far-field microphone channels. Nineteen MFCC features are then extracted from the beam-formed signal.

The system performance is evaluated in terms of Diarization Error Rates (DERs). DER is the sum of missed speech errors (speech classified as nonspeech), false alarm speech error (nonspeech classified as speech), and speaker error [28]. Speech/nonspeech (spnsp) error is the sum of missed speech and false alarm speech. For all experiments reported in this paper, we include the overlapped speech in the evaluation.

Speech/nonspeech segmentation is obtained using a forced alignment of the reference transcripts on close talking microphone data using the AMI RT06 first pass ASR models [29]. Results are scored against manual references force aligned by an ASR system. Being interested in comparing the clustering algorithms, the same speech/nonspeech segmentation will be used across all experiments. The missed speech, false alarm speech and total speech/nonspeech error for all meetings in the development dataset and evaluation dataset are listed in Table I and Table II, respectively.

B. Tradeoff β

The parameter β represents the tradeoff 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 clusters 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 Rate (DER) of the development dataset

TABLE I
MISSED SPEECH, FALSE ALARM, AND TOTAL SPEECH/NONSPEECH ERROR FOR THE DEVELOPMENT DATASET

Meeting	Miss	FA	spnsp
AMI_20041210-1052	0.40	1.20	1.60
AMI_20050204-1206	2.60	2.10	4.70
CMU_20050228-1615	9.40	1.10	0.50
CMU_20050301-1415	3.80	1.60	5.40
ICSI_20000807-1000	4.70	0.30	5.00
ICSI_20010208-1430	3.70	1.00	4.70
LDC_20011116-1400	2.10	1.70	3.80
LDC_20011116-1500	5.90	1.00	6.90
NIST_20030623-1409	1.00	0.60	1.60
NIST_20030925-1517	7.70	5.70	3.40
VT_20050304-1300	0.60	1.00	1.60
VT_20050318-1430	1.40	6.20	7.60
ALL	3.50	1.80	5.30

TABLE II
MISSED SPEECH, FALSE ALARM, AND TOTAL SPEECH/NONSPEECH ERROR FOR THE EVALUATION DATASET

Meeting	Miss	FA	spnsp
CMU_20050912-0900	11.60	0.20	11.80
CMU_20050914-0900	10.30	0.00	10.30
EDI_20050216-1051	4.90	0.10	5.00
EDI_20050218-0900	4.30	0.10	4.40
NIST_20051024-0930	7.00	0.20	7.20
NIST_20051102-1323	6.10	0.10	6.20
TNO_20041103-1130	3.80	0.10	3.90
VT_20050623-1400	5.20	0.20	5.40
VT_20051027-1400	3.50	0.30	3.80
ALL	6.50	0.10	6.60

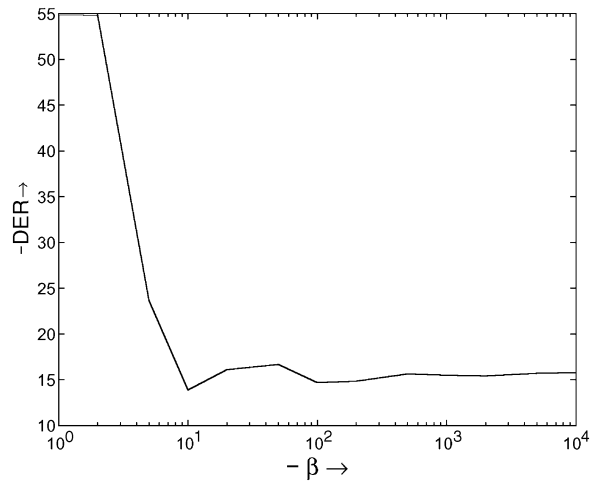

Fig. 4. Effect of varying parameter β on the diarization error for the development dataset. The optimal β is chosen as $\beta = 10$.

for different values of beta is presented in Fig. 4. These results do not include Viterbi realignment. The value of $\beta = 10$ produce the lowest DER. In order to understand how the optimal value of β changes across different meetings, we report in Table III optimal β for each meeting, DER for the optimal β and for $\beta = 10$. In eight meetings out of the 12, the β that produces the lowest DER is equal to 10. In four meetings the optimal β is different from 10, but only in one (CMU_20050228) the DER is significantly different from the one obtained using $\beta = 10$. To summarize the optimal value of β seems to be consistent across different meetings.

TABLE III
OPTIMAL VALUE FOR β FOR EACH MEETING IN THE DEVELOPMENT DATASET. DER FOR THE OPTIMAL β AS WELL AS $\beta = 10$ ARE REPORTED

Meeting	optimal β	DER at optimal β	DER at $\beta = 10$
AMI_20041210-1052	10	4.6	4.6
AMI_20050204-1206	10	10.0	10.0
CMU_20050228-1615	50	20.4	25.3
CMU_20050301-1415	10	9.4	9.4
ICSI_20000807-1000	100	11.9	12.3
ICSI_20010208-1430	10	12.9	12.9
LDC_20011116-1400	1000	6.2	8.7
LDC_20011116-1500	10	18.7	18.7
NIST_20030623-1409	10	6.0	6.0
NIST_20030925-1517	10	24.3	24.3
VT_20050304-1300	10	7.3	7.3
VT_20050318-1430	100	28.5	29.7

Fig. 5 shows the DER curve w.r.t. number of clusters for two meetings (LDC_20011116-1400 and CMU_20050301-1415). It can be seen that the DER is flat for $\beta = 1$ and does not decrease with the increase in number of clusters. This low value of β implies more weighting to the regularization term $(1/\beta)I(C,X)$ of the objective function in (1). Thus, the optimization tries to minimize $I(C,X)$. The algorithm uses hard partitions, i.e., $p(c|x) \in \{0,1\}$, this leads to $H(C|X) = -\sum_{x\in X} p(x)\sum_{c\in C} p(c|x)\log p(c|x) = 0$ and as a result $I(C,X) = H(C) - H(C|X) = H(C)$. Hence, minimizing $I(C,X)$ is equivalent to minimizing $H(C)$. Thus, $H(C)$ is minimized while clustering with low values of β. This leads to a highly unbalanced distribution where most of the elements are assigned to one single cluster($H(C) \approx 0$). Thus, the algorithm always converges towards one large cluster followed by several spurious clusters and the DER stays almost constant. Conversely, when β is high (e.g., $\beta = \infty$), the effect of this regularization term vanishes. The optimization criterion focuses only on the relevance variable set $I(Y,C)$ regardless of the data compression. The DER curve thus becomes less smooth.

For intermediate values of β, the clustering seeks the most informative *and* compact representation. 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 a function of number of clusters (see Fig. 5).

C. Agglomerative and Sequential Clustering

In this section, we compare the agglomerative and sequential clustering described in Sections II-A and II-B on the development data. As before, model selection is performed using an oracle and the value of β is fixed at 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 IV. Improvements are obtained on eight of the 12 meetings included in the development data.

Also the additional computation introduced by the sequential clustering is small when initialized with aIB output. The sIB algorithm converges faster in this case than using random initial

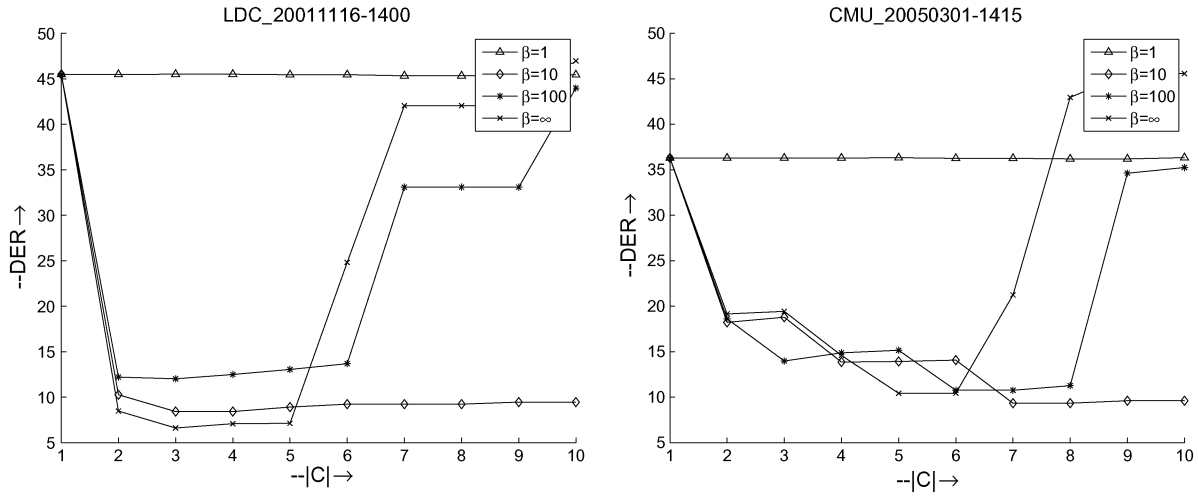

Fig. 5. DER as a function of number of clusters ($|C|$) for different values of parameter β.

TABLE IV
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

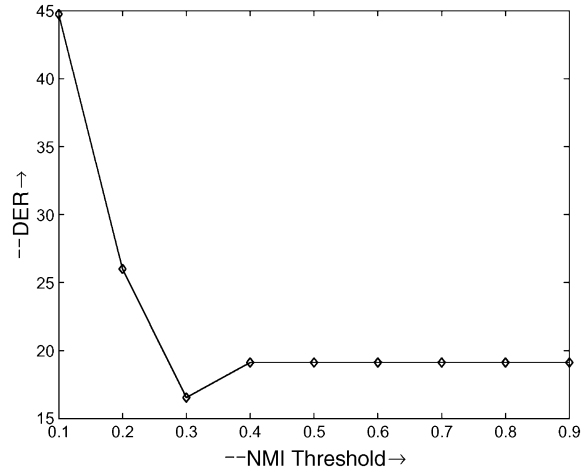

Fig. 6. Effect of varying NMI threshold on the diarization error for the development dataset. The optimal threshold is fixed as 0.3.

partitions (four iterations as compared to six iterations on an average across the development dataset).

D. Model Selection

In this section, we discuss experimental results with the model selection algorithms presented in Section III. Two different model selection criteria— NMI and MDL—are investigated to select the number of clusters. They are compared with an oracle model selection which manually chooses the clustering with the lowest DER. The NMI is a monotonically increasing function with the number of clusters. The NMI value is compared against a threshold to determine the optimal number of clusters in the model. Fig. 6 illustrates the change of overall DER over the whole development dataset for changing the value of this threshold. The lowest DER is obtained for the value of 0.3. In order to understand how the optimal value of the threshold changes across different meetings, we report in Table V optimal threshold for each meeting, DER for the optimal threshold and for threshold equal to 0.3. In eight out the 12 meetings in the development data set, the threshold that produces the lowest DER is equal to 0.3. Only in two meetings (ICSI_20000807-1000 and NIST_20030925-1517) results obtained with the optimal threshold are significantly different from those obtained with the value 0.3. To summarize the optimal value of the threshold seems to be consistent across different meetings.

The MDL criterion described in (20) is also explored for performing model selection. Speaker error rates corresponding to both the methods are reported in Table VI. The NMI criterion outperforms the MDL model selection by $\sim$2%. The NMI criterion is 2.5% worse than the oracle model selection.

E. Viterbi Realignment

The Viterbi realignment is carried out using an ergodic HMM as discussed in Section IV-E. The number of components of each GMM is fixed at 30 based on experiments on the development dataset. The performance after Viterbi realignment is presented in Table VI. 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.

VI. 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

TABLE V
OPTIMAL VALUE FOR NMI THRESHOLD FOR EACH MEETING IN THE DEVELOPMENT DATASET. THE DER IS REPORTED FOR THE OPTIMAL VALUE AS WELL AS FOR 0.3. THE CLUSTERING IS PERFORMED WITH $\beta = 10$

Meeting	optimal NMI threshold	DER at opt th.	DER at thres 0.3
AMI_20041210-1052	0.3	9.6	9.6
AMI_20050204-1206	0.3	14.9	14.9
CMU_20050228-1615	0.3	26.5	26.5
CMU_20050301-1415	0.3	9.6	9.6
ICSI_20000807-1000	0.4	13.5	20.0
ICSI_20010208-1430	0.3	14.4	14.4
LDC_20011116-1400	0.3	9.2	9.2
LDC_20011116-1500	0.2	20.6	21.9
NIST_20030623-1409	0.4	7.8	11.9
NIST_20030925-1517	0.4	25.2	30.6
VT_20050304-1300	0.3	5.9	5.9
VT_20050318-1430	0.3	34.9	34.9

TABLE VI
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

provided for the NIST RT06 evaluation data. Section VI-A describes the baseline system while Section VI-B describes the results of the IB-based system. Section VI-C compares the computational complexity of the two systems.

A. Baseline System

The baseline system is an ergodic HMM as described in [3], [15]. Each HMM state represents a cluster. The state emission probabilities are modeled by GMMs with a minimum duration constrain of 2.5 s. Nineteen MFCC coefficients extracted from the beam-formed signal are used as the input features. The algorithm follows an agglomerative framework, i.e., it starts with a large number of clusters (hypothesized speakers) and then iteratively merges similar clusters until it reaches the best model. After each merge, data are realigned using a Viterbi algorithm to refine speaker boundaries.

The initial HMM model is built using uniform linear segmentation and each cluster is modeled with a five-component GMM. The algorithm then proceeds with bottom-up agglomerative clustering of the initial cluster models [1]. At each step, all possible cluster merges are compared using a modified version of the BIC criterion [2], [3] 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 respectively 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 the maximum increase in the BIC criterion [given by (16)] are merged

$$(i', j') = \arg\max_{i,j} \mathrm{BIC}(m) - [\mathrm{BIC}(m_j) + \mathrm{BIC}(m_i)]. \quad (22)$$

TABLE VII
RESULTS OF THE BASELINE SYSTEM

File	Miss	FA	spnsp	spkr err	DER
All meetings	6.5	0.1	6.6	17.0	23.6
Without TNO meeting	6.8	0.1	6.9	15.7	22.7

TABLE VIII
DIARIZATION ERROR RATE FOR RT06 EVALUATION DATA

Model selection	aIB+ Viterbi	sIB+ Viterbi
All meetings		
MDL	24.4	23.8
NMI	23.7	**23.2**
Without TNO meeting		
MDL	23.9	23.5
NMI	23.0	**22.8**

In [3], 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 the same, i.e., as the sum of number of Gaussians in m_j and m_i, i.e., $p = p_i + p_j$. Under this condition, (22) reduces to

$$(i', j') = \arg\max_{i,j} \log \frac{p(D|m)}{p(D_i|m_i)p(D_j|m_j)}. \quad (23)$$

This eliminates the need of the penalty term from the BIC. Following the merge, all cluster models are updated using an EM algorithm. The merge/reestimation continues until no merge results in any further increase in the BIC criterion. This determines the number of clusters in the final model. This approach yields state-of-the art results [15] in several diarization evaluations. The performance of the baseline system is presented in Table VII. The table lists missed speech, false alarm, speaker error, and diarization error.[1]

B. Results

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

Table VIII reports results for aIB and aIB+sIB clustering both with/without TNO meeting. Conclusions are drawn on the original data set. Results for both NMI and MDL criteria are reported. NMI is more effective than MDL by 0.7%. Sequential clustering (aIB+sIB) outperforms agglomerative clustering by 0.5%. As in the development data, the best results are obtained by aIB+sIB clustering with NMI model selection. This system achieves a DER of 23.2% as compared to 23.6% for the baseline system.

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

[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 [16].

TABLE IX
DIARIZATION ERROR RATE FOR INDIVIDUAL MEETINGS USING NMI MODEL SELECTION

Meeting	Baseline	Viterbi realign	
		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

TABLE X
ESTIMATED NUMBER OF SPEAKERS BY DIFFERENT MODEL SELECTION CRITERIA

Meeting	#speakers	aIB + sIB	
		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

attributed to the different optimization criteria used by the two systems—BIC criterion for the baseline system and IB criterion for the proposed system.

Furthermore, the IB clustering is based on the use of a set of relevance variables defined as the components of a background GMM. The GMM is estimated using data from the same meeting. As variations in signal properties like signal-to-noise-ratio (SNR) and amount of overlapping speech can deteriorate the quality of the GMM thus the clustering results. For instance, the performance of the IB system are comparatively low for CMU meetings which contain large amounts of overlapping speech and low SNR. On the other hand, IB performs considerably better then the baseline system on VT meetings that have high SNR and TNO meeting which has very less overlapping speech.

Table X shows the number of speakers estimated by different algorithms for the RT06 eval data. The number of speakers is mostly higher than the actual. This is due to the presence of small spurious clusters with very short duration (typically less than 5 s). However those small clusters does not significantly affect the final DER.

C. Algorithm Complexity

Both the IB bottleneck algorithm and the baseline HMM/GMM system use 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., $(1/2)K(K-1)$ distance calculations. Let us consider the difference between the two methods.

- In the HMM/GMM model, each distance calculation involves computing the BIC criterion as given by (23). Thus, a new parametric model m has to be estimated for every possible merge. This requires training a GMM model for every pair of clusters. The training is done using the EM algorithm which is computationally demanding. In other words, this method involves the use of EM parameter estimation for every possible cluster merge.
- In the IB framework, the distance measure is the sum of two Jensen–Shannon divergences as described by (7). The JS divergence calculation is straightforward and computationally very efficient. Thus, the distance calculation in the IB frame work is much faster as compared to the HMM/GMM approach. The distribution obtained merging two clusters is given by (12), (13) which simply consists in averaging distributions of individual clusters.

In summary, while the HMM/GMM systems make intensive use of the EM algorithm, the IB-based system performs the clustering in the space of discrete distributions using closed form equations for distance calculation and cluster distribution update. Thus, the proposed approach require less computation than the baseline.

TABLE XI
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

We perform benchmark experiments on a desktop machine with AMD Athlon 2.4-GHz 64 X2 Dual Core Processor and 2 GB of RAM. Table XI lists the real time factors for the baseline and IB-based diarization systems for the RT06 meeting diarization task. 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. Also, introducing the sequential clustering contributes very little to the total algorithm time ($\approx$8%). Overall the proposed diarization system is considerably faster than-real time.

VII. DISCUSSIONS AND CONCLUSION

We have presented speaker diarization systems based on the information theoretic framework known as the Information Bottleneck. This system can achieve Diarization Error rates close to those obtained with conventional HMM/GMM agglomerative clustering. In the following, we discuss main differences between this framework and traditional approaches.

- *Distance measure*: in the literature, several distance measures have already been proposed for clustering speakers, e.g., BIC, generalized log-likelihood ratio, KL divergence and cross-likelihood distances. The IB principle states that when the clustering seeks the solution that preserves 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 (8)]. JS divergence can be written as the sum of two KL divergences and has many appealing properties related to Bayesian error (see

[30] for detailed discussion). This similarity measure between clusters is not arbitrarily introduced but is naturally derived from the IB objective function (see [9]).

- *Regularization*: The tradeoff parameter β between amount of mutual information and compression regularizes the clustering solution as shown in Section V-B. We verified that this term can reduce the DER and make the DER curve more smooth against 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 assumes 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 resulting clustering techniques are significantly faster than conventional systems given that merges are estimated in a space of discrete probabilities.
- *Sequential clustering*: Conventional systems based on agglomerative clustering (aIB) can produce suboptimal solutions due to their greedy nature. Conversely, sequential clustering (sIB) seeks a global optimum of the objective function. In Sections V-C and VI-B, it is shown that sequential clustering outperforms agglomerative clustering by $\sim$1% on development and $\sim$0.5% evaluation data sets. The sequential clustering can be seen as a "purification" algorithm. In the literature, methods aiming at obtaining clusters that contain speech from a single speaker are referred to as "purification" methods. They refine the agglomerative solution according to smoothed log-likelihood [31] or cross expectation-maximization between models [32] for finding 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 XI, its computational complexity is only marginally higher than the one obtained using agglomerative clustering.

In conclusion, the proposed system based on the IB principle can achieve on RT06 evaluation data a DER of 23.2% as compared to 23.6% of HMM/GMM baseline while running $0.3 \times \mathrm{RT}$, i.e., significantly faster than the baseline system.

ACKNOWLEDGMENT

Authors would like to thank Dr. C. Wooters and Dr. X. Anguera for their help with baseline system and beam-forming toolkit. They would also like to thank Dr. J. Dines for his help with the speech/non-speech segmentation and Dr. P. Garner for his suggestions for improving the paper as well as all project partners for a fruitful collaboration. The Authors would like to thank all the reviewers for their valuable comments and suggestions which improved the paper considerably.

REFERENCES

[1] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Speech Recognition Workshop*, 1998, pp. 127–138.
[2] G. Schwartz, "Estimation of the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.
[3] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. IEEE Autom. Speech Recognition Understanding Workshop*, 2003, pp. 411–416.
[4] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE Signal Process. Lett.*, vol. 11, no. 8, pp. 649–651, Aug. 2004.
[5] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Multistage speaker diarization of broadcast news," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1505–1512, Oct. 2006.
[6] M. Ben and F. Bimbot, "D-MAP: A distance-normalized MAP estimation of speaker models for automatic speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03).* , 2003, vol. 2, pp. 69–72.
[7] D. Reynolds, E. Singer, B. Carlson, G. O'Leary, J. McLaughlin, and M. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," in *Proc. 5th Int. Conf. Spoken Lang. Process.*, 1998, ISCA.
[8] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," NEC Research Institute TR, 1998.
[9] N. Slonim, N. Friedman, and N. Tishby, "Agglomerative information bottleneck," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 617–623, MIT Press.
[10] F. F. Slonim and T. N. , "Unsupervised document classification using sequential information maximization," in *Proc. SIGIR'02, 25th ACM Int. Conf. Res. Develop. Inf. Retireval*, 2002.
[11] J. Goldberger, H. Greenspan, and S. Gordon, "Unsupervised image clustering using the information bottleneck method," in *Proc. 24th DAGM Symp. Pattern Recognition*, 2002, pp. 158–165.
[12] N. Slonim, "The information bottleneck: Theory and applications," Ph.D. dissertation, Hebrew Univ. of Jerusalem, Jerusalem, Israel, 2002.
[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[14] [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2006/spring/
[15] X. Anguera, "Robust Speaker Diarization for Meetings," Ph.D. dissertation, Univ. Politecnica de Catalunya, Catalunya, Spain, 2006.
[16] D. Vijayasenan, F. Valente, and H. Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in *Proc. IEEE Autom. Speech Recognition Understanding Workshop (ASRU)*, 2007, pp. 250–255.
[17] Y. Huang, O. Vinyals, G. Friedland, C. Müller, N. Mirghafori, and C. Wooters, "A fast-match approach for robust, faster than real-time speaker diarization," in *Proc. IEEE Autom. Speech Recognition Understanding Workshop (ASRU)*, 2007, pp. 693–698.
[18] D. Vijayasenan, F. Valente, and H. Bourlard, "Combination of agglomerative and sequential clustering for speaker diarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 4361–4364.
[19] H. Ning, M. Liu, H. Tang, and T. Huang, "A spectral clustering approach to speaker diarization," in *Proc. 9th Int. Conf. Spoken Lang. Process.*, 2006, pp. 2178–2181, ISCA.
[20] S. Meignier, J. Bonastre, C. Fredouille, and T. Merlin, "Evolutive HMM for multi-speaker tracking system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'00)* , 2000, vol. 2, pp. 201–204.
[21] J. Ajmera, "Robust Audio Segmentation," Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, 2004.
[22] Y. Seldin, N. Slonim, and N. Tishby, "nformation bottleneck for non co-occurrence data," in *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press, 2007.
[23] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.
[24] J. Fiscus, J. Ajot, M. Michel, and J. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," *Lecture Notes Comput. Sci.*, vol. 4299, p. 309, 2006.

[25] J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Trans. Comput.*, vol. 56, no. 9, p. 1189, Sep. 2007.
[26] X. Anguera, C. Wooters, and J. H. Hernando, "Speaker diarization for multi-party meetings using acoustic fusion," in *Proc. Autom. Speech Recognition Understanding*, 2006, pp. 426–431.
[27] X. Anguera, "Beamformit, the fast and robust acoustic beamformer," 2006 [Online]. Available: http://www.icsi.berkeley.edu/xanguera/BeamformIt
[28] [Online]. Available: http://nist.gov/speech/tests/rt/rt2004/fall/
[29] T. Hain *et al.*, "The ami meeting transcription system: Progress and performance," in *Proc. NIST RT'O6 Workshop*, 2006.
[30] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.
[31] A. X. , J. Wooters, C. , and Hernando, "Purity algorithms for speaker diarization of meetings data," in *Proc. ICASSP*, 2006.
[32] H. Ning, W. Xu, Y. Gong, and T. Huang, "Improving speaker diarization by cross EM refinement," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2006, pp. 1901–1904.

Deepu Vijayasenan (S'08) received the B.Tech. degree in electronics and communication from University of Kerala, Thiruvananthapuram, India, in 2000 and the M.E. degree in signal processing from Indian Institute of Science, Bangalore, India, in 2003. He is currently pursuing his Ph.D. degree atthe IDIAP Research Institute, Martigny, Switzerland.

From 2003 to 2006, he was a Research Associate with Hewlett Packard Laboratories Bangalore. His research interests include speech signal processing and machine learning.

Fabio Valente (M'06) received the M.Sc. degree (summa cum laude) in communication systems from Politecnico di Torino, Turin, Italy, in 2001, the M.Sc. degree in image processing from University of Nice-Sophia Antipolis, Nice, France, in 2002, and the Ph.D. degree in signal processing from University of Nice-Sophia Antipolis for his work on variational Bayesian methods for speaker diarization done at the Institut Eurecom, France, in 2005.

In 2001, he was with the Motorola Human Interface Lab (HIL), Palo Alto, CA. Since 2006, he has been with the Idiap Research Institute, Martigny, Switzerland, and is involved in several EU and U.S. projects on speech and audio processing. His main interests are in machine learning and speech recognition. He is the author or coauthor of several papers in international conferences and journals with contributions in feature extraction and selection for speech recognition, multistream ASR and Bayesian statistics for speaker diarization.

Hervé Bourlard (M'98–SM'95–F'00) received the Electrical and Computer Science Engineering degree and the Ph.D. degree in applied sciences, both from the Faculté Polytechnique de Mons, Mons, Belgium.

After having been a Member of the Scientific Staff at the Philips Research Laboratory of Brussels and an R&D Manager at L&H SpeechProducts, he is now Director of the Idiap Research Institute, Full Professor at the Swiss Federal Institute of Technology, Lausanne (EPFL), and Director of the National Center of Competence in Research in "Interactive Multimodal Information Management" (IM2). Having spent (since 1988) several long-term and short-term visits (initially as a Guest Scientist) at the International Computer Science Institute (ICSI) in Berkeley, CA, he is now a member of the ICSI Board of Trustees. His main interests are in signal processing, statistical pattern classification, multichannel processing, artificial neural networks, and applied mathematics, with applications to speech and natural language modeling, speech and speaker recognition, computer vision, and multimodal processing. He is the author, coauthor, or editor of four books and over 250 reviewed papers (including one IEEE paper award) and book chapters. Over the last 20 years, he has initiated and coordinated numerous large international research projects, as well as multiple collaborative projects with industries. He is an appointed expert for the European Commission and, from 2002 to 2007, was also part of the European Information Society Technology Advisory Group (ISTAG).

Dr. Bourlard is an IEEE Fellow for "contributions in the fields of statistical speech recognition and neural networks." He is (or has been) a member of the program/scientific committees of numerous international conferences (e.g., General Chairman of IEEE Workshop on Neural Networks for Signal Processing 2002, Co-Technical Chairman of ICASSP'02, General Chairman of Interspeech'03) and on the editorial board of several journals (e.g., past Co-Editor-in-Chief of *Speech Communication*).