IDIAP RESEARCH REPORT

AGGLOMERATIVE INFORMATION BOTTLENECK FOR SPEAKER DIARIZATION OF MEETINGS DATA

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

IDIAP–RR 07-31

July 2007

SUBMITTED FOR PUBLICATION

[a] IDIAP Research Institute, Martigny, Switzerland

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

IDIAP Research Report 07-31

AGGLOMERATIVE INFORMATION BOTTLENECK FOR SPEAKER DIARIZATION OF MEETINGS DATA

Deepu Vijayasenan, Fabio Valente, Hervé Bourlard

July 2007

SUBMITTED FOR PUBLICATION

Abstract. In this paper, we investigate the use of agglomerative Information Bottleneck (aIB) clustering for the speaker diarization task of meetings data. In contrary to the state-of-the-art diarization systems that models individual speakers with Gaussian Mixture Models, the proposed algorithm is completely non parametric . Both clustering and model selection issues of non-parametric models are addressed in this work. The proposed algorithm is evaluated on meeting data on the RT06 evaluation data set. The system is able to achieve Diarization Error Rates comparable to state-of-the-art systems at a much lower computational complexity.

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 ASR systems, speaker based indexing and retrieval. It involves determining the number of speakers and identification of the speech segments corresponding to each speaker. The number of speakers is not a priori known and must be estimated from data in an unsupervised manner. This is generally achieved using a model selection criterion for inferring the number of clusters (speakers).

Conventional diarization systems are based on an ergodic HMM in which each state represents a speaker. Emission probabilities are Gaussian Mixture Models (GMM). The initial audio stream is segmented into several regions both using speaker change detection methods or uniform segmentation. The diarization algorithm is then based on bottom-up agglomerative clustering of those initial segments [15]. Segments are merged according to some measure till 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 [9] as model complexity metric has been proposed in [15] and a modified version of BIC that keep the model complexity constant has been proposed in [3],[4]. These approaches yield state-of-the art results [11] in several diarization evaluations.

BIC criterion is computed using the ratio of likelihoods of the individual segments to the merged segment. Hence at each step, the algorithm has to estimate likelihood of individual clusters and of all possible merges. This assumes availability of enough data at each cluster to estimate the parametric model. Also this can be computationally very demanding since the parametric models(GMM) has to be re-estimated in each step.

In some applications like browsing meeting recording while the meeting is taking place, there is need for very fast diarization systems and HMM/GMM based systems may need significant optimization to achieve this goal. We instead propose the use of a non-parametric clustering algorithm with reduced computational complexity.

In this paper, we investigate an alternative solution based on the agglomerative Information Bottleneck (aIB) method proposed in [16]. aIB is a clustering algorithm based on information theoretic framework. The idea is to minimize the mutual information loss between successive clusterings while preserving the mutual information to a relevance variable(for details see [13]). The main advantage w.r.t. conventional HMM/GMM agglomerative approach is that there is no explicit representation of a speaker model. The aIB depends only on the representation of the relevant variables for classification . This results in dramatic reduction in cost of the clustering algorithms, since GMM parameters re-estimations are avoided in the clustering.

Given that there is no explicit parametric speaker model, the BIC cannot be directly applied in this case. However, other model selection criteria based on information theory are considered. We investigate stopping criteria based on the normalized mutual information as well as Minimum Description Length principle [7].

Experiments are performed on NIST RT06 “Meeting Recognition Diarization” task based on data from Multiple Distant Microphones (MDM) [2].

The remainder of the paper is organized as follows. In section 2 we describe the Information Bottleneck (IB) principle, in section 3 we describe the agglomerative Information Bottleneck which provides an approximated solution to the IB problem, in section 4 we address model selection issues and in 5 we describe how aIB can be applied to speaker clustering. In section 6 we give details on a speaker diarization system based on aIB clustering, in section 7 we provide several experiments to investigate the previously described algorithm comparing results with a state-of-the art baseline system both in terms of DER and computational time. Finally in section 8 we discuss conclusions and future works.

2 INFORMATION BOTTLENECK PRINCIPLE

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

IB method [16] is inspired from Rate-Distortion theory and aims at finding the most compact representation C of data X that minimizes the mutual information $I(X,C)$ and preserves as much information as possible about Y (maximizing $I(C,Y)$). Thus the IB objective function can be formulated as minimization of the Lagrangian,

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

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

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

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

As shown in [13], this minimization leads to the following equations that define conditional distributions needed for computing mutual informations (2) and (3):

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

$$p(y|c) = \sum_x p(y|x)p(c|x)\frac{p(x)}{p(c)} \tag{5}$$

$$p(c) = \sum_x p(c|x)p(x) \tag{6}$$

Where $Z(\beta,x)$ is a normalization function, $p(x)$ is prior distribution of x and the functional $D_{KL}[p(y|x)||p(y|c)]$ is the Kullback-Liebler divergence. Function (1) defines a concave curve in the (I_x, I_y) plane.

The limit $\beta \to \infty$ of equations (4-6) induces a hard partition of the input space i.e. the probabilistic map $p(c|x)$, takes values of 0 and 1 only. This is equivalent to minimizing only the information loss in the clustering i.e. $I(Y,C)$.

3 AGGLOMERATIVE INFORMATION BOTTLENECK

The agglomerative Information Bottleneck (aIB) [13] focuses on generating hard partitions of the data X using a greedy approach such that objective function of equation (1) is minimized. The algorithm is initialized with the trivial clustering of $|X|$ clusters; i.e. each data point is considered as a cluster. Subsequently the clusters are merged iteratively such that after each step the loss of mutual information w.r.t the relevant variables Y is minimum.

The loss of mutual information δI_y obtained by merging x_i and x_j is given by Jensen-Shannon divergence between $p(Y|x_i)$ and $p(Y|x_j)$:

$$\delta I_y = (p(x_i) + p(x_j)) \cdot JS(p(Y|x_i), p(Y|x_j) \tag{7}$$

where JS denotes the Jensen-Shannon divergence defined as:

$$JS(p(Y|x_i), p(Y|x_j)) = \pi_i \, D_{KL}[p(Y|x_i)||q(Y)] + \\ +\pi_j \, D_{KL}[p(Y|x_j||q(Y)] \quad (8)$$

$$\text{with } q(Y) = \pi_i \, p(Y|x_i) + \pi_j \, p(Y|x_j) \quad (9)$$

with $\pi_i = p(x_i)/(p(x_i) + p(x_j))$ and $\pi_j = p(x_j)/(p(x_i) + p(x_j))$. In case of discrete probabilities, this divergence (7) is straightforward to compute.

This algorithm produces a clustering that provides a good approximation to optimal IB solution. The information preserved in each step decreases monotonically.

Details about implementation of aIB algorithm can be found in [13] and will not be further discussed here.

The objective function 1 decreases monotonically with number of clusters . However, this does not give any further information on the optimal number of clusters. This will be addressed in the next section.

4 Model Selection

Model selection chooses the best model that represent a given data set. In case of parametric models, BIC [9] or modified BIC [3],[4] are very common choices.

In the case of aIB there is no parametric model that represent the data and BIC criterion cannot be applied. Several alternative solutions have been considered in literature. For instance the normalized mutual information $\frac{I(C,Y)}{I(X,Y)}$ gives useful information on the clustering quality. This quantity decreases rapidly when dissimilar clusters (most likely different speakers) are merged. Hence, we investigate a simple thresholding of $\frac{I(C,Y)}{I(X,Y)}$ as possible choice of the number of clusters.

Because of the information theoretic basis of the information bottleneck method, it is straightforward to apply the minimum description length (MDL) principle as proposed in [12]. The MDL principle states that the optimal model is the one that encodes the data and model with minimum code length. [7]. The MDL criterion is given by

$$\mathcal{F}_{MDL} = L(H) + L(D|H) \quad (10)$$

Where $L(H)$ is the code length to encode the hypothesis with a fixed length code and $L(D|H)$ is the code length required to encode the data given the hypothesis. In case of parametric models MDL reduces to the BIC where $L(D|H)$ is equivalent to likelihood of data and $L(H)$ is equivalent to the penalty term.

Let $N = |X|$ be the number of input samples, and $W = |C|$ the number of clusters. The number of bits required to code these samples with a fixed length code is $L(H) = N \log \frac{N}{W}$. The clustering itself can be coded with $L(D|H) = N[H(Y|C)+H(C)]$ bits. Since $H(Y|C)$ can be written as $H(Y)-I(C,Y)$ the model selection criterion is given by

$$\mathcal{F}_{MDL} = N[H(Y) - I(C,Y) + H(C)] + N \log \frac{N}{W} \quad (11)$$

Expression (11) provides the criterion according to which number of clusters (i.e. speakers) can be selected. Penalty term is analogous to the BIC penalty term and it penalizes codes that uses too many clusters.

5 APPLYING AIB TO SPEAKER CLUSTERING

The IB methods has been applied to clustering of different type of data like documents [13] or images [10]. In this paper we investigate the use of aIB to clustering of speech segments.

Consider acoustic features extracted from an audio file and a segmentation into regions containing only speech from a single speaker. In relation to the notation used in previous sections, those regions would represent the data set X.

Let us now consider a Gaussian Mixture Model
$f(x) = \sum_{j=1}^{M} w_j \mathcal{N}(x, \mu_j, C_j)$ where M is the number of components, w_j are weights, μ_j means and C_j covariance matrices. It is possible to project each speech frame in X onto the space of Gaussian components of the GMM. In relation to the notation used in previous sections, the space induced by GMM components would represent the relevance variable Y.

Computation of $p(y_i|x)$ is thus straightforward:

$$p(y_i|x) = \frac{w_i \mathcal{N}(x, \mu_i, C_i)}{\sum_{j=1}^{N} w_j \mathcal{N}(x, \mu_j, C_j)};\ i = 1, \ldots, N \tag{12}$$

The probability $p(y_i|x)$ estimates the relevance that the i^{th} component in the GMM has for speech frame x. If segment X is composed of several speech frames, distributions $p(y_i|x)$ can be averaged over the length of the segment. $p(y_i|x)$ is in this case a discrete distribution and aIB clustering can be easily applied.

aIB tries to find the optimal clustering C that preserves as much information as possible on the space induced by the components of the GMM. This is obtained by clustering together segments that have the smallest Jensen-Shannon distance i.e. segments that are the closest in the space of the GMM components.

Note that in the case of a parametric model, model parameters have to be estimated for each segment $\{x_i\}$, $\{x_j\}$ and for $\{x_i \bigcup x_j\}$ then penalized log-likelihood ratio is computed to decide whether segments should be clustered. This is repeated until a merging decision is found. On the other hand, in aIB, the sequence of merging is determined on the basis of similarity in between clusters estimated using Jensen-Shannon divergence. This similarity is not based on a model for each segment but on how close those segments are in the space of relevance variables i.e. in the space of components of a GMM. There is a clear advantage in terms of computational complexity because $p(y|x)$ are computed only once and there is no intermediate model estimation.

At this point a parallel can be drawn with document clustering which is one of the first applications of aIB. Let us denote with X a set of documents and with Y a vocabulary of words with associated conditional distributions $p(y|x)$ i.e. the probability of having word y in document x. Documents can be clustered together using aIB according to the fact that similar documents will have similar conditional probabilities of containing the same words. In a similar way we can cluster together speech segments if their conditional probabilities of containing the same components are close. In this case Gaussian components can be interpreted as vocabulary of "words" that compose each speech segment.

6 aIB BASED DIARIZATION SYSTEM

In this section we briefly describe the diarization system used for experiments in the paper. Schematically it can be summarized as follows:

1 Extract acoustic features 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.

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

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

6 aIB clustering + model selection i.e. inferring the number of speaker and assigning each element of X to a speaker.

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

Steps 1 and 2 are common to any diarization system. In step 3 the speech is uniformly segmented into chunks of fixed length. This step aims at obtaining an initial number of segments containing speech from a single speaker. In this work we used a uniform segmentation but other solutions may also be considered like using a speaker change detector or simple K-means algorithm. In this work we will limit our investigation to uniform segmentation.

In step 4 a Gaussian Mixture Model with shared covariance matrix is trained on the speech data. GMM components are used as relevance variables for clustering of speech chunks. Here we trained the GMM on data from the same file but it could actually be trained on a large independent data set i.e. like Universal Background Model (UBM) for speaker verification. As preliminary investigation we used a GMM estimated on data from the current file only.

In step 5 conditional distributions $p(Y|X)$ are estimated and in step 6 aIB clustering is performed followed by a model selection. This step provides the speaker clustering i.e. the deterministic map from the data set X to the inferred number of clusters C. The advantage of such clustering in a discrete probability space is that explicit model estimation is avoided with large gain in terms of computational resources.

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

In order to refine the initial segmentation, step 7 perform a set of Viterbi realignment on the data. 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 that were assigned from the aIB clustering. Then data are re-aligned using Viterbi algorithm. This step does not change the number of speakers and mapping from X to C but modifies boundaries that were obtained arbitrarily in the step 3. In experiments we tested the system with and without Viterbi realignment and observed an absolute improvement of 4.5% with Viterbi realignment.

Improvements in step 3 (for instance with a speaker change detector) would mitigate the need of step 7.

7 EXPERIMENTS AND RESULTS

We performed all the experiments on NIST RT06 evaluation data for "Meeting Recognition Diarization" task based on data from Multiple Distant Microphones (MDM) [2] and results are provided in terms of Diarization Error Rates (DER). DER is the sum of missed speech errors, false alarm speech error and speaker error (for details on DER see [1]). Speech/non-speech (spnsp) is the sum of missed speech and false alarm speech. System parameters are tuned on the development data.

Pre-processing of the data consists of Wiener filter denoising for individual channels followed by a beam-forming algorithm (delay and sum) as described in [6],[11]. This was performed using the *BeamformIt* toolkit [5]. 19 MFCC features are then extracted from the beam-formed signal.

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

We conducted the following set of experiments to compare the aIB based clustering with conventional HMM/GMM based clustering. In section 7.1 we present results of baseline HMM/GMM system, in section 7.2 we investigate impact of trade-off factor β performing model selection with an oracle and in section 7.3 we investigate two different model selection framework based on Normalized Mutual Information and on Minimum Description Length.

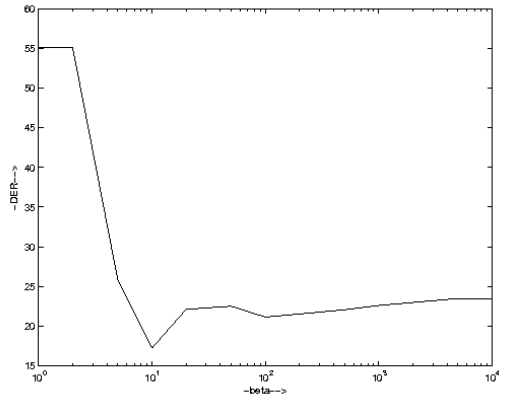

Figure 1: DER as a function of beta parameter for a single meeting recordings.

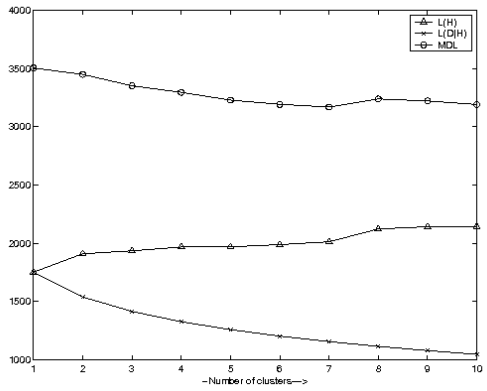

Figure 2: MDL model selection plotted as function of number of clusters for a single meeting recording.

7.1 Baseline system

The baseline system is based on 'bottom-up' clustering using HMM/GMM framework [3]. It is based on a modified version of the BIC criterion in which model complexity is kept constant across different mergings for avoiding fine tuning of BIC penalty term.

The clustering is obtained using an iterative algorithm based on segment merging and Viterbi re-alignment imposing a duration constraint of 2.5 seconds. This system has shown very competitive results in several NIST evaluation and will be used as baseline system for comparison purposes.

The results of the baseline system on RT06 eval data is listed in Table 1. The table lists missed speech, false alarm, speaker error and diarization error for all the meetings in the database.

File	Miss	FA	spnsp	spkr err	DER
ALL	6.50	0.10	6.60	18.90	25.54

Table 1: Results of the baseline system

7.2 aIB experiment 1: setting β

In this experiments we aim to study the impact of the trade-off factor β on final DER. We changed the value of β from 1 to 10^4 on a log scale and plotted DER obtained in the development data (figure 1). The optimal value of β obtained on development data is then applied to evaluation data. We manually chose trough an oracle the clustering that yields the lower DER in each meeting. No Viterbi re-alignment is performed.

We found that the optimal trade-off value of β is equal to 10. Results are shown in table 2.

	Miss	FA	spnsp	spkr err	DER
w/o Viterbi	6.50	0.10	6.60	22.20	28.78
with Viterbi	6.50	0.10	6.60	17.90	24.54

Table 2: Results of the oracle system (with and without Viterbi realignment)

Overall Speaker Error is 3.2% absolute worst then the baseline system. In table 2, we also report results after performing Viterbi re-alignment as described in section 6.

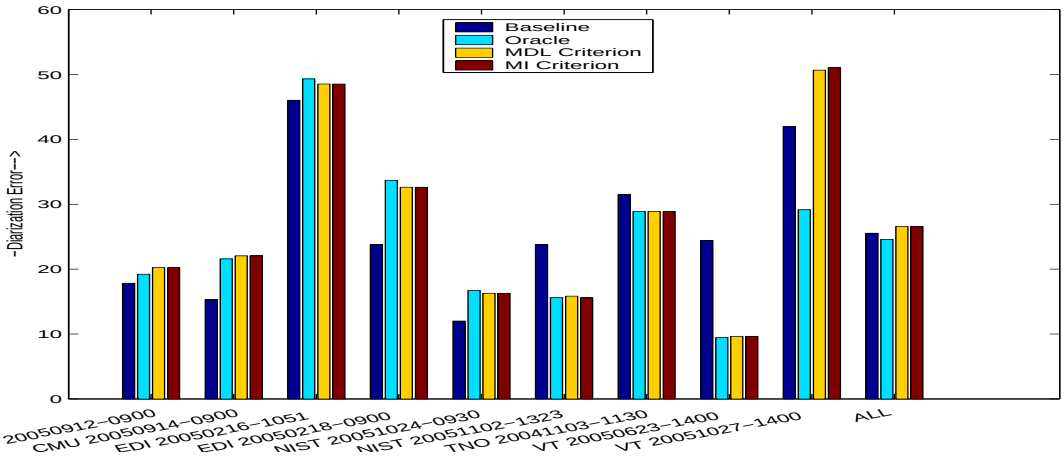

Figure 3: DER summary for all meetings: Baseline system, aIB + oracle, aIB + Normalized Mutual Information, aIB + Minimum Description Length.

After re-alignment the aIB based system improves of 4.5% absolute, over-performing the baseline system by 1% absolute. It is interesting to notice that while overall performances are similar for the two systems, results per meeting are very different (see figure 3 for results per meeting).

7.3 aIB experiment 2: model selection

In this section, we present experimental results with model selection as described in section 4. Two different model selection criteria –Normalized Mutual Information and Minimum Description Length– are investigated. Experimental framework is the same as in previous section but the number of clusters is chosen with one of the model selection criteria rather than the oracle.

Normalized Mutual Information decreases with the number of clusters. It can be observed that this quantity increases more rapidly in the beginning and then flatten out. A threshold tuned on the development data is used to determine the number of clusters. Optimal threshold value is fixed to be 0.25. Viterbi realignment is performed after model selection. Corresponding results are listed in Table 3

File	Miss	FA	spnsp	spkr err	DER
w/o Viterbi	6.50	0.10	6.60	24.70	31.36
w Viterbi	6.50	0.10	6.60	19.90	26.54

Table 3: Results of the system with model selection using normalized mutual information (with and without Viterbi decoding)

MDL criterion in equation 11 is also investigated as model selection criterion. Figure 2 illustrates the variation of code length of hypothesis $L(H)$, data $L(D|H)$ and the total MDL length $\mathcal{F}_{MDL}$. Note that the term $L(H)$ increases with the number of clusters, while $L(D|H)$ decreases. Results are listed in Table 4.

File	Miss	FA	spnsp	spkr err	DER
w/o Viterbi	6.50	0.10	6.60	25.00	31.61
w Viterbi	6.50	0.10	6.60	20.00	26.65

Table 4: Results of the system with model selection using MDL (with and without Viterbi decoding)

Both model selection criteria hold very similar performances but results show a 1% degradation respect to the oracle experiment. Figure 3 summarizes results for all meetings in the RT06 evaluation data in terms of baseline system, oracle experiment and model selection experiment. The model selection is able to select a model close to the best one obtained through oracle in 8 of the 9 meetings. Although overall performance for the aIB and HMM/GMM is the same, per meeting performance is significantly different from the state of the art system. In fact only in two of the evaluation meetings they have similar performance. Perhaps the two systems (GMM based and aIB based) show good complementarity properties.

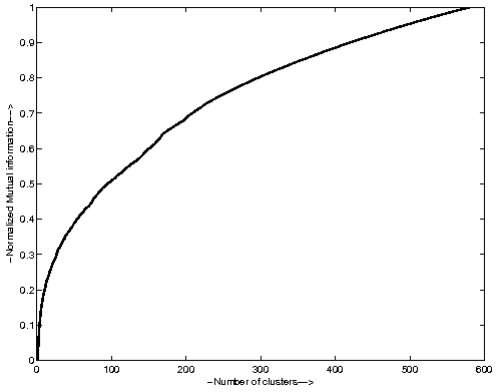

Figure 4: $\frac{I(\tilde{X},Y)}{I(X,Y)}$ Vs Number of clusters.

7.4 Computational time

In order to verify that the aIB clustering provides a significant reduction in terms of complexity, we report in table 5 computational time on RT06 evaluation data for aIB with and without Viterbi realignment and for conventional HMM/GMM system. Feature extraction and beam-forming are common to both methods and are not taken into account for measuring performances. aIB processing time includes computation of shared covariance matrix GMM. Processes are run on a Dual Core AMD Opteron(tm) Processor 2216 MHz machine.

	aIB w/o Viterbi	aIB with Viterbi	HMM/GMM
Time	74 min.	131 min.	424 min.

Table 5: Computational Time for RT06 evaluation data (in minutes) for aIB with and without Viterbi realignment and conventional HMM/GMM system.

aIB based clustering runs 6 times faster compared to the HMM/GMM system. If Viterbi realignment is performed the speed-up factor reduces to 3.

8 CONCLUSIONS AND FUTURE WORK

In this work, we proposed a new framework for speaker diarization based on agglomerative Information Bottleneck (aIB) method. aIB generates a clustering that minimize the loss in terms of mutual information between input variables and a given set of relevance variables. When the feature space X and relevance variable space Y are discrete and the probability distribution $p(Y|X)$ is available, the information loss can be easily computed (Section 3). In order to apply it to unsupervised clustering

of speakers, two model selection criteria are studied based on Normalized Mutual Information and Minimum Description Length.

We compare aIB based speaker diarization system with a baseline system based on agglomerative clustering with HMM/GMM. The main advantage of aIB clustering is that it does not build any parametric model (GMM) for deciding if two clusters should be merged. This avoids explicit GMM computation at each merging step, thus significantly reduces the complexity of the algorithm.

Experiments are performed on RT06 evaluation data and results are provided in terms of Diarization Error Rate. In the first set of experiments, model selection is performed using an oracle; in this case performances are better 1% absolute with respect to the baseline system. In the second set of experiments model selection is investigated using Normalized Mutual Information and Minimum Description Length criteria. In this case the system has a drop in performances by 2% w.r.t oracle model selection and by 1% w.r.t the baseline system. To summarize, the non-parametric agglomerative clustering (aIB) is found to achieve DER close to conventional HMM/GMM systems with reduced computation.

This preliminary work on information theory based clustering is based on a series of assumptions that will be further investigated in future works. For instance, we used an initial uniform segmentation in blocks of fixed length. This step can be improved through the use of a speaker change detection algorithm or an initial K-means clustering to obtain "pure" initial segments i.e. segments containing a single speaker. Improving initial segmentation would help to avoid the need of Viterbi re-alignment of data.

Furthermore, the relevance variables space is obtained from a GMM trained on the same data that is used for clustering; the use of a Universal Background Model (UBM) would increase the robustness to different amount of training data. Also algorithms based on sequential optimization rather than agglomerative greedy search has been proposed for optimizing the Information Bottleneck criterion (e.g. [14]) and could be worth exploring for speaker diarization task.

9 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 beam-forming toolkit. Authors also would like to thank Dr. John Dines for his help with the speech/non-speech segmentation

References

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

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

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

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

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

[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] A. Barron, J. Rissanen, and Yu B. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44:2743–2760, October 1998.

[8] Thomas Hain et. al. The ami meeting transcription system: Progress and performance. In *Proceedings of NIST RT'O6 Workshop*, 2006.

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

[10] 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.

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

[12] 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.

[13] 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.

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

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

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