IDIAP RESEARCH REPORT

COMBINATION OF AGGLOMERATIVE AND SEQUENTIAL CLUSTERING FOR SPEAKER DIARIZATION

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

IDIAP–RR 07-51

October 2007

to appear in
ICASSP 2008

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

IDIAP Research Report 07-51

COMBINATION OF AGGLOMERATIVE AND SEQUENTIAL CLUSTERING FOR SPEAKER DIARIZATION

Deepu Vijayasenan Fabio Valente Hervé Bourlard

October 2007

to appear in
ICASSP 2008

Abstract. This paper aims at investigating the use of sequential clustering for speaker diarization. Conventional diarization systems are based on parametric models and agglomerative clustering. In our previous work we proposed a non-parametric method based on the agglomerative Information Bottleneck for very fast diarization. Here we consider the combination of sequential and agglomerative clustering for avoiding local maxima of the objective function and for purification. Experiments are run on the RT06 eval data. Sequential Clustering with oracle model selection can reduce the speaker error by 10% w.r.t. agglomerative clustering. When the model selection is based on Normalized Mutual Information criterion, a relative improvement of 5% is obtained using a combination of agglomerative and sequential clustering.

1 Introduction

Speaker Diarization is the task of deciding *who spoke when* in an audio stream. It involves determining the number of speakers and identifying the speech segments corresponding to each speaker.

Conventional diarization systems are based on ergodic HMMs in which each state represents a speaker. Emission probabilities are Gaussian Mixture Models (GMM). The diarization algorithm is based on bottom-up agglomerative clustering of those initial segments [13]. Segments are merged according to some measure till a stopping criterion is met. Bayesian Information Criterion (BIC) [9], [13] and modified BIC [3],[4] are very common choices. Those systems are based on parametric models (GMMs) which assume availability of enough data at each cluster to estimate the model parameters.

In our previous work [15], we introduced a non-parametric diarization system based on the agglomerative information bottleneck (aIB) framework [11]. aIB is a clustering algorithm based on information theoretic framework. The system clusters segments based on their distance in a space of relevance variables. Furthermore it is not based on any explicit model estimation for speaker models. Tests on the RT06 eval data show that aIB based diarization provides performances similar to conventional HMM/GMM systems, with significant reduction in computational load.

However both HMM/GMM and aIB based systems use agglomerative clustering methods. Agglomerative clustering is a greedy procedure that takes the optimal merging decision at each step. This algorithm can easily get stuck in a local maxima and there is no guarantee that the optimal decision at each step will provide a globally optimal solution. This weakness is overcome in two different ways: repeating the agglomerative clustering with several initializations or improving the final partition with purification algorithms. The first approach attempts to find a reasonable initialization for diarization algorithms [6]. However this is computationally expensive. On the other hand, purification algorithms [16] try to identify wrongly assigned segments (by agglomerative method) using some confidence measures. Those segments are then re-assigned to another speaker.

In this paper we investigate the use of the sequential Information Bottleneck (sIB) [12] method for the diarization of meeting data. In contrast to the agglomerative method, it aims at finding the global maximum of the Information Bottleneck objective function. The sIB method operates on a fixed partition of K clusters in the data space. Given that the number of speakers is not known a priori, sIB cannot be directly applied. We propose a set of experiments that explore the combination of agglomerative and sequential clustering with model selection criteria. Furthermore we draw a parallel between sIB and speaker purification algorithms [16].

The paper is organized as follows: in section 2 we describe the information bottleneck principle, in section 3 we describe the agglomerative and sequential clustering algorithms and how they can be used in the diarization system, in section 4 we describe the diarization system used for experiments and in section 5 we compare agglomerative and sequential clustering on RT06 data.

2 INFORMATION BOTTLENECK PRINCIPLE

Let X denote a set of elements that we want to cluster into 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. 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. 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 IB method [14] 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,

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

where β is the trade-off parameter 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 written as:

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

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

Formal solution that maximizes the function (1) is given by an equation system that relates $p(c|x)$,$p(y|c)$ and $p(c)$ (for details see [11]). Two different algorithms have been proposed for maximization of function (1) based on agglomerative and sequential clustering and will be detailed in the next section.

3 AGGLOMERATIVE AND SEQUENTIAL IB

The agglomerative Information Bottleneck (aIB) [11] 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 the Jensen-Shannon divergence between $p(Y|x_i)$ and $p(Y|x_j)$:

$$\delta I_y = (p(x_i) + p(x_j)) \cdot JS(p(Y|x_i), p(Y|x_j)) \tag{4}$$

where JS denotes the Jensen-Shannon divergence defined as:

$$\begin{aligned} JS(p(Y|x_i), p(Y|x_j)) = \pi_i\, D_{KL}[p(Y|x_i)||q(Y)] + \\ + \pi_j\, D_{KL}[p(Y|x_j||q(Y)] \end{aligned} \tag{5}$$

$$\text{with } q(Y) = \pi_i\, p(Y|x_i) + \pi_j\, p(Y|x_j) \tag{6}$$

with $\pi_i = p(x_i)/(p(x_i) + p(x_j))$, $\pi_j = p(x_j)/(p(x_i) + p(x_j))$ and D_{KL} is the KL divergence. In case of discrete probabilities, JS divergence (4) is straightforward to compute.

This algorithm produces a clustering that provides a good approximation to the optimal IB solution. Details about implementation of the aIB algorithm can be found in [11] and will not be further discussed here. The objective function (1) decreases monotonically with the number of clusters. However, this does not give any further information on the optimal number of clusters which must be estimated with a model selection criterion. In [15] we considered two different model selection metrics: the Minimum Description Length (MDL) and a thresholded Normalized Mutual Information (NMI). NMI can be written as $\frac{I(C,Y)}{I(X,Y)}$. MDL for information bottleneck can be formalized as:

$$\mathcal{F}_{MDL} = N[H(Y) - I(C, Y) + H(C)] + N \log \frac{N}{W} \tag{7}$$

where $H(Y)$ entropy of Y, $H(C)$ entropy of C, $N = |X|$ is the number of input samples and $W = |C|$ is the number of clusters. Expression (7) provides the criterion according to which number of clusters (i.e., speakers) can be selected. The last term is the penalty term and is analogous to the BIC penalty term and it penalizes codes that use too many clusters. Thus the algorithm scores the quality of the clustering using NMI or MDL after each merging and selects the best one from all possible models. However for a given number of clusters W, aIB is not guaranteed to find the best partition because of the greedy nature of the search.

On the other hand, sequential Information Bottleneck [12] aims at finding the global maximum of the objective function. It starts with an initial partition of the space into W clusters $\{c_1, ..., c_W\}$. This partition can be random or can be obtained for instance with agglomerative clustering . The sIB method draws some element x out of its cluster c_{old} and represents it as a new singleton cluster. x is then merged into the cluster c_{new} such that $c_{new} = argmin_{c\epsilon C}\, d_F(x, c)$ where $d(.,.)$ is the Jensen-Shannon distance previously defined. It can be verified that if $c_{new} \neq c_{old}$ then $F(C_{new}) < F(C_{old})$ i.e., at each step either the objective function (1) improves or stays unchanged. This step is repeated several times until there is no change in the clustering assignment. To avoid local maxima, the procedure can be repeated with several random initializations.

sIB operates with a fixed number of clusters W. In case of diarization systems, this number is not known a priori and must be estimated from the data. We study three different ways of clustering based on combination of aIB, sIB and model selection:

- Method 1 (aIB): based on conventional aIB+model selection [15]. This method starts with the trivial partition of each element of X in a cluster and performs agglomerative IB clustering until all elements in the space are grouped into a single cluster. The best partition is then selected using a model selection criterion (MDL or NMI).
- Method 2 (sIB): based on sIB + model selection. This method starts with a random partition of the space into W clusters, where W is large enough. sIB is applied to find the optimal clustering into W classes. The number of initial clusters is then progressively reduced from W to 1 and sIB is performed for each of the initial number of clusters. Model selection is applied to select the best model. From a theoretical point of view, this method is optimal because it estimates the best clustering for each possible initial number of classes. However running sIB a number of times equal to W, can be a computationally demanding task.
- Method 3 (aIB+sIB): based on aIB+model selection followed by sIB. This methods uses aIB and model selection for finding the number of clusters and a given partition of the data into W clusters. Successively sIB is applied on the previous partition. Obviously in this case the quality of the sIB clustering will depend on the quality of aIB clustering. This method is in spirit close to the conventional purification algorithms in the sense that it tries to improve the agglomerative clustering once a partition is found.

4 SPEAKER DIARIZATION SYSTEM

In [15] we introduced an aIB based speaker diarization system. In this section we briefly summarize the same and present an extension using sIB clustering. Information Bottleneck methods cluster the input data $X = \{x_i\}$, w.r.t. a set of relevance variables $Y = \{y_i\}$ using the conditional probability distribution $p(y_i|x)$. x_i is defined as a speech segment. In order to estimate relevance variables a shared covariance matrix GMM is estimated from the data in the audio file. Each gaussian mixture component is considered as a relevance variable y_i and conditional probabilities $p(y_i|x)$ are estimated in a straightforward way using the Bayes rule.

aIB clustering is based on the Jensen-Shannon distance between two speech segments x_i, x_j in the space of gaussian mixture components. In contrast to the conventional HMM/GMM system, there is no explicit computation of speaker models thus resulting in a much faster diarization system with similar diarization error rates (for details see [15]).

In this work we extend the previous system based on aIB, with the use of sIB clustering. It can be summarized as follows:

1 Acoustic feature extraction from the audio file.

2 Speech/non-speech segmentation and rejection of 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 Clustering and model selection using one of the three methods described in section 3.

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

Step 7 performs a Viterbi realignment on the data given the segmentation obtained in the previous steps. This step does not change the number of speakers but modifies boundaries that were obtained arbitrarily with step 3.

Optionally, sIB can be applied after the Viterbi re-alignment as a purification algorithm. We will refer to this as purification sIB.

5 EXPERIMENTS AND RESULTS

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

Pre-processing of the data consists of a Wiener filter denoising for individual channels followed by a beam-forming algorithm (delay and sum) as described in [7],[10]. 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 force aligned by an ASR system. The same speech/non-speech segmentation is used across all experiments.

The baseline system is based on 'bottom-up' clustering using HMM/GMM framework [3]. It uses a modified version of the BIC criterion in which the model complexity is kept constant while merging to avoid fine tuning the BIC penalty term.

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.

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

Table 1: Results of the baseline system

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. 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 [15].

5.1 Information Bottleneck Experiments

We set the trade off factor β equal to 10 by tuning on the development data (for details about tuning see [15]). Normalized Mutual Information (NMI) threshold for model selection is tuned on development data as well. MDL does not require any tuning.

We propose a set of experiments that aim at investigating the use of sIB in the diarization task. Table 2 reports speaker error for aIB (method 1), sIB (method 2) and aIB+sIB (method 3), with and

	Method 1 (aIB)		Method 2 (sIB)		Method 3 (aIB+sIB)	
Model selection	w/o Viterbi	w Viterbi	w/o Viterbi	w Viterbi	w/o Viterbi	w Viterbi
Oracle	21.6	17.1	16.6	**15.2**	17.6	16.6
MDL	22.6	17.8	19.2	17.4	18.9	17.1
NMI	22.1	17.1	19.2	17.4	18.3	**16.6**

Table 2: Speaker Error Rate for RT06 evaluation data. Results are reported for the three proposed clustering methods based on aIB, sIB and aIB+sIB with and without Viterbi realignment. The same speech/non-speech segmentation is used for all methods and for the baseline system thus it is reported only in table 1. DER is sum of speech/non-speech error and speaker error.

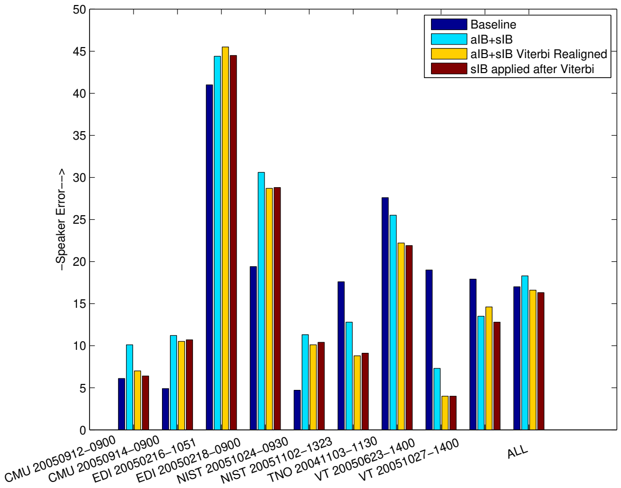

Figure 1: Speaker Error summary for all meetings: Baseline system, method 3 + NMI model selection without Viterbi, with Viterbi and with sIB purification.

without Viterbi re-alignment for MDL, NMI and Oracle model selection (DER is computed for all the different methods and the model with the lowest DER is chosen). The same speech/non-speech segmentation is used for all the methods and the baseline system, thus only speaker error is reported in table 2.

Without Viterbi re-alignment and with oracle model selection , sIB provides 4% absolute improvement compared to aIB while aIB+sIB (method 3) provides 3% absolute improvement. It is perhaps interesting to note that in case of method 2 with oracle, the sIB system is already outperforming the baseline system without using HMM/GMM.

When model selection is performed using MDL or NMI there is degradation of roughly 1% for method 1 and 3 and of 2.5% for method 2. We can notice that model selection criteria are more effective when used in conjunction with the agglomerative clustering. In this case the lowest speaker error is obtained using method 3 which combines aIB and sIB.

Viterbi re-alignment further improves the results; however the effect of the sIB is less significant. Also in the case of Viterbi re-alignment, sIB provides the best results with oracle model selection. In case of NMI and MDL model selection, the aIB+sIB method performs the best. The Viterbi re-alignment has a double effect of improving the segmentation and smoothing out the differences obtained using the method 3.

As stated before sIB can also be applied after Viterbi re-alignment in the same fashion as most purification algorithms work. Results are reported in table 3. In this case only small improvements are observed.

To summarize, when oracle model selection is used the sequential clustering (method 2) outperforms methods that uses agglomerative clustering (1 and 3) and improves the speaker error by 10% relative w.r.t. the baseline system (i.e., from 17.00% to 15.10%) . On the other hand when model selection (NMI) is used, method 3 (that combines aIB and sIB) achieves the lowest speaker error, improving the baseline system by 5% relative (i.e., from 17.00% to 16.30%). Out of the two model selection criteria, NMI is found being more effective compared to MDL. Figure 1 summarizes results for method 3 with NMI model selection for all the RT06 eval meetings across different steps i.e., with and without Viterbi re-alignment and with sIB purification. Improvement in overall speaker error is evaluated after each step.

6 Conclusion

In this work we compare the use of agglomerative and sequential clustering for speaker diarization based on the information bottleneck principle. We propose three different systems based on agglomerative IB (aIB), sequential IB (sIB), and aIB followed by sIB. Model selection is addressed using Minimum Description Length (MDL), Normalized Mutual Information (NMI) and the oracle that select the best model. Results are obtained on RT06 evaluation data.

With oracle model selection, sIB provides consistent improvement of 4% absolute w.r.t. aIB and outperforms HMM/GMM system by 0.4% without the use of re-alignment. Viterbi re-alignment further reduces the overall speaker error. Sequential clustering applied after Viterbi further purifies the obtained partitions. With this second level of purification, method 2 with oracle model selection achieves a speaker error rate of 15.2% that is better than the baseline system by 10% relative. However, with NMI model selection , method 3 gives the lowest error rate of 16.3% that is a 5% relative improvement over the baseline.

It is observed that model selection is not very effective with the partitions issued by the sequential clustering. Method 3 (where model selection is done after aIB and sIB is then applied) performs the best both with NMI and MDL. Further investigations have to be carried out as to why model selection performance is better while using aIB or aIB+sIB instead of sIB.

7 ACKNOWLEDGEMENTS

This work was supported by the European Union under the integrated projects AMIDA, Augmented Multi-party Interaction with Distance Access, contract number IST-033812, as well as KERSEQ project under the Indo Swiss Joint Research Program (ISJRP) financed by the Swiss National Science Foundation. This project is pursued in collaboration with EPFL under contract number IT02. The authors gratefully thank the EU and Switzerland for their financial support, and all project partners for a fruitful collaboration.

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

References

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

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

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

Model selection	Method 1	Method 2	Method 3
Oracle	17.1	**15.1**	16.3
MDL	17.2	17.6	16.9
NMI	16.7	17.2	**16.3**

Table 3: Results for purification sIB applied after Viterbi re-alignment for the 3 different systems: Method 1 (aIB), Method 2 (sIB) and Method 3 (aIB+sIB).

[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. Hernando. Friends and enemies: A novel initialization for speaker diarization. In *Proceedings of ICSLP-Interspeech*, 2006.

[7] 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.

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

[11] 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.

[12] 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.

[13] 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.

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

[15] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard. Agglomerative information bottleneck for speaker diarization of meetings data. In *IDIAP Technical Report IDIAP-RR-07-31 ftp://ftp.idiap.ch/pub/reports/2007/vijayasenan-idiap-rr-07-31.ps.gz to be published in ASRU*, 2007.

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