IDIAP RESEARCH REPORT

ADVANCES IN FAST MULTISTREAM DIARIZATION BASED ON THE INFORMATION BOTTLENECK FRAMEWORK

Deepu Vijayasenan Fabio Valente Hervé Bourlard

Idiap-RR-23-2010

JULY 2010

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Advances in Fast Multistream Diarization based on the Information Bottleneck Framework

Deepu Vijayasenan[1,2], *Fabio Valente* [1], *Hervé Bourlard* [1,2]

[1] Idiap Research Institute, Martigny, CH
[2] Ecole Polytechnique Fédérale de Lausanne, Lausanne, CH
dvijaya@idiap.ch, fvalente@idiap.ch, bourlard@idiap.ch

Abstract

Multistream diarization is an effective way to improve the diarization performance, MFCC and Time Delay Of Arrivals (TDOA) being the most commonly used features. This paper extends our previous work on information bottleneck diarization aiming to include large number of features besides MFCC and TDOA while keeping computational costs low. At first HMM/GMM and IB systems are compared in case of two and four feature streams and analysis of errors is performed. Results on a dataset of 17 meetings show that, in spite of comparable oracle performances, the IB system is more robust to feature weight variations. Then a sequential optimization is introduced that further improves the speaker error by $5-8\%$ relative. In the last part, computational issues are discussed. The proposed approach is significantly faster and its complexity marginally grows with the number of feature streams running in 0.75 real time even with four streams achieving a speaker error equal to 6%.

Index Terms: Speaker diarization, Information bottleneck, Multistream diarization, Sequential IB.

1. Introduction

Including multiple feature streams is an effective method to improve the performance of speaker diarization systems recorded in meeting rooms using multiple distant microphones. The most common feature combination is based on spectral features, e.g., MFCC and Time delay of arrivals features (TDOA). Other studies have proposed the combination of MFCC with prosodic or long term features, i.e. extracted from a long time span of the signal [1]. However, the speaker error reduction with multiple feature streams happens at the cost of an increased computational complexity. In our previous work [2], it was shown that up to four feature streams (MFCC, TDOA, Modulation Spectrum features and Frequency Domain Linear Prediction[3]) can simultaneously be integrated in a non-parametric diarization system for further reduction in the speaker error. This paper advances the previous work by three contributions. The IB system is compared side by side with a conventional HMM/GMM system both using two and four features streams. The study aims to analyze the robustness of the two systems to the stream weights. While a greedy agglomerative method was used in previous works, this paper introduces a sequential optimization method to find the global minimum of the objective function. The sequential method (sIB) acts as purification step to improve the partition produced by the agglomerative clustering. The application of sIB to multistream system is investigated. The last contribution of the paper is the study of computational costs of the proposed non-parametric system versus a parametric HMM/GMM system when multiple feature streams are used.

The remainder of the paper is organized as follows: Section 2 and Section 3 describe the HMM/GMM system and the agglomerative IB system. Section 4 describes the sequential optimization method. Experiments both on two and four feature streams are presented in Section 5 and complexity analysis is presented in Section 6. The paper is concluded in Section 7.

2. GMM based diarization

Conventional speaker diarization systems are based on HMM/GMM models in which each speaker is represented by an HMM state with GMM emission probability [4]. The diarization starts with a uniform linear segmentation of the input into a large number of clusters (speakers). Successively at each step a cluster pair is merged based on a distance measure like the BIC or its modified version [4]. The merging stops when all the BIC values are less than zero. After each merge, a realignment of speaker boundaries is performed with the estimated speaker models. Whenever multiple feature streams $\{x_t^i\}, i = 1, \ldots, M$ are available, the system can be extended by considering a separate GMM model for each stream. Let $b_c^i(x_t^i)$ be the GMM model of cluster c corresponding to the feature stream x_t^i. The BIC criterion is extended using a combined likelihood $l_c(x_t)$ computed as a weighted linear combination of individual likelihoods:

$$l_c(x_t) = \sum_i P_i \log \left[b_c^i(x_t^i) \right] \quad (1)$$

where P_i represents the weight of the feature stream x_t^i and is estimated on a development dataset. The most common features used are MFCC and TDOA [5] but also other feature sets have been considered recently [1]. Details on the initialization and the number of gaussian components per feature stream can be found in [6],[5].

3. IB based Speaker Diarization

This section briefly summarizes the IB speaker diarization system that operates in a space of relevance variables proposed in [7]. The Information Bottleneck is a distributional clustering technique introduced in [8]. Consider a set of input variables X. The Information Bottleneck principle depends on a relevance variables' set Y that carries important information about the problem. According to IB principle, any clustering C should be compact with respect to the input representation (minimum $I(X, C)$) and preserve as much mutual information as possible about relevance variables Y (maximum $I(C, Y)$). This corre-

sponds to the maximization of:

$$\mathcal{F} = I(C,Y) - \frac{1}{\beta} I(X,C) \quad (2)$$

where β is a Lagrange multiplier. The IB criterion is optimized w.r.t. the stochastic mapping $p(c|x)$ using iterative optimization techniques. The agglomerative Information Bottleneck (aIB) clustering is a greedy way of optimizing the IB objective function [9]. The algorithm is initialized with each input element $x \in X$ as a separate cluster. At each step, two clusters are merged such that the reduction in mutual information w.r.t relevance variables is minimum. It can be proved that the loss in mutual information in merging any two clusters c_1 and c_2 is given in terms of a Jensen-Shannon divergence that can directly be computed from the distribution $p(y|x)$ as:

$$\Delta\mathcal{F}(c_1,c_2) = [p(c_1)+p(c_2)]JS[p(y|c_1),p(y|c_2)] \quad (3)$$

The Jensen-Shannon divergence $JS[p(y|c_1),p(y|c_2)]$ is given by:

$$\pi_1 D_{kl}\left[p(y|c_1)||q(y)\right] + \pi_2 D_{kl}\left[p(y|c_2)||q(y)\right] \quad (4)$$

where $\pi_j = \frac{p(c_j)}{p(c_1)+p(c_2)}$, $q(y)$ represents the distribution of relevance variables after the cluster merge and D_{kl} denotes the Kullback-Leibler divergence between two distributions. The number of clusters is determined by using a threshold on the Normalized Mutual Information given by $\frac{I(C,Y)}{I(X,Y)}$.

In order to apply this method to speaker diarization, the set of relevance variables $Y = \{y_n\}$ is defined as the components of a background GMM ($\mathcal{M}$) trained on the entire audio recording [7]. The input to the clustering algorithm is uniformly segmented speech segments x_t. The posterior probability $p(y_n|x_t)$ is computed using Bayes' rule. The speech segments with the smallest distance (the Jensen-Shannon divergence) are then iteratively merged until the model selection criterion is satisfied.

Whenever multiple features are available, the combination is performed in the space of relevance variables y [10]. Separate GMMs with the same number of components are trained for each feature stream. The individual components are kept aligned. i.e, the same component of two different GMMs are estimated using the features with same time indices. In other words, there is a one-to-one correspondence between the GMM components. Let $\{\mathcal{M}_i\}$ be the background model for the feature stream x^i. The combined distribution $p(y|x)$ is then estimated as:

$$p(y|x) = \sum_i p(y|x^i,\mathcal{M}_i)P_i \quad (5)$$

where P_i corresponds to the weights of i^{th} feature stream ($\sum P_i = 1$). This corresponds to averaging the different $p(y|x^i,\mathcal{M}_i)$ obtained with GMMs trained on different feature streams.

After clustering, the speaker boundaries are realigned. Instead of using HMM/GMMs, the realignment is performed in the space of relevance variables $p(y|x)$ using an HMM-KL divergence (Kullback-Leibler) based system described in [10].

The entire diarization algorithm including clustering, feature combination and realignment depends only on the relevance variable distribution $p(y|x)$.

4. Sequential IB

Being a greedy algorithm, the aIB may not converge to the global optimum of the objective function. A sequential optimization referred as sequential Information Bottleneck(sIB) was proposed in [11] and aims at finding the global maximum of the objective function.

Consider an initial partition of the data X into K clusters $\{c_1,\ldots,c_K\}$. An element $x \in X$ is drawn at random out of its cluster c_{old} and is represented as a singleton cluster. This singleton cluster is then merged into a new cluster c_{new} according to:

$$c_{new} = \arg\min_c \Delta\mathcal{F}(x,c) \quad (6)$$

where $\Delta\mathcal{F}(x,c)$ is the loss in IB function in merging the singleton cluster x with any cluster c. This information loss is again represented in terms of a Jensen-Shannon divergence i.e. Eqn. 3. It can be shown that if $c_{new} \neq c_{old}$ the IB objective function improves [11]. Thus in each step the IB functional improves or stays the same. This reassignment is repeated several times until there is no change in the clustering assignments.

We propose here the use of this sequential optimization on the agglomerative clustering partition. In this scenario, sIB refines the clusters by reassigning the elements similar to the cluster purification algorithms [12]. In case of multistream diarization, the distribution $p(y|x)$ calculated by (5) can be employed. As the aIB, the sIB algorithm also requires only the distribution $p(y|x)$ as the input.

5. Experiments

The experiments are conducted on 17 meeting recordings from five different meeting rooms (CMU,EDI,NIST,TNO,VT) corresponding to data collected for the NIST RT06/RT07 evaluations [13]. The amount of speech data is more than twice more as compared to our previous experiments [2]. At first multiple channels are beamformed using the *BeamformIt* toolkit. MFCC and TDOA features are then extracted from the beamformed output (details about the front-end are available in [5]). Two additional sets of features – filtered trajectories of critical band energies, i.e., Modulation Spectrum (MS), and FDLP features[3] are also extracted from long temporal windows, differing from both location features like TDOA or spectral features like MFCC. The current work studies speaker diarization based on the combination of two features (MFCC and TDOA) and four features (MFCC, TDOA, FDLP and MS) as well as their complementarity. The MFCC, and FDLP features have a dimensionality of 19 while MS features are 26 dimensional and the dimensionality of TDOA features changes with the number of microphones in the array.

A critical part of multi-stream methods consists of determining the weights of different feature sets. In this work, these weights are estimated from a development dataset composed of 12 recordings across 6 meetings rooms. The weights that minimize a smoothed version of speaker error [2] are selected in order to avoid local minima. The system performance is evaluated using Diarization Error Rate (DER) that is the sum of speech/non-speech segmentation and speaker errors. Since we use the same speech non-speech segmentation across all the experiments only speaker error is reported for the purpose of comparison.

Experiments report the performance obtained by both optimal weights and estimated weights from development data. Optimal weights represent the best performance possible with the feature combination. They are obtained by varying the feature weights P_i from 0 to 1 under the constraint $\sum_i P_i = 1$ and choosing the weights that corresponds to the minimum speaker error. Results for the HMM/GMM, the agglomerative and sequential IB systems are reported in the following.

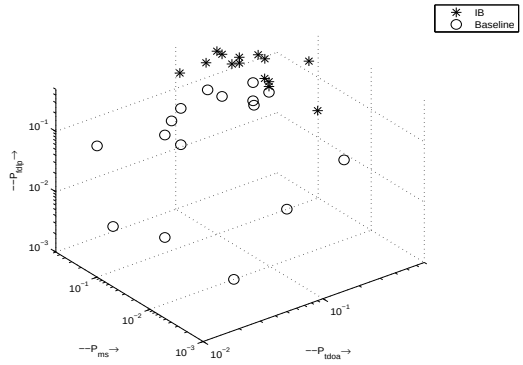

Figure 1: Optimal weights for IB and baseline systems obtained with an oracle experiment for the four stream system.

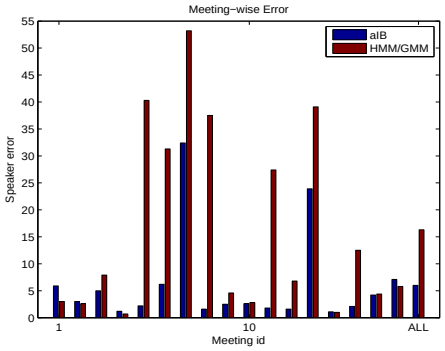

Figure 2: Meeting-wise Speaker Error in case of HMM/GMM and aIB with weights estimated from development data.

Table 1: Baseline and aIB speaker errors corresponding to two and four feature combination systems (optimal and estimated weights).

	HMM/GMM		aIB	
	2 feat	4 feats	2 feats	4 feats
optimal wts	5.1	2.6	5.8	2.8
estimated wts	12.4	16.3	11.6	6.3

5.1. HMM/GMM system

Table 1(second column) reports the results corresponding to the MFCC and TDOA combination with estimated weights and oracle weights. Weights estimated on development data are $(P_{mfcc}, P_{tdoa}) = (0.9, 0.1)$. The speaker error obtained with estimated weights is more than 7% worse than performance with optimal weights. Let us now consider the combination of four features (MFCC, TDOA, MS and FDLP). The results are presented in Table 1 (third column). The estimated weights from the development data are $(P_{mfcc}, P_{tdoa}, P_{ms}, P_{fdlp}) = (0.69, 0.20, 0.01, 0.10)$. In spite of the improvement in the optimal performance, the speaker error of the baseline system increases when weights are estimated from development data.

5.2. Agglomerative IB system

Let us now consider the performance of the agglomerative IB system reported in Table 1 in case of MFCC and TDOA features (fourth column) and four features (fifth column). The optimal performances of the IB system are similar to the optimal performance of the HMM/GMM showing that in case of oracle stream weighting the two systems are equivalent. However in case of estimated weights it consistently outperforms the baseline.

The selected weights in case of two $[(P_{mfcc}, P_{tdoa}) = (0.7, 0.3)]$ and four feature streams $[(P_{mfcc}, P_{tdoa}, P_{ms}, P_{fdlp}) = (0.5, 0.20, 0.05, 0.25)]$ are quite different from the GMM scenario. The speaker error decreases in case of four features largely outperforming the baseline system.

To study the system performance in detail, let us consider the meeting-wise optimal weights depicted in Figure 1 in case of four streams. Optimal weights for the HMM/GMM system span a wider range compared to those of the aIB. A possible reason for this could be the variable dimension of the TDOA features which affects the order of magnitude of the log-likelihoods of Eqn. (1). In case of aIB, the combination is performed using probabilities (Eqn. 5) rather than log-likelihoods and optimal weights are observed to be in the similar magnitude range. Figure 2 compares the meeting-wise speaker errors; the baseline system performs considerably worse in case of meetings with higher number of microphones while it has comparable performances on the remaining. In summary, while the two systems have comparable oracle performances, the aIB seems considerably more robust when weights are obtained from a development data set.

To further analyze the two systems, we investigate the variance of the speaker error on the development data in a ± 0.05 neighborhood of the estimated weights . The variance is equal to 2.0 in the HMM/GMM case and 0.68 in case of aIB. This implies that the second one is less sensitive to weights than the first.

5.3. Sequential IB

The sIB algorithm is performed to further improve the diarization output as described in Section 4. Table 2 reports the results in case of two and four feature streams. The sequential framework improves the performance by 8% relative in the first case (from 11.6% to 10.7%) and by 5% relative in the second case (from 6.3% to 6.0%). The improvement obtained by the sequential optimization decreases while the number of streams increases and the speaker error becomes very low.

	aIB	aIB + sIB
2 feats	11.6	10.7 (+8%)
4 feats	6.3	6.0 (+5%)

Table 2: aIB and sIB performance in case of two and four feature streams.

6. Complexity

This section investigates the computational complexity of the HMM/GMM and the IB diarization systems. Both systems use an agglomerative clustering and this requires the estimation of the distance between every pair of clusters, i.e., $\frac{1}{2}k(k-1)$ distance calculations where k is the number of clusters. In case of HMM/GMM, the distance is represented by the BIC distance. Its calculation involves the estimation of a new GMM model using the Expectation-Maximization algorithm. Whenever multiple streams are available, a GMM must be estimated for each of

Table 3: Complexity analysis - Real time factors:
(a)algorithm time used by different steps in the IB system

	aIB			aIB+sIB		
	estimate $p(y\|x)$	IB clstrng	KL realgn	estimate $p(y\|x)$	IB clstrng	KL realgn
2 feat	0.24	0.08	0.09	0.25	0.10	0.09
4 feat	0.52	0.09	0.11	0.52	0.10	0.11

(b) comparison with baseline

	Baseline	aIB	aIB+sIB
2 feats	3.8	0.41	0.43
4 feats	11.3	0.72	0.75

them (see Eqn. 1) thus increasing the computational complexity.

In contrast to this, the IB system estimates a background GMM for each feature only once (before the clustering) and the combination happens in the space of distributions $p(y|x)$ (Eqn. 5). The distance measure (the Jensen-Shannon divergence of Eqn. 3) is obtained in close form and does not depend on the number of features streams. This happens since the dimension of the relevance variables Y depends only on the number of components in the mixture model. Thus the clustering and the realignment complexities remain the same. The extra cost comes only in the estimation of distribution $p(y|x)$.

The algorithms are benchmarked on a normal desktop Machine (AMD Athlon^{TM} 64 X2 Dual Core Processor 2.6GHz, 2GB RAM). The run-time of the algorithms are averaged across multiple iterations. Table 3(a) reports the real time factors taken by various steps in the IB diarization for two and four feature streams. The clustering and the realignment complexities remain almost constant with the addition of new features. The largest part of the computational time is spent in the distribution estimation step – roughly 60% in case of two stream combination and 70% in case of four streams. In both cases the additional complexity introduced by sIB is minimal (12% of the clustering time). Table 3(b) compares the real time factors for IB and baseline systems.

The IB diarization is 8 times faster than the HMM/GMM system in the two stream case and 14 times faster in the four stream case. It can also be noticed that the introduction of two additional features increases the computing time by a factor of 3 in the HMM/GMM system while this factor is only 1.7 in the IB system.

7. Conclusions

Speaker diarization based on combination of multiple streams has been an active field during last years. In the previous work [2] we have shown that up to four feature streams can be simultaneously integrated in a non-parametric diarization system for further reducing the speaker error. To our best knowledge this was the first successful attempt of including other features together with MFCC and TDOA.

This work aims at comparing a conventional HMM/GMM diarization system with a system based on the Information Bottleneck principle. While the first performs the combination by averaging log-likelihoods the second one operates in a space of relevance variables and avoids any log-likelihood combination. The investigation is carried on 17 meeting recordings from five meeting rooms in case of two features (MFCC and TDOA) and with four features (MFCC, TDOA, MS and FDLP). Results reveal that the two systems have comparable oracle performance (obtained manually choosing the optimal weights) in both cases.

Whenever weights are estimated from the development data, the baseline performance degrades considerably in case of four features. Comparatively the IB system performance degrades only by 3.5%(Table 1) achieving a speaker error equal to 6.3%. Analysis of sensitivity to the weights shows that the IB combination scheme is more robust to variations in feature weights as revealed by the speaker error variance.

The paper also proposes and investigates a sequential optimization method for refining the partition obtained by the agglomerative clustering. In contrast to the greedy aIB algorithm that might converge to a local minimum, sIB tries to find the global optimum. Experiments reveal that the algorithm improves the performance by 8% relative in the two stream case and by 5% relative in case of four streams. This shows that purification methods are effective even at very low speaker errors.

In addition, the analysis of the algorithm complexities shows that the IB algorithms are much faster than the baseline system. The algorithms perform in realtime, the majority of the running time being spent by the estimation of distributions $p(y|x)$. The clustering and realignment algorithm complexities remain almost same in spite of increase in number of features. The sequential optimization only marginally increases the computational time. Remarkably, even when four feature streams are used, the system runs in 0.75 times real-time achieving a speaker error of 6.0%.

In summary, results show that the proposed system provides a very robust way of integrating multiple features with a limited increase in the computational complexity.

8. Acknowledgements

This work is supported by the Swiss National Science Foundation under the NCCR on Interactive Multimodal Information Management and the MULTI grants.

9. References

[1] O. Vinyals, G. Friedland, "Modulation spectrogram features for speaker diarization," in *Proceedings of Interspeech*, 2008.

[2] D. Vijayasenan, F. Valente, and H. Bourlard, "Multistream Speaker Diarization beyond Two Acoustic Feature Streams," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4950–4953.

[3] M. Athineos and D. Ellis, "Frequency-domain linear prediction for temporal features," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU '03.*, 2003.

[4] J. Ajmera, "Robust audio segmentation," Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne (EPFL), 2004.

[5] X. Anguera, "Robust speaker diarization for meetings," Ph.D. dissertation, Universitat Politecnica de Catalunya, 2006.

[6] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *IEEE Automatic Speech Recognition Understanding Workshop*, 2003, pp. 411–416.

[7] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 7, pp. 1382 – 1393, 2009.

[8] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *NEC Research Institute TR*, 1998.

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

[10] D. Vijayasenan, F. Valente, and H. Bourlard, "KL realignment for speaker diarization with multiple feature streams," in *10th Annual Conference of the International Speech Communication Association*, 2009.

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

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

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