Information Bottleneck Features for HMM/GMM Speaker Diarization of Meetings Recordings

Sree Harsha Yella[1,2], *Fabio Valente* [1]

[1] Idiap Research Institute, CH-1920 Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
sree.yella@idiap.ch, fabio.valente@idiap.ch

Abstract

Improved diarization results can be obtained through combination of multiple systems. Several combination techniques have been proposed based on output voting, initialization and also integrated approaches. This paper proposes and investigates a novel approach to combine diarization systems through the use of *features*. A first diarization system, based on the Information Bottleneck, is used to generate a set of features that contain information relevant to the clustering. Those features are later used in conjunction with conventional MFCC in a second diarization system. This method is inspired from the TANDEM framework in ASR. While being fully integrated, the approach does not need modifications to any of the two systems in order to integrate the information. Experiments on 24 recordings from the NIST RT06/RT07/RT09 evaluations collected in five meeting rooms reveal that when the IB features are used together with MFCC, the total speaker error is reduced from 12% to 9.7%, i.e., by approximatively 19% relative.

Index Terms: Speaker diarization, Meetings, Information Bottleneck, System Combination, TANDEM features.

1. Introduction and Motivation

Speaker diarization deals with the task of identifying "who spoke when" in a given multi-party speech recording. The task is unsupervised as there is no knowledge of number of speakers in the recording. Several methods have been proposed in the literature to solve this problem [1], however, the most common approaches are based on HMM/GMM modeling which achieve state-of-the-art performance on several types of data [2]. Recently a non-parametric method based on the Information Bottleneck (IB) framework has been proposed in [3]. The system, based on a completely different objective function, provides comparable results to state-of-the-art HMM/GMM diarization with a significant speed-up. The current work will investigate if and how these two approaches can be combined.

A number of studies on Broadcast data have discussed the combination of speaker diarization outputs from different systems to improve results. The simplest combination consists of voting schemes [4] between outputs of multiple systems. Also, a system can be initialized with the output of another one, like in case of bottom-up and top-down diarization as proposed in [5]. Finally integrated approaches [6], i.e., systems that integrate two different diarization methods into a single one, have been considered for broadcast data. Recently they have also been revisited in the context of meeting recordings [7]. While combinations are able to outperform the individual diarization systems, each combination technique has advantages and pitfalls; in particular the voting scheme performs only late combination, i.e. at the output level, the initialization approaches only benefit from a different starting point and the integrated approaches require modifications to all parts of the systems.

This paper proposes and investigates a novel approach of combination through the use of *features*. A first diarization system is used to generate a set of features that contain information relevant to the clustering. Those are later used in conjunction with conventional spectral features in a second diarization system. The rationale behind this is that the new feature set will complement the second system at each step with the information provided from the output of the first. The approach does not need modifications to any of the two systems in order to integrate the information. This idea is largely inspired from the TANDEM framework used in Automatic Speech Recognition (ASR) [8]. TANDEM aims at using probabilistic output of a Multi Layer Perceptron that estimates phoneme posterior probabilities, as features to a conventional HMM/GMM system. Given an input speech frame X and a set of phonetic targets Y, the MLP estimates the posterior probabilities $p(Y|X)$. After that, $p(Y|X)$ are first gaussianized using a logarithm and then de-correlated with a PCA transform followed by a dimensionality reduction. Those are referred as TANDEM features. After concatenation with MFCC, they are used to train a standard HMM/GMM system. TANDEM features are able to reduce the Word Error Rate by $10-15\%$ relative (see [9] for a review of tasks and improvements) thus complementing well the standard spectral features. However, contrary to ASR, speaker diarization is an unsupervised task thus there is no direct equivalent to the phoneme posterior probabilities $p(Y|X)$.

This work proposes to generate TANDEM-like features using the probabilistic output of the Information Bottleneck system described in [3]. The IB diarization is based on the use of a set of relevance variables Y on which speech segments X are projected. Its output produces an assignment of each speech segment X to a cluster C, i.e., $p(C|X)$ as well as the probability of the relevance variables Y per each cluster C, i.e., $p(Y|C)$. The estimates $p(C|X)$ and $p(Y|C)$ will be used to generate a feature set representative of the clustering and to be integrated into HMM/GMM system. The remainder of the paper is organized as follows, Section 2 presents briefly the HMM/GMM diarization system, Section 3 describes the IB system, Section 4 introduces the TANDEM-IB features. Experiments are then presented in Section 5 and the paper is concluded in Section 6.

2. HMM/GMM Speaker Diarization

Conventional diarization systems are based on agglomerative clustering framework using HMM/GMM where each speaker is modeled as a HMM state and each state distribution is modelled using a GMM. The system discussed here achieved state-of-

the-art performance in several NIST evaluations [10]. It is initialized by uniformly segmenting a given audio recording into segments treated as initial clusters (speakers). Their number is much higher than the actual number of speakers in the recording. Then at each iteration, the closest clusters obtained using distance measures such as BIC or modified BIC are merged. The process continues until cluster pairs are found suitable for merging, i.e., until a stopping criterion is met. After each merge speaker boundaries are realigned based on the estimated speaker models using a Viterbi decoder. The emission probability distribution b_{c_i}, corresponding to speaker cluster c_i is modeled as a GMM:

$$\log b_{c_i}(s_t) = \log \sum_r w^r_{c_i} \mathcal{N}(s_t, \mu^r_{c_i}, \Sigma^r_{c_i}) \quad (1)$$

where s_t is input feature, $\mathcal{N}(.)$ is Gaussian pdf and $w^r_{c_i}$, $\mu^r_{c_i}$, $\Sigma^r_{c_i}$ are the weights, means and covariance matrices of r^{th} mixture Gaussian of cluster c_i. The modified BIC criterion (see [11]) for a pair of clusters c_i and c_j with respective GMM models $b_{c_i}(.)$ and $b_{c_j}(.)$ is given by

$$\sum_{s_t \in c_i \cup c_j} \log b_{c_{i+j}}(s_t) - \sum_{s_t \in c_j} \log b_{c_j}(s_t) - \sum_{s_t \in c_i} \log b_{c_i}(s_t) \quad (2)$$

Where $b_{c_{i+j}}(.)$ represents the GMM model estimated from combined data of clusters c_i and c_j. The number of Gaussian components in the model $b_{c_{i+j}}$ is equal to the sum of the Gaussian components in b_{c_i} and b_{c_j}.

3. IB based Speaker Diarization

This section briefly summarizes the Information Bottleneck (IB) speaker diarization system proposed in [3]. The IB is a distributional clustering technique introduced in [12]. Consider a set of input variables X to be clustered into C clusters. 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 corresponds to the maximization of:

$$\mathcal{F} = I(C,Y) - \frac{1}{\beta} I(X,C) \quad (3)$$

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 [13]. 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 $\Delta\mathcal{F}$ by merging any two clusters c_1 and c_2 is given in terms of a Jensen-Shannon divergence that can be directly computed from the distribution $p(Y|x)$ as:

$$\Delta\mathcal{F}(c_1, c_2) = [p(c_1) + p(c_2)] JS[p(Y|c_1), p(Y|c_2)] \quad (4)$$

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

$$\pi_1 D_{kl}\left[p(Y|c_1)||q(Y)\right] + \pi_2 D_{kl}\left[p(Y|c_2)||q(Y)\right] \quad (5)$$

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. After each merge, $p(Y|c_1)$ and $p(Y|c_2)$ are averaged to form the distribution of the new cluster $p(Y|c_{new})$. The number of clusters is determined by using a threshold on the Normalized Mutual Information given by $\frac{I(C,Y)}{I(X,Y)}$ (see [3] for details).

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 trained on the entire audio recording [3]. The input to the clustering algorithm is uniformly segmented speech segments ($X = \{x_j\}$), each composed of D consecutive speech frames. The posterior probability $p(y_n|x_j)$, i.e., the probability of each Gaussian component conditioned to the speech segment can be 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. The algorithm produces a partition of the data (i.e. a clustering) $p(C|X)$ as well as the distribution of relevance variables $p(Y|C)$ for each cluster c. The partition of data is a hard partition, i.e., $p(c_i|x_j) \in \{0,1\}$, meaning that each segment is assigned to only one cluster (a speaker). The distribution $p(Y|c_i)$ is obtained averaging the distributions $p(Y|x_j)$ for all the segments x_j assigned to the clustering c_i. Let us briefly summarize the differences between the two systems in Tab. 1:

Table 1: Main differences between the HMM/GMM and the IB diarization systems.

	HMM/GMM	IB
Modeling	a separate GMM for each speaker c	relevance variables Y from a background GMM
Distance	Modified BIC (Eqn. 2)	JS divergence (Eqn. 5)
Output	mapping $X \to C$	mapping $X \to C$ and $p(Y\|C)$

4. Information Bottleneck features

The HMM/GMM and IB system differ in a number of implementation issues (see Tab. 1) thus we could expect complementarity between them. This section describes how the output of IB system can be used as features in HMM/GMM diarization. Let us consider MFCC feature vectors $S = \{s_1, \ldots, s_T\}$ where s_t denotes the feature vector at time t; those are then segmented in $X = \{x_j\}$ chunks each containing D consecutive speech frames (feature vectors). The feature vectors S can be re-designated as $S = \{s^j_t\}$, where the superscript j denotes to which segment the feature vector belongs to. The output of the IB diarization is a hard partition of speech segments $x_j \in X$ into C clusters, i.e., $p(c_i|x_j) \in \{0,1\}$, meaning that each segment x_j is assigned to only one cluster. For each cluster, the associated relevance variable distribution $p(Y|c_i)$ is available (see previous section).

Thus each feature vector s^j_t belonging to segment x_j (given by the initial segmentation) can be associated to a cluster z_t obtained from the diarization output, i.e.,

$$z_t = \{c_i | s^j_t \in x_j, p(c_i|x_j) = 1\}, \;\; t = 1, ..., T. \quad (6)$$

Let us denote with F a matrix that contains the relevance variable distributions $p(Y|z_t)$ associated with each z_t, i.e.,

$$F = [p(Y|z_1), ..., p(Y|z_T)], \;\; t = 1, ..., T. \quad (7)$$

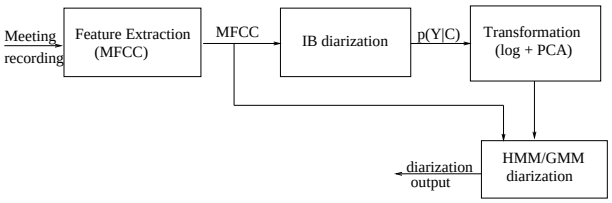

Figure 1: *Block diagram of the proposed method.*

F is a $|Y| \times T$ matrix where T is the number of speech frames and $|Y|$ is the cardinality of the relevance variable space.

F contains both information on the clustering output (if two feature vectors s_t and $s_{t'}$ belong to the same cluster), and characterizes each cluster with the distribution $p(Y|z_t)$ (different clusters will have different $p(Y|z_t)$). Thus TANDEM processing [8] can be applied, probabilities $p(Y|z_t)$ are gaussianized by a logarithm on their individual elements and then de-correlated using Principal Component Analysis (PCA). The PCA is also used to reduce initial dimensionality, equal to the relevance variable space cardinality ($|Y|$). The resulting matrix, designated as F_{IB} and referred as Information Bottleneck (IB) features can be used as input to a conventional diarization system where the GMM speaker models can be learnt from these features. The integration with MFCC can happen in two possible ways:

1 by concatenating IB features with MFCC features (as done in ASR) thus forming a single input vector to HMM/GMM system. This approach will be referred as IB_aug (the IB feature stream is augmented with MFCC features).

2 by multi-stream modeling, i.e., estimating a separate GMM model for each feature stream and combining their log-likelihoods [14]. This approach is used for instance in diarizing with features having very different statistics (like MFCC and Time Delay of Arrival features) and will be referred as IB_multistr. In this case, the clustering is based on the combined log-likelihood:

$$w_{mfcc} \log b_c^{mfcc} + w_{F_{IB}} \log b_c^{F_{IB}} \quad (8)$$

where b_c^{mfcc} and $b_c^{F_{IB}}$ are GMMs trained on MFCC and F_{IB} features and $(w_{mfcc}, w_{F_{IB}})$ are the combination weights.

The overall method can be summarized in three main steps given below and a block diagram of the proposed approach is shown in Fig. 1:

1 Perform IB diarization and estimate $p(C|X)$ and $p(Y|C)$.

2 Map $p(Y|C)$ to input frames S and apply TANDEM processing to obtain IB features (F_{IB})

3 Use F_{IB} as complementary features to MFCC in a conventional HMM/GMM system.

5. Experiments and Results

The experiments are conducted on 24 meetings recorded at different meeting room environments (CMU,EDI,NIST,VT,TNO) which were collected for the purpose of NIST RT06, RT07, RT09 evaluations [15]. The audio from multiple distant microphone channels of each meeting is beamformed using *BeamformIt* toolkit [16]. The beamformed output of each meeting is used for speech, non-speech detection and feature extraction. Acoustic features consist of 19 MFCC. The speech/non-speech detection is based on the AMIDA system and evaluated in terms of missed speech rate (Miss) and false alarm rate (FA) summing into the speech/non-speech error rate (SpNsp) (see Tab. 2). The performance is evaluated in terms of Diarization Error Rate

Table 2: *Speech/non-speech error rate in the evaluation data set.*

meeting	Miss	FA	SpNsp
ALL	7.3	0.4	7.7

(DER) which is the sum of speech/non-speech error and speaker error. For the purpose of comparison, only speaker error is reported here as same speech/non-speech is used for all the systems.

The number of principal components to be kept after PCA and the weights $(w_{mfcc}, w_{F_{IB}})$ are selected as the ones that minimize the speaker error on a separate development data set. The optimal number of principal components is found to be equal to two, covering more than 80% of the PCA variability. The feature weights $(w_{mfcc}, w_{F_{IB}})$ are found to be equal to $(0.9, 0.1)$. These values are then used for evaluation on RT06, RT07, RT09 meetings. Tab. 3 reports speaker error for the baseline system as well as the IB_aug and IB_multistr approaches. The meeting-wise performance is reported in Fig. 2. The baseline HMM/GMM system achieves a speaker error equal to 12%. The first approach IB_aug, which concatenates MFCC and F_{IB} features, degrades the performance producing an error equal to 13.5%. On the other hand, the second approach IB_multistr which estimates separate GMM models for MFCC and F_{IB} features, reduces speaker error to 9.7%, i.e., an improvement of approximatively 19% relative compared to the baseline. The degradation in performance produced by concatenation can be explained by the very different statistical properties of MFCC and F_{IB} features. In fact, F_{IB} features have smaller dimensionality compared to MFCC and are compact representation of IB diarization output, thus they do not share the same distribution of MFCC. Therefore, whenever the modeling is done using separate GMMs, speaker error decreases from 13.5% (IB_aug) to 9.7% (IB_multistr). This is similar to what was observed in case of TDOA features, as they also become affective only through multistream modeling [14].

Table 3: *Total speaker error with relative improvement over baseline in parenthesis on the evaluation data sets (RT06, RT07, RT09 combined) for various diarization systems.*

system	Baseline	IB_aug	IB_multistr
spkr err	12.0 (-)	13.5 (-12.5%)	9.7 (+19%)

It can be noticed from Fig. 2 that the IB_multistr shows significant improvement upon the baseline system in meetings with high error (over 15%). It is observed that the IB features have an effect on purity of clusters, i.e., assignment of segments uttered by different speakers to the same clusters is reduced thus producing much purer clusters compared to MFCC only (baseline). Reversely IB_aug often degrades the performances.

Let us investigate the effect of the F_{IB} features at different stages of the clustering. Fig. 3 plots speaker error for the baseline and IB_multistr after each merge, for the meeting EDI_20061113-1500. It can be noticed that both the systems have similar error rates in initial iterations but after few iterations, the F_{IB} features avoid wrong cluster merges, which increase error rate and produce a smooth and decreasing error curve. Similar trends are verified for other meetings where

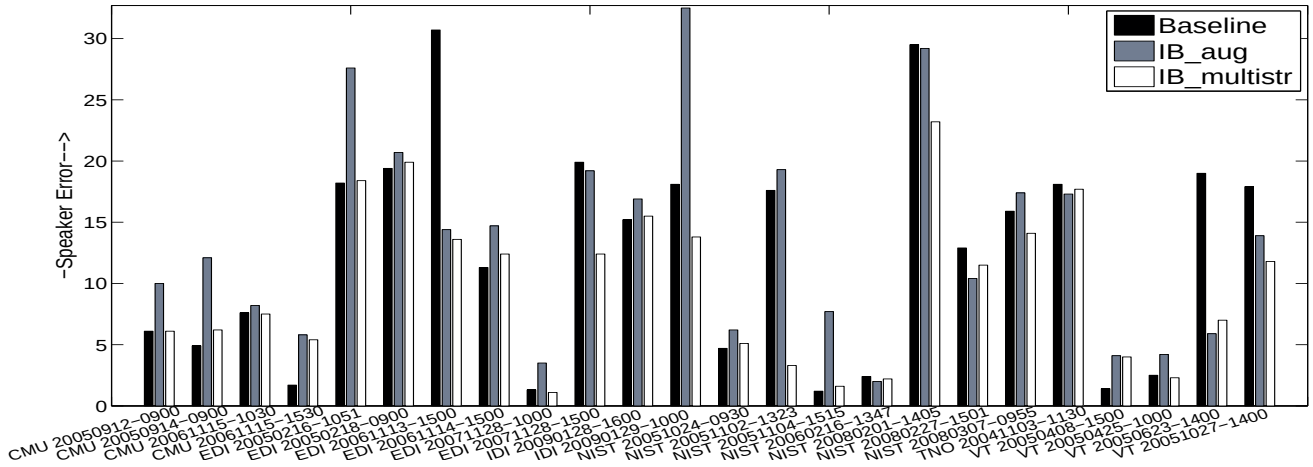

Figure 2: *Meeting wise speaker error values for baseline HMM/GMM diarization system and IB_multistr.*

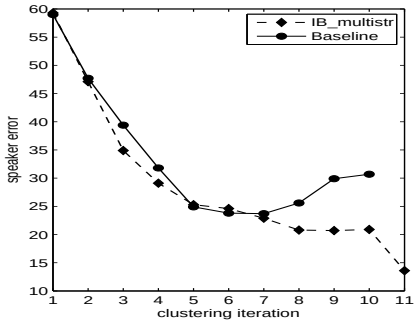

Figure 3: *Speaker error after each merge for baseline and IB_multistr systems for meeting EDI_20061113-1500*

IB_multistr achieves improvements over baseline.

6. Discussions and Conclusions

This paper proposes and investigates a novel approach to combine diarization systems through the use of *features*. The Information Bottleneck system is used to generate a set of features that contain information relevant to the clustering and characterize each speaker in terms of probabilities; these features are later used to complement MFCC in a conventional HMM/GMM system. The approach is largely inspired from TANDEM framework used in ASR and has the advantage of being fully integrated (features are used at all steps of agglomerative clustering) while it does not require any change to individual diarization components.

The combination with MFCC features is investigated using simple concatenation and using multi-stream modeling. Results on 24 meetings from the NIST RT06/RT07/RT09 evaluation campaigns reveal that the Information Bottleneck features reduce the speaker error from 12% to 9.7%, i.e., a 19% relative improvement when they are used together with MFCC in multi-stream fashion. The approach is particularly effective in meetings where the baseline system has speaker error higher than 15%. On the other hand, simple concatenation increases speaker error to 13.5% as F_{IB} and MFCC have very different statistical distributions to be modeled using same GMM. In summary the IB system provides complementary information to the HMM/GMM whenever the integration happens by multi-stream modeling.[1]

7. References

[1] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557–1565, September 2006.

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

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

[4] S. E. Tranter, "Two-way cluster voting to improve speaker diarisation performance," in *ICASSP*, 2005, pp. 753–756.

[5] D. Moraru and al., "The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation," in *ICASSP*, vol. 2, 2003, pp. 89–92.

[6] D. Moraru and al., "The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation," in *ICASSP*, vol. 1, 2004, pp. 373–376.

[7] S. Bozonnet and al., "System output combination for improved speaker diarization," in *INTERSPEECH*, 2010, pp. 2642–2645.

[8] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *ICASSP*, vol. 3, 2000, pp. 1635–1638.

[9] N. Morgan and al., "Pushing the envelope - aside," *IEEE Signal Processing Magazine*, vol. 22, no. 5, 2005.

[10] C. Wooters and M. Huijbregts, "Multimodal technologies for perception of humans," R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, ch. The ICSI RT07s Speaker Diarization System, pp. 509–519.

[11] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *Signal Processing Letters, IEEE*, vol. 11, no. 8, pp. 649–651, August 2004.

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

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

[14] J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Trans. Comput.*, vol. 56, pp. 1189–1224, September 2007.

[15] "http://www.itl.nist.gov/iad/mig/tests/rt/."

[16] X. Anguera, "Beamformit, the fast and robust acoustic beamformer," in *http://www.icsi.berkeley.edu/xanguera/BeamformIt*, 2006.

[1] Authors would like to thank Dr. Deepu Vijayasenan for his help with IB system. This work was funded by the Swiss Science Foundation through IM2 grant, by the EU through SSPnet grant and by the Hasler foundation through the SESAME grant.