MULTISTREAM SPEAKER DIARIZATION THROUGH INFORMATION BOTTLENECK SYSTEM OUTPUTS COMBINATION

Deepu Vijayasenan, Fabio Valente, Petr Motlicek

Idiap Research Institute, 1920, Martigny, Switzerland
{deepu.vijayasenan,fabio.valente, petr.motlicek}@idiap.ch

ABSTRACT

Speaker diarization of meetings recorded with Multiple Distant Microphones makes extensive use of multiple feature streams like MFCC and Time Delay of Arrivals (TDOA). Typically the combination happens using separate models for each feature stream. This work investigates if the combination of multiple feature streams can happen through the combination of multiple diarization systems performed using those features. The paper extends the previously proposed Information Bottleneck method to handle the combination of several probabilistic diarization outputs. In contrast to the conventional model-based feature combination, this technique is referred as system-based combination. Furthermore the paper introduces an hybrid model-system combination. Experiments are run on data from the Rich Transcription campaigns and show that the system based combination largely outperforms the model based combination by 37% relative. The hybrid approaches improve by $10 - 20\%$. The analysis of errors shows that the improvements come from the recordings where the individual MFCC and TDOA systems provide very different performances.

Index Terms— Speaker diarization, Information bottleneck principle, Feature combination, TDOA features, diarization system combination.

1. INTRODUCTION

Speaker diarization is an unsupervised learning paradigm with the objective of finding *"who spoke when"* in a given audio recording. Both the number of speakers and speech segments corresponding to each speaker need to be learnt. Whenever diarization is applied to meeting recordings, the system often make use of multiple feature streams like MFCC and Time Delay Of Arrivals (TDOA) extracted from a microphone array. The combination happens weighting the log-likelihoods of GMMs trained on different features [1], i.e., at model level. A number of studies on Broadcast data have discussed the combination of speaker diarization outputs including voting schemes [2], initialization of a system using the output of another one, like in case of bottom-up and top-down systems [3] or integrated approaches [4]. Recently they have been revisited also in the context of meeting recordings diarization [5]. Those approaches are able to outperform the individual diarization systems although no attempt has been done in using them to combine multiple feature streams.

This work investigates if the combination of multiple feature streams can happen through the combination of multiple diarization systems performed using those features. In our previous works, we introduced a multi-stream diarization method based on the Information Bottleneck (IB) principle [6]. The combination of multiple streams happens using intermediate variables that carry relevant information about the problem, referred as relevance variables. The system was shown to outperform conventional HMM/GMM systems and being effective in combining up to four different feature streams [6]. This paper proposes a novel combination framework based on the IB principle which aims at combining the output of multiple diarization systems. This novel combination will be referred as *system based* combination in contrast with what previously proposed in [6] referred as *model based* combination. Furthermore the paper introduces an hybrid model-system combination similar to the piped approaches described in [3] and [5]. Experiments are run on meetings recordings and aims at comparing model and system based combination of MFCC and TDOA features. The remainder of the paper is organized as follows: section 2 briefly describes the IB diarization system and section 3 describes its multi-stream extension. Sections 4 and 5 introduce the system based combination and the hybrid model-system combinations. Experiments are then reported in section 6 and the paper is concluded in section 7.

This work was funded by the SNF IM2 project, the EU FP7 SSPNet project and the Hasler fundation SESAME grant.

2. IB BASED DIARIZATION

This section briefly summarizes the IB speaker diarization system proposed in [7]. The Information Bottleneck is a distributional clustering technique introduced in [8]. 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:

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

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:

$$\Delta\mathcal{F}(c_1, c_2) = [p(c_1) + p(c_2)] JS[p(Y|c_1), p(Y|c_2)] \tag{2}$$

The Jensen-Shannon divergence $JS[p(Y|c_1), p(Y|c_2)]$ is given by: $\pi_1 D_{kl}\left[p(Y|c_1)||q(Y)\right] + \pi_2 D_{kl}\left[p(Y|c_2)||q(Y)\right]$ 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 $\frac{I(C,Y)}{I(X,Y)}$.

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 composed of D consecutive speech frames. The posterior probability $p(y_n|x_t)$, i.e., the probability of each gaussian component conditioned to the speech segment is computed in straightforward way 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 for each cluster c i.e $p(Y|C)$. The partition of the data is an hard partition, i.e., $p(c_i|x_t) \in \{0, 1\}$, meaning that each segment is assigned to a cluster (a speaker).

The distribution $p(Y|c_i)$ is obtained averaging the distributions $p(Y|x_t)$ for all the segments x_t assigned to the clustering c_i. The complete algorithm is summarized as follows.

1 MFCC feature extraction from the raw audio data.

2 Speech/non-speech segmentation and rejection of non-speech frames.

3 Uniform segmentation of speech in chunks of length $D = 250ms$, i.e., set X.

4 Estimation of a Gaussian component with shared diagonal covariance matrix for each segment i.e., set Y.

5 Estimation of conditional distribution $p(Y|x)$.

6 aIB clustering and model selection to determine the speaker clusters (Diarization output).

7 Realignment of the speaker boundaries using Viterbi decoding as described in [10].

3. MULTIPLE FEATURES COMBINATION

Multiple feature streams can be combined in the relevance variable space, i.e, using the posterior probabilities $p(Y|x_t)$. Let us consider here the use of MFCC and TDOA features. A separate GMM with the same number of components is trained for each feature stream. Their individual components are kept aligned. Let $\{\mathcal{M}_{mfcc}\}$ and $\{\mathcal{M}_{tdoa}\}$ be the background models estimated using the MFCC and TDOA features. The combined distribution $p(Y|x_t)$ is then obtained as:

$$p(Y|x_t) = W_{mfcc} \cdot p(Y|x_t, \mathcal{M}_{mfcc}) + W_{tdoa} \cdot p(Y|x_t, \mathcal{M}_{tdoa}) \quad (3)$$

where (W_{mfcc}, W_{tdoa}) are weights and $W_{mfcc} + W_{tdoa} = 1$. Once the relevance variable distributions $p(Y|x_t)$ are estimated using Eq. 3, the aIB clustering can be performed as before, generating thus the diarization output. In the remainder of the paper, it will be referred as *model based combination*, as the combination happens using separate GMM models.

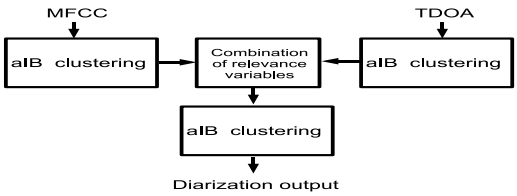

Fig. 1. Schematic representation of the system combination method based on the IB system.

4. MULTIPLE SYSTEMS COMBINATION

This section introduces a novel method for combining multiple diarization systems based on the IB principle where the relevance variables space is formed using the diarization outputs. Toward this end, let us first consider the output of two independent diarization systems S_{mfcc} and S_{tdoa} based on aligned GMMs $\{\mathcal{M}_{mfcc}\}$ and $\{\mathcal{M}_{tdoa}\}$ trained with MFCC and TDOA features. They respectively produce two hard partitions, i.e., two cluster assignments of segments x_t into clusters c_i:

$$p(c_i|x_t, S_{mfcc}) \in \{0, 1\} \quad p(Y|c_i, S_{tdoa}) \in \{0, 1\} \quad (4)$$

and two relevance variable distributions for each cluster:

$$p(Y|c_i, S_{mfcc}) \quad p(Y|c_i, S_{tdoa}) \quad (5)$$

obtained averaging $p(Y|x_i, M_{mfcc})$ and $p(Y|x_i, M_{tdoa})$ that are assigned to the same clusters by distributions (4). Two new distributions of relevance variables $P(Y|x_t)$ can be obtained from (4) and (5) as:

$$P(Y|x_t, S_{mfcc}) = \sum_{c_i} p(Y|c_i, S_{mfcc}) \cdot p(c_i|x_t, S_{mfcc}) \quad (6)$$

$$P(Y|x_t, S_{tdoa}) = \sum_{c_i} p(Y|c_i, S_{tdoa}) \cdot p(c_i|x_t, S_{tdoa}) \quad (7)$$

Note that $p(c_i|x_t)$ is equal to one for a cluster and zero elsewhere, thus $P(Y|x_t, S_.)$ will be equal to the distribution of the corresponding cluster $p(Y|c_i, S_.)$. The probabilistic output of the diarization systems in Eq.(6) and (7) can be combined into a single relevance variable distributions. The method is schematically depicted in Figure 4 and summarized as follows:

1 Perform single-stream IB diarization (as in section 2) using MFCC features and estimate $p(Y|c_i, S_{mfcc})$, $p(c_i|x_t, S_{mfcc})$.

2 Perform single-stream IB diarization (as in section 2) using TDOA features and estimate $p(Y|c_i, S_{tdoa})$, $p(c_i|x_t, S_{tdoa})$.

3 Estimate a new $p(Y|x_t)$ combining the output of the two single stream systems:

$$p(Y|x_t) = W_{mfcc} P(Y|x_t, S_{mfcc}) + W_{tdoa} P(Y|x_t, S_{tdoa}) \quad (8)$$

where W_{mfcc} and W_{tdoa} are the weights of the combination.

4 Perform aIB clustering and model selection using $p(Y|x_t)$ thus obtaining another partition in speakers (the diarization output).

In other words, instead of combining the MFCC and TDOA features using separate background GMMs (see Eq. 3) , the combination uses the the distributions of relevance variables after the clustering, i.e., the output of the diarization systems (see Eq. 8). In the following, we will refer to it as *system based* combination and briefly discuss its properties. The variable c in Eq.(6) and (7) is a 'dummy' variable thus the combination can happen even when the number of speakers (clusters) produced by the two systems is different. It is easy to verify that in the extreme cases, $(W_{mfcc}, W_{tdoa}) = (1, 0)$ and $(W_{mfcc}, W_{tdoa}) = (0, 1)$, the output after Step 4 will be equal to the output of the MFCC and TDOA system respectively.

The rationale behind this type of combination is related to the amount of data used to estimate $p(Y|x_t)$. In case of model combination, $p(Y|x_t)$ is estimated using the frames that compose the segment x_t, i.e., $p(Y|x_t, \mathcal{M}_{mfcc})$ and $p(Y|x_t, \mathcal{M}_{tdoa})$ (see Eq. 3). In case of system combination, $p(Y|x_t)$ is estimated using all the frames that are assigned to the same cluster, i.e., $p(Y|c_i, S_{mfcc})$ and $p(Y|c_i, S_{tdoa})$ (see Eq. 8). The amount of data is significantly larger compared to the first case. This could potentially provide a better estimate of the relevance variables distributions as they are averaged over several clusters. The combination comes at the price of an increased computational complexity; in fact, at first two independent diarization outputs must be obtained and finally the combined distribution is fed into a third system.

5. HYBRID SYSTEM-MODEL COMBINATION

Instead of combining the relevance variables from two background models or from two diarization systems, a third hybrid solution can be considered. Let us define the following combination:

$$p(Y|x_t) \quad = \quad W_{mfcc}\, p(Y|x_t, S_{mfcc}) + W_{tdoa}\, p(Y|x_t, M_{tdoa}) \tag{9}$$

where the output of the relevance variables distributions $p(Y|x_t, S_{mfcc})$ from the MFCC diarization system $p(Y|x_t, S_{mfcc})$ are combined with those from the TDOA-GMM, $p(y|x_t, M_{tdoa})$. The algorithm can be summarized as follows:

1 Perform single-stream IB diarization (as in section 2) using MFCC features and estimate $p(Y|c_i, S_{mfcc})$, $p(c_i|x_t, S_{mfcc})$.

2 Estimate a background GMM M_{tdoa} using TDOA features and the set of relevance variables $p(y|\mathcal{M}_{tdoa}, x_t)$.

3 Estimate a new $p(Y|X)$ using Eq. 9.

4 Perform aIB clustering using $p(Y|X)$ thus obtaining another partition in speakers (the diarization output).

A similar combination can be obtained inverting the order of MFCC and TDOA. This method will be referred as *hybrid system-model* combination. This approach is close in spirit to what proposed originally in [3] and later in [5] where the output of a first system is used as initialization for a second system; however the combination is here probabilistic.

6. 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 [11]. 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 [1]).

A critical part of multi-stream methods consists in determining the weights of different feature sets. In this work, the weights are estimated from a development dataset composed of 12 recordings across 6 meetings rooms. The weights are selected as those that minimize the speaker error on the development data set. 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. Table 1 reports the results of the conventional model based combination as described in section 3. Weights estimated on the development data are $(P_{mfcc}, P_{tdoa}) = (0.7, 0.3)$. For comparison purposes, the same table also reports the performance of a diarization system based on conventional HMM/GMM [12],[1] where the combination happens at log-likelihood level. Results show that the

Table 1. Speaker Error for the aIB model based combination and for a conventional HMM/GMM system that use MFCC and TDOA features.

	aIB	HMM
Speaker Error	11.6	12.4

aIB model-based combination outperforms the HMM/GMM system by 0.8% achieving state-of-the-art results; this system will be used as baseline and the proposed techniques will be benchmarked w.r.t. it. Table 2 table reports the performances in case of model combination (Case 1), system combination (Case 2) and hybrid model-system combination (Cases 3 and 4). The weights obtained from tuning on the independent development data set are also reported. The relative improvements in the brackets are computed w.r.t. the baseline system (Case 1). Results reveal that the system combination largely

Table 2. Speaker Error for the proposed combination schemes: model based, system based and the two hybrid combinations.

Case	MFCC	TDOA	(W_{mfcc}, W_{tdoa})	Speaker Error
1	Model	Model	(0.7,0.3)	11.6 (–)
2	System	System	(0.7,0.3)	7.3 (+37%)
3	System	Model	(0.8,0.2)	10.5 (+9%)
4	Model	System	(0.6,0.4)	9.4 (+19%)

outperforms the model combination by 37%. Figure 2 plots the per-meeting results in case of system and model combination. On the other hand, the hybrid system-model combination is effective in reducing the speaker error by 10% and 20% whenever the diarization output is from the MFCC or the TDOA systems respectively.

In order to investigate the reason of this effect, Table 3 reports the speaker error of the single-stream MFCC and TDOA systems and Figure 3 reports their per-meeting performance. As already discussed in several studies, the MFCC system largely outperforms the TDOA system, as the TDOA features are significantly more noisy. Furthermore Figure 3 shows that the difference in performance can vary a lot from meeting to meeting; in fact while in some recordings

Table 3. Speaker Error for the single stream diarization systems based on MFCC and TDOA features.

	MFCC	TDOA
Speaker Error	15.5	28.1

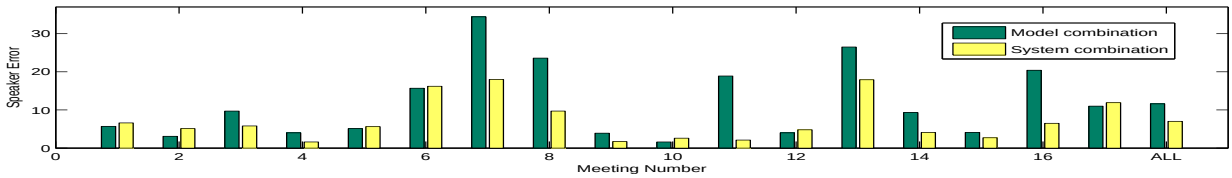

Fig. 2. Meeting wise speaker error for the model based and system based combination. The system based combination appears superior to the model based one.

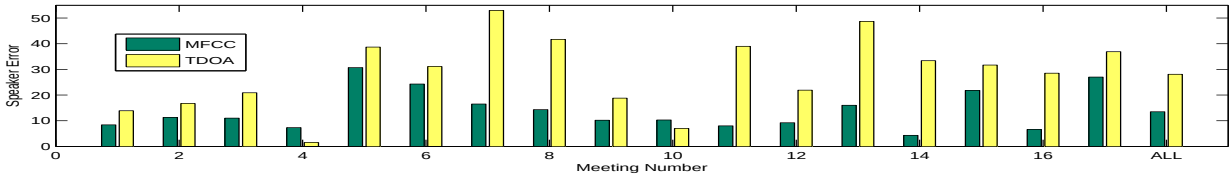

Fig. 3. Meeting wise speaker error for the single stream diarization systems based on MFCC and TDOA features. The MFCC system largely outperforms the TDOA system.

the performance of MFCC and TDOA is similar, on others, the error is almost three times larger. Figure 2 shows that the improvements of the system over the model combination, come from meetings where the individual performance of MFCC and TDOA features is very different.

The reason of this effect can be related to the estimation of $p(Y|x_t)$. In the model based combination, $p(Y|x_t)$ is obtained weighting $p(Y|x_t, M_{mfcc})$ and $p(Y|x_t, M_{tdoa})$ estimated using observations from the segment x_t. In the system based combination, $p(Y|x_t)$ is obtained weighting $p(Y|x_t, S_{mfcc})$ and $p(Y|x_t, S_{tdoa})$ estimated using the output of systems S_{mfcc} and S_{tdoa} thus significantly more data. If the features in the segment x_t are noisy like in case of TDOA, the estimation of $p(Y|x_t)$ will benefit from more data. This explanation is also supported by the performance of the hybrid combination schemes. In case 4, where the TDOA system output $p(Y|x_t, S_{tdoa})$ is combined with the MFCC model estimates $p(Y|x_t, M_{mfcc})$, the improvement over the baseline is 19%. In case 3, where the MFCC system output $p(Y|x_t, S_{mfcc})$ is combined with the TDOA model estimates $p(Y|x_t, M_{tdoa})$, the improvement becomes 9%. In both cases, it is beneficial to use the diarization output, the improvements being larger when the features are noisy like in case of TDOA. Further evidence of this can be noticed considering the combination weights obtained minimizing the error on an independent development data set. Weights are equal to (0.7,0.3) both in case of model and system based combination. On the other hand, they become (0.8,0.2) in case 3 and (0.6,0.4) in case 4 (hybrid model system combinations). Thus whenever the relevance variables are estimated using the same amount of data, their stream weighting is (0.7,0.3); in case of hybrid combination, the weighting moves towards the estimates done on larger amount of data, i.e., towards the output of the diarization system.

7. CONCLUSION AND DISCUSSIONS

Motivated by previous studies on combination of diarization systems outputs [4], [2],[5], this work investigates if the combination of multiple feature streams can happen throught the combination of multiple diarziation systems. Towards this end, the paper exthends the Information Bottleneck combination [6]. The method forms the space of relevance variables, necessary for the aIB clustering, using the output of two separate diarization systems. The rationale behind this is based on the fact that the probability of relevance variables $p(Y|x_t)$ is estimated using all the x_t assigned to the same cluster thus considerably more data. The investigation also covers an hybrid model-system combination similar to the piped diarization already discussed in [4],[5]. Experiments compare the model based and the system based combination whenever MFCC and TDOA features are used. Results reveal that the system combination largely outperforms the model combination by 37%; The hybrid solutions improve over the baseline by $10 - 20\%$ relative without outperforming the system based combination. The improvements come from the recordings where the difference in performance between the MFCC system and the TDOA system is large. The analysis of the proposed method suggests that using the system outputs is particularly useful for averaging the performances of noisy features like TDOA. In summary the results show that the information from multiple feature streams can be more effectively used through system combination rather then model combination. The improvements come at the cost of running multiple diarization.

While this investigation has been limited to two diarization systems trained on MFCC and TDOA features, in future we plan to experiment with a larger number of streams/systems like in [6] where the MFCC is combined with a larger number of poor-performance features (up to four).

8. REFERENCES

[1] Xavier Anguera, *Robust Speaker Diarization for Meetings*, Ph.D. thesis, Universitat Politecnica de Catalunya, 2006.

[2] Tranter S.E., "Two-way cluster voting to improve speaker diarisation performance," in *Proc. of ICASSP*, 2005.

[3] Moraru D. et al., "The elisa consortium approaches in speaker segmentation during the nist 2002 speaker recognition evaluation," in *Proc. of ICASSP*, 2003.

[4] Moraru D. at al., "The elisa consortium approaches in broadcast news speaker segmentation during the nist 2003 rich transcription evaluation," in *Proc. of ICASSP*, 2004.

[5] Bozzonet S. et al., "System output combination for improved speaker diarization," in *Proc. of Interspeech*, 2010.

[6] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard, "Multistream speaker diarization beyond two acoustic feature streams," in *Proc. of ICASSP*, 2010.

[7] Deepu Vijayasenan, Fabio Valente, and Hervé 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.C. 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] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard, "KL realignment for speaker diarization with multiple feature streams," in *10th Annual Conference of the International Speech Communication Association*, 2009.

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

[12] J.M. Pardo , X. Anguera, C. Wooters, "Speaker Diarization For Multiple-Distant-Microphone Meetings Using Several Sources of Information," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1189, 2007.