MULTISTREAM SPEAKER DIARIZATION BEYOND TWO ACOUSTIC FEATURE STREAMS

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

[1]Idiap Research Institute, 1920, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), 1015, Lausanne,Switzerland
{deepu.vijayasenan,fabio.valente, herve.bourlard}@idiap.ch

ABSTRACT

Speaker diarization for meetings data are recently converging towards multistream systems. The most common complementary features used in combination with MFCC are Time Delay of Arrival (TDOA). Also other features have been proposed although, there are no reported improvements on top of MFCC+TDOA systems. In this work we investigate the combination of other feature sets along with MFCC+TDOA. We discuss issues and problems related to the weighting of four different streams proposing a solution based on a smoothed version of the speaker error. Experiments are presented on NIST RT06 meeting diarization evaluation. Results reveal that the combination of four acoustic feature streams results in a 30% relative improvement with respect to the MFCC+TDOA feature combination. To the authors' best knowledge, this is the first successful attempt to improve the MFCC+TDOA baseline including other feature streams.

***Index Terms*—** Speaker diarization, Information bottleneck principle, Feature 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. Conventional systems use short term spectral features such as mel frequency cepstral coefficients (MFCC) and follow an agglomerative approach for this purpose [1].

Recently diarization systems are converging to a multistream approach. Various other features that carry complimentary information have been explored in combination with MFCC features. In case of data recorded with Multiple Distance Microphones (MDM), Time Delay of Arrivals (TDOA) of the sound in different microphones carry information about the location of the speaker [2]. Most of the state of the art systems in speaker diarization uses a combination of MFCC and TDOA features. Different other features including long term and prosodic features were used in combination with MFCC to improve the diarization performance [3, 4]. However according to the authors' best knowledge there was little attempt to incorporate additional features to MFCC+TDOA baseline.

Conventional systems use an ergodic HMM where each speaker is modeled using an HMM state with minimum duration. State emission probabilities are modeled with Gaussian Mixture Models (GMM). In case of multistream diarization, separate models are built for individual feature streams. The feature combination is performed by a linear combination of the two individual log likelihoods. However, different feature streams possess very diverse statistical properties. For example, the dimension of TDOA features depends on number of microphones used, which varies across meeting rooms. This could lead to two different issues. Different features might require GMMs of different complexities (number of gaussians). In addition the dynamic range of individual feature streams could be totally different. Thus using a linear combination of log-likelihoods may not be appropriate with multiple feature streams (see [5] for details).

This study builds on our recent works [5], [6], [7]. In [6] we proposed a non-parametric approach based on Information Bottleneck principle for speaker diarization. The system was then extended to multiple streams (MFCC+TDOA) [7] and later a Kullback-Leibler divergence based realignment was introduced [5]. Such a system solely depends on posterior probability distributions of each feature stream and avoids combination of GMM log-likelihoods.

In this work, we extend this system beyond the combination of two sets of acoustic features streams. We investigate the use of cepstral like features obtained from frequency domain linear prediction(FDLP) and modulation spectrum(MS) features in addition to the conventional MFCC and TDOA. The paper examines issues in estimating the feature weights of four different streams. In particular we show that, when more than two acoustic streams are used, the Diarization Error becomes a non-smooth function of the weights without a well defined minimum. We investigate a solution to overcome the problem. The paper is organized as follows. The next section introduces speaker diarization based on Information Bottleneck (IB) principle. Section 3 describes the feature combination scheme. Performing a speaker realignment with multiple feature streams is described in Section 4. Section 5 describes the baseline feature combination of MFCC and TDOA features. Section 6 investigates the feature combination of four acoustic features. The paper is concluded in Section 7.

2. IB BASED DIARIZATION

Consider a set of speech segments $X = \{x_1, \ldots, x_T\}$ obtained from uniform linear segmentation of an input audio stream, to be clustered into set of clusters $C = \{c_1, \ldots, c_K\}$. Let Y be a set of relevance variables, that contain relevant information about the problem. Motivated by the wide success of GMM for speaker recognition, we had proposed to use the components of a background GMMs as the set of relevance variables for speaker diarization [6]. According to IB principle the best clustering compresses the input variables while preserving as much mutual information as possible about the relevance variables Y [8]. This corresponds to the minimization of:

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

Where β is a Lagrange multiplier. The clustering operates using probabilities $p(y|x)$ that are obtained using Bayes' rule. This criterion is optimized with respect to the stochastic mapping $p(c|x)$ using iterative optimization techniques [8].

The optimization of the objective function (1) can be done in a greedy fashion using the agglomerative Information Bottleneck method [9]. The algorithm is initialized with the trivial clustering of each point considered as a separate cluster ($|X|$ clusters). At each step of the algorithm a cluster merge is performed such that the information loss with respect to the relevance variables is minimum. The loss of mutual information at each step is given by a Jensen-Shannon divergence which is straightforward to compute from the posterior distribution $p(y|x)$. The optimal number of clusters are selected based on a threshold on the Normalized Mutual Information (NMI) $\frac{I(C,Y)}{I(X,Y)}$. The complete algorithm is summarized as follows.

1 Feature extraction from the beamformed audio.

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

3 Uniform segmentation of speech in chunks of fixed size 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)

Full details can be found in [6].

3. MULTIPLE FEATURE COMBINATION

The feature combination is performed in the relevance variable space, i.e, using the posterior probabilities $p(y|x)$. For each feature stream F_i, a background GMM $\mathcal{M}_i$ is estimated, and a posterior distribution $p(y|\mathcal{M}_i, x)$ calculated. The combined distribution is then calculated as:

$$p(y|x) = \sum_i p(y|\mathcal{M}_i, x)P_i \qquad (2)$$

Where P_i is the feature weight corresponding to i^{th} feature stream ($\sum_i P_i = 1$). The combination is more robust to different dimensionality or different statistics of the features, since the combination happens at the posterior level rather than combining log likelihoods [7].

4. REALIGNMENT

The aIB algorithm produces clusters that are aligned to the boundaries of the initial segments. Those boundaries can be realigned using the speaker models to further improve the performance. Typically an HMM/GMM based realignment is performed. The optimal speaker segmentation $\mathbf{c}_{opt}$ is obtained as:

$$\mathbf{c}_{opt} = \arg\min_{\mathbf{c}} \sum_t -\log(b_{c_t}(x_t)) - \log(a_{c_t c_{t+1}}) \qquad (3)$$

Where c_t denotes the cluster ID at time index t, $b_{c_t}(.)$ the emission probability (GMM) of cluster c_t and $a_{c_i c_j}$ the transition probability from cluster c_i to cluster c_j. In case of multiple feature streams, the log likelihood $\log(b_{c_t}(x_t))$ is computed as the linear combination of likelihoods of individual GMM models. This might not be appropriate for streams with different number of features (e.g. different number of delays as estimated from different microphones).

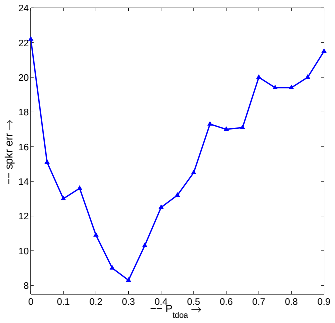

Fig. 1. speaker error of the development data as a function of TDOA weight in MFCC+TDOA feature combination ($P_{mfcc} = 1 - P_{tdoa}$)

In [5] we had proposed an alternative algorithm that depends only on posterior distribution of the relevance variables. The optimal speaker segmentation is computed as:

$$\mathbf{c}_{opt} = \arg\min_{\mathbf{c}} \sum_t KL(p(Y|x_t)||p(Y|c_t)) - \log(a_{c_t c_{t+1}}) \qquad (4)$$

Where $p(Y|x_t)$ is the posterior distribution of the relevance variables at a given speech segment x_t and $p(Y|c_t)$ is the posterior distribution of relevance variables in a given cluster c_t. It can be shown that Equation 4 optimizes a special case of IB criterion with a minimum duration constraint [5]. The optimization is solved by an EM algorithm that operates in the posterior space. In case of multistream diarization, the combined posteriors estimated as in Equation 2 are used for the realignment thus avoiding issues with log likelihood combination.

The entire diarization system (IB clustering, feature combination and realignment)works in the space of relevance variables and avoids the log-likelihood combination.

5. BASELINE MFCC AND TDOA FEATURES

The baseline system is based on the combination of MFCC and TDOA features. We use the NIST RT06 evaluation data for meeting diarization task for evaluating the algorithm. The dataset consists of nine meetings recorded across different locations and the number of channels and number of speakers vary across different meetings. The data preprocessing and beamforming is performed with *BeamformIt* [10] toolkit. The bug fixed version of *BeamformIt 2.2* is used for this purpose which provides different features compared to those used in [7]. We verified an improvement with the new beamforming in the MFCC based system as compared to what was reported in [7].

Diarization systems are evaluated using Diarization Error Rate (DER). DER is the sum of speech/non-speech error (missed speech and false alarm errors) and speaker errors. The same speech/non-speech segmentation is used across all the experiments. The total speech/non-speech error is 6.6% in RT06 evaluation data. Hence only the speaker error will be reported in all results.

Feature stream weights $\{P_i\}$ are optimized minimizing the speaker error on a development data. The development data contains 10 meetings. Figure 1 represents speaker error as a function of TDOA weight. It can be seen that the function has a well defined minimum at $P_{tdoa} = 0.3$. Speaker error obtained on RT06 evaluation data is reported in Table 1 and is equal to 11.6%.

Table 1. Baseline MFCC+TDOA speaker error $(P_{mfcc}, P_{tdoa}) = (0.7, 0.3)$

Feature	no realgn	hmm/gmm	kl
mfcc+tdoa	11.6	10.7	9.9

Table 2. Evaluation data results – The minimum obtained from development data is 3.3% worst than the oracle.

Weight selection scheme	spkr err
Devdata tun.	10.1
Oracle	6.8

The speaker boundaries can be realigned using an HMM/GMM system or a KL divergence based system [5]. Table 1 reports results with both realignments; the KL based realignment outperforms the HMM/GMM realignment by 0.8%. Thus the baseline system used for this study has a speaker error equal to 9.9%.

6. COMBINATION OF FOUR FEATURES

In addition to conventional MFCC and TDOA features, we explore two feature sets that are extracted from long time windows:

- **Modulation Spectrum Features** – The modulation spectrum (MS) represents the slowly varying components of the short term spectrum. The critical band energies trajectories are filtered using a gaussian low pass filter and the resulting features are decorrelated [4].
- **FDLP Features** – Frequency Domain Linear Prediction (FDLP) provides a smoothed temporal envelope [11]. FDLP is performed over the sub-bands of the audio signal over a large time window (typically 1 second) that yields a parametric model of the temporal envelope. Short term temporal energy integration is performed over the smoothed envelope, and the short term spectral energies are converted to short term cepstrum like features. A gain normalization step in the linear prediction helps to remove the artifacts from the reverberrent speech. Details can be found in [12].

Let us first, consider the four features combination without realignment. As before, the weights are tuned on the development data and the minimum is achieved for $(P_{mfcc}, P_{tdoa}, P_{ms}, P_{fdlp}) = (0.70, 0.25, 0.00, 0.05)$. Most of the weight is concentrated on MFCC and TDOA features; FDLP features receive a very small weight and MS features are discarded. The results obtained on the evaluation data are reported in Table 2 (first line). The four streams reduce the speaker error from 11.6% to 10.1%. However, if we consider the lowest possible speaker error by oracle weight selection (second line of Table 2), we can see that weights obtained from tuning perform worse by more than 3% absolute. This shows that weight selection on the development data is not effective. In order to understand how the speaker error varies on the development data, let us study the error as a function of the feature weights. The four weights ($P_{mfcc}, P_{tdoa}, P_{ms}, P_{fdlp}$) lie in a three dimensional subspace ($\sum_i P_i = 1$). The speaker error can be visualized by fixing one of the weights and plotting as a function of the other two. Figure 2 illustrates the speaker error as the function of two feature stream weights in a local neighborhood of the global minimum. It can be observed that the speaker error is a highly non smooth function with no well defined minima unlike in the case of MFCC+TDOA features. Note that the speaker error varies considerably (up to 3%) in the nine nearest points itself (Table 3).

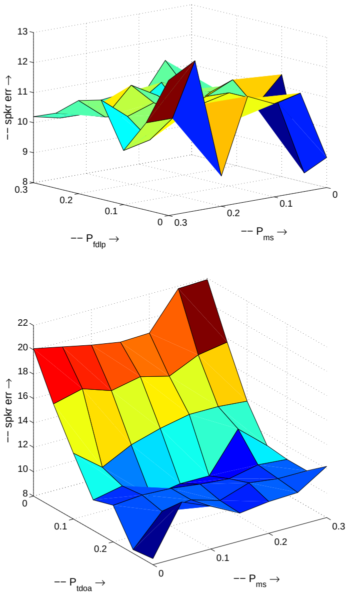

Fig. 2. Speaker error as a function of feature stream weights and fdlp features in the neighborhood of global minimum. (Top) Fixed $P_{tdoa} = 0.25$ (Bottom) Fixed $P_{fdlp} = 0.05$. $P_{mfcc} = 1-(P_{tdoa}+ P_{ms} + P_{fdlp})$

In order to avoid points with considerable variation of speaker error in the neighborhood, we proposed to use a smoothed version of the speaker error. This would force the algorithm to select feature weights in a region that has low speaker error as compared to an isolated minima with lot of variation in the neighborhood. The original speaker error function is convolved with a multidimensional

Table 3. Speaker error in the nine nearest points of the global minima (denoted in bold). Other entries are at the same distance from the minima

P_{mfcc}	P_{tdoa}	P_{ms}	P_{fdlp}	spkr err
0.65	0.25	0.00	0.10	11.4
0.65	0.25	0.05	0.05	10.8
0.65	0.30	0.00	0.05	8.5
0.70	0.20	0.00	0.10	10.6
0.70	0.20	0.05	0.05	11.0
0.70	**0.25**	**0.00**	**0.05**	**8.3**
0.70	0.25	0.05	0.00	11.3
0.70	0.30	0.00	0.00	8.3
0.75	0.25	0.00	0.00	9.0
0.75	0.20	0.00	0.05	11.0

Table 4. Comparison of speaker error : selecting the feature stream weights based on minimum value of speaker error Vs minimum value of smoothed speaker error

	using spkr err	smoothed spkr err
validation data	10.9	7.5
evaluation data	10.1	8.3

Table 5. Meeting wise speaker error for two as well as four feature stream combination with and without KL realignment

	mfcc+tdoa		four feats	
Meeting	no realgn	realgn	no realgn	realgn
CMU_20050912-0900	7.6	5.7	9.0	5.9
CMU_20050914-0900	4.8	3.1	4.4	3.1
EDI_20050216-1051	7.1	5.1	5.5	**3.9**
EDI_20050218-0900	18.6	15.7	7.4	**5.5**
NIST_20051024-0930	5.5	3.9	3.9	**3.2**
NIST_20051102-1323	2.5	1.6	4.1	2.8
TNO_20041103-1130	28.3	26.5	26.1	**24.4**
VT_20050623-1400	22.0	20.4	6.2	**4.1**
VT_20051027-1400	12.1	11.0	8.4	**8.1**
ALL	11.6	9.9	8.3	**6.7**

Gaussian given by:

$$\begin{aligned} g[l,m,n] &= e^{-(l^2+m^2+n^2)}, |l|,|m|,|n| \leq 1 \\ &= 0, \text{otherwise} \end{aligned}$$

The filter is a low pass filter centered at origin. The filter computes a weighted average of the center point with its 26 nearest neighbours in the three dimensional input space. The weights are then chosen as the point where the smoothed speaker error is minimum. Using such an approach, the weights are $(P_{mfcc}, P_{tdoa}, P_{ms}, P_{fdlp}) = (0.50, 0.20, 0.05, 0.25)$ which are quite different from the global minimum. We use a separate validation data (independent of both development and evaluation data) to verify the effectiveness of the approach. The validation data consists of a set of 8 meetings used in NIST evaluations. The results are present in Table 4 (first line). Using the smoothed version of speaker error results in a 3.4% absolute improvement in the validation data (first line). When tested on the evaluation data, this approach produces an improvement of 1.8% absolute over unsmoothed speaker error. Furthermore the obtained error 8.3% is only 1.5% worst then the oracle error. Also note that this smoothing does not alter the weight selection in case of the baseline MFCC+TDOA feature combination for which the minimum stays in the same point.

After the feature combination and clustering, a KL realignment is performed to refine the speaker boundaries. The realignment improves the speaker error consistently across all meetings by 1.6%. Table 5 provides meeting-wise results for the baseline (MFCC+TDOA) and for the four features with and without realignment.

In summary before realignment, the four feature streams system outperform the MFCC+TDOA baseline by 3.3% absolute (from 11.6% to 8.3%). After realignment the improvement is 3.2% absolute (from 9.9% to 6.7%) i.e. 30% relative better then the baseline.

7. CONCLUSIONS

In this work we proposed a speaker diarization system that incorporates four feature streams extending previous work on IB based diarization. In addition to the conventional MFCC and TDOA features, we combine two other features: modulation spectrum and FDLP.

When the weights are estimated according to minimum speaker error on the development dataset, the four stream improves by 1.5% absolute on the MFCC+TDOA baseline. However speaker error of four stream system is not a smoothly varying convex function as in case of two features. If the speaker error on development data is smoothed with a gaussian filter before weight selection, the speaker error goes from 10.1% to 8.3%. After the KL realignment, the four feature system achieves an error equal to 6.7% i.e. 30% relative improvement over the MFCC+TDOA baseline.

According to the authors' best knowledge, this is the first successful attempt in combining more acoustic features beyond the MFCC and TDOA features in speaker diarization.

8. 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 would like to thank our colleagues at Idiap and ICSI for their help with setting up the system.

9. REFERENCES

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

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

[3] Gerald Friedland, O. Vinyals, Yan Huang, and C. Müller, "Prosodic and other Long-Term Features for Speaker Diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 985–993, 2009.

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

[5] 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.

[6] 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.

[7] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard, "Integration of tdoa features in information bottleneck framework for fast speaker diarization," in *Interspeech 2008*, 2008.

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

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

[12] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of Reverberant Speech Using Frequency Domain Linear Prediction," *IEEE Signal Processing Letters*, vol. 15, pp. 681–684, 2008.