FILTERBANK SLOPE BASED FEATURES FOR SPEAKER DIARIZATION

Srikanth Madikeri[1] *and Hervé Bourlard*[1,2]

[1] Idiap Research Institute, CH-1920 Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
srikanth.madikeri@idiap.ch, herve.bourlard@idiap.ch

ABSTRACT

In this paper, filterbank slope based features are applied to the Information Bottleneck based system for speaker diarization. The filterbank slope based features have shown promise in the context of speaker recognition systems owing to their ability to emphasize formants. Hence, it is proposed to study their use in the context of speaker diarization as well, where speaker discrimination is equally important. The feature is explored using two different filterbank arrangements, linear and Mel, to form the Linear Filterbank Slope (LFS) and Mel Filterbank Slope (MFS), respectively. Both arrangements are shown to be inherently better at speaker discrimination compared with MFCC (Mel Frequency Cepstral Co-efficients). The feature streams are tested on the NIST RT06, 07 and 09 datasets. A best case relative improvement of 22.1% and 37.1% is observed for LFS and MFS, respectively, when compared with the MFCC-based baseline. The combination with time domain features is also studied and further improvements are observed. Finally, results on the fusion of multiple features are presented.

Index Terms— Filterbank slope, Speaker Diarization, Information Bottleneck

1. INTRODUCTION

Speaker diarization systems address the problem of finding "who spoke when" in an audio recording. Conventional speaker diarization systems take an unsupervised approach to solving the problem. The HMM/GMM (Hidden Markov Model/Gaussian Mixture Model) based system for diarization is a popular technique used in this context [1, 2]. Alternately, an information theoretic-approach, called the Information Bottleneck (IB) method, is also applied [3]. In this paper, we focus on a diarization system that uses the latter approach.

The existing approaches to speaker diarization depend on short-term cepstral features extracted from the audio in the form of MFCC (Mel Frequency Cepstral Co-efficients) feature streams. MFCC is ubiquitously used in most of the speech technology related tasks, including speaker recognition and speech recognition, and depending on the task the configuration of the front-end used to extract MFCC varies. For speaker diarization systems, the front-end parameters are tuned such that the speaker characteristics are emphasized. This is all the more important given that the modelling approaches in speaker diarization are unsupervised in nature [4].

Speaker characteristics are known to be available in the high order formants [5, 6]. The design of the Mel filterbank, which forms the basis of MFCC feature extraction, is such that there is high resolution at low frequencies but low resolution at high frequencies [7]. To compensate, the Mel Filterbank Slope (MFS) feature is used to provide equal emphasis to formants across all frequencies [8, 9]. In MFS, the slope across consecutive filterbank energies are calculated in order to provide equal emphasis to all formants. Results on speaker recognition tasks indicate that MFS is a suitable feature for speaker recognition and performs better than the MFCC based system [9, 10]. These results suggest that speaker characteristics are inherently emphasized by MFS and can be used in related tasks. Thus, it is proposed to apply filterbank slope based features to the speaker diarization task. In the attempt to emphasize formants, it will be interesting to study if a linear filterbank arrangement, in which filters have equal width at all frequency bands, will be more suitable for this task. In [11], it is shown that the linear filterbank arrangement is better suited compared with the Mel scale for discriminating female speakers. Analysis on the effect of linear filterbank arrangement is, therefore, proposed. This feature is termed Linear Filterbank Slope (LFS).

In this paper, it is shown that MFS and LFS can be used along with TDOA to outperform the equivalent MFCC-based system for speaker diarization. The IB based system is used for all the experiments unless mentioned otherwise. The rest of the paper is organized as follows: Section 2 provides the feature extraction procedure for MFS. Section 3 gives the details of IB based diarization system used in this work. Results of experiments conducted on NIST RT 05, 06, 07 and 09 datasets are discussed in Section 4.

2. FILTERBANK SLOPE BASED FEATURES

The extraction of the MFS feature vector from a frame of speech is shown in Figure 1. In the case of LFS extraction,

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

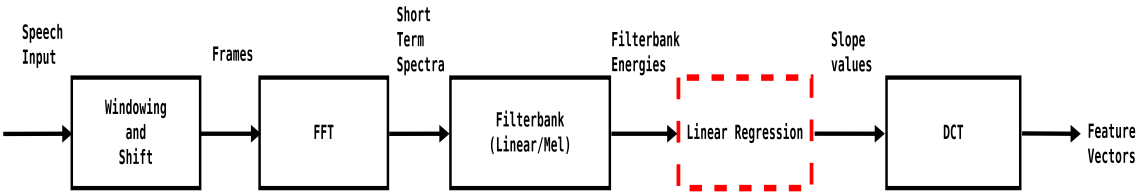

Fig. 1. Block diagram representing Filterbank slope feature extraction

the Mel filterbank is replace by a linear filterbank. For every frame of speech, the filterbank energies (FBE) are computed. Log of these filterbank energies are used. The filterbank energies are subtracted from the mean of filterbank energies computed over the entire utterance. This is analogous to applying Cepstral Mean Subtraction on MFCC feature streams [12]. The first set of n_{reg} FBE values, from the 1^{st} FBE to n_{reg}^{th} FBE, are used to compute the first slope value. The value of n_{reg}, typically varies from 3 to 7 [9]. In [10], a value of 4 is observed to be optimal and this is retained throughout this paper. The procedure is continued to obtain a vector of slope values, of dimension $F-1$ (F being the number of filterbanks), by shifting one filterbank at a time. The slope value is directly computed through a linear regression fit. While computing the final set of slope values for a frame of speech, fewer than n_{reg} values are considered for the line fit. It is observed that including these slope values is indeed beneficial as opposed to ignoring them.

DCT is applied on the slope vector to compress the data as there is significant overlap in information across consecutive slope values for a frame. The DCT is shown to achieve significant decorrelation with negligible reduction in performance. Moreover, should covariance matrices need be computed, it suffices to assume them to be diagonal.

The extraction procedure for both filterbank slope based features is similar to that of MFCC, the difference being that slope is computed from consecutive filterbank energies before applying the Discrete Cosine Transform (DCT). Computing slope from consecutive filterbank energies emphasizes formant locations. All formants get equal emphasis through the presence of zero-crossings in the slope values.

As mentioned earlier, two different filterbank arrangements are explored - the linear filterbank and the Mel filterbank. These filterbank arrangements are used to compute the Mel Filterbank Slope (MFS) and Linear Filterbank Slope (LFS), respectively. The divergence from the Mel filterbank towards linear filterbank is proposed in order to study if the filterbank arrangement makes a difference for diarization. In case of LFS feature extraction linear filterbanks are used instead of Mel filterbanks.

3. INFORMATION BOTTLENECK SYSTEM

The Information Bottleneck (IB) method, introduced in [3], performs diarization by optimizing the clusters with respect to a set of relevance variables. The optimization criterion is given as follows:

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

where X is the feature set, Y is the set of relevance variables and C is the set of clusters. β is the Lagrangian multiplier that controls the trade-off between information preserved in the clusters and the cluster size. The term I refers to mutual information between two random variables.

The IB system is used to present our results throughout this work. The conventional IB system is built primarily on MFCC based features with time domain information from the Time Delay Of Arrival (TDOA) features ([13]) when multiple distant microphone (MDM) recordings are involved [14]. Features such as Frequency Domain Linear Prediction (FDLP) and Modulation Spectrum (MS) are also used [15] to add complementary information. The advantages of using an IB based system for diarization in place of the conventional HMM/GMM based system are as follows: the IB system is faster compared to the HMM/GMM system [15]. Moreover, it is easier and advantageous to combine multiple features in the IB framework compared to the HMM/GMM framework. Combining multiple features involves fusing the posteriors before clustering. Whereas in the HMM/GMM framework, a common way to combine information from different features is to fuse the individual likelihood scores prior to Viterbi decoding [16], which requires model re-estimation at every iteration and is thus compute intensive. The primary focus of this paper is to study replacing the primary feature of the IB system with Filterbank slope based features.

In the IB framework, an audio recording is split into short segments based on the Voice Activity Detector's (VAD) output and an upper limit on the length of a segment. The upper limit is typically $2.5s$. Each segment is parameterized by a multivariate Gaussian distribution based on the features with respect to the segment. While the mean is computed from the segment, the covariance is computed from the entire utterance and used as a shared parameter. The parameters of the Gaussians are used to compute the posterior of each segment. The posteriors form the relevance variables Y in Equation 1. The agglomerative information bottleneck (aIB) clustering algorithm is applied to these segments [17, 18].

In a greedy approach such as aIB, it is important that each segment being combined has sufficient information about the speaker in the segment. As the Gaussians representing the segments are required to be inherently discriminative, it is

Table 1. Mean and variance measures of inter-class and intra class KL-divergence values for segments within a recording. Observations are show as mean/variance in the scale of 10^{-1}

Feature	Intra-class	Inter-class
MFCC	4.1/8.8	5.0/9.9
LFS	3.0/6.8	3.6/7.1
MFS	3.0/7.7	3.6/8.4

proposed to use a feature that captures speaker specific information better than MFCC. It is hypothesized that Filterbank slope based features are inherently better in doing so. This assumption is derived from the design of the filterbank slope based feature that has an explicit step in the feature extraction procedure to emphasize formants. In the following subsection this is empirically quantified.

3.1. Analysis: MFCC vs Filterbank Slope features

To show that filterbank slope based features are effective in capturing speaker information the following experiment is conducted: the parameters of the Gaussians of the short segments formed at the start of the IB method are analysed for each of the MFCC and filterbank slope based systems. The parameters defining these three systems are later described in Section 4. Recordings from RT05 dataset are used for this purpose. The Kullback-Leibler (KL) divergence measure between two segments is computed [19]. The pairwise comparison is treated as either within-class distance or inter-class distance based on the identity of speaker in each of the two segments. As the Gaussians share the covariance, the KL divergence measure simplifies to the Mahalanobis distance between the mean parameters. The average of intra-class divergence and inter-class divergence are computed. This average is computed by considering all such distances across all the recordings in the RT05 dataset. The KL divergence values are reported in Table 1.

As the results suggest, the mean and variance of KL divergence values of both inter-class and intra-class are observed to reduce for the filterbank slope based features. The reduction in the variance increases the separability between the inter-class and intra-class pairs indicating that filterbank slope features can indeed be beneficial in the IB framework. Even though the means of the inter-class and intra-class distances are approximately the same for the LFS and MFS, the variances are less for LFS compared to that for MFS. However, the separability between speakers provided by both MFS and LFS has appreciably increased compared to that provided by MFCC.

4. EXPERIMENTS

Speaker diarization experiments are performed on the NIST RT 05, 06, 07 and 2009 benchmark datasets. The NIST RT05 is used as a development dataset while others form the test set. The development set is used to tune β for IB clustering, fusion weights while combining multiple feature streams and select optimal number of cepstral co-efficients for the Filterbank slope features. Multiple Distant Microphone (MDM) recordings are used for the experiments after their enhancement using *Beamformit* [20].

Three different baseline systems are developed. The first baseline system is a conventional HMM/GMM system based on MFCC to compare the proposed features on the state-of-the-art methods. The implementation is based on the procedure described in [2] with the initial number of segments set to 16. The other two baseline systems are based on MFCC and TDOA features built using the IB system: one that uses only MFCC and the other includes TDOA with MFCC. MFS and LFS are compared in both modelling frameworks: HMM/GMM and IB. Furthermore, in the testing phase different combinations of the features are tested in the IB system as the framework lends itself to simple fusion of different features. The different combinations include: MFS, MFS + TDOA, LFS, LFS + TDOA, MFS + MFCC + TDOA, LFS + MFCC + TDOA and MFS +LFS +TDOA; where the symbol '+' is used to refer to fusion. The combination of all 4 features is not presented as the gains are not appreciable.

MFCC features are extracted from the utterance with the following parameters: 25 ms window, 10 ms shift, 26 filters and 19 cepstral co-efficients. The parameters remain the same for MFS extraction. In case of LFS extraction, it is observed that increasing the number of filters is helpful. This is mostly due to the fact that using the same number of filterbanks in the linear arrangement would decrease the resolution in the lower frequencies. A feature with 23 cepstral co-efficients from 40 filters is observed to provide optimal performance on the development set. For the MFS and LFS based systems, $\beta = 15$ is chosen even when they are combined with MFCC for score fusion. However, $\beta = 10$ is set for MFCC-based systems.

4.1. Results

The NIST RT 2005 dataset is used as the development set. The results of speaker diarization experiments are presented in the initial columns of Table 2. The results are presented in the form of speaker error rate (SER) with speech/non-speech output obtained from groundtruth. The results on LFS and MFS provide a fair indication that Filterbank slope features, by design, are much more suited to represent speakers compared to MFCC. In the HMM/GMM system, relative improvements of 10.1% and 11.8% are obtained with respect to LFS and MFS, respectively. In the IB system, a relative improvement of 5.7% is obtained for LFS and that of 8.6% for MFS. To include time domain features, TDOA features is used. A weight of 0.8 on the primary feature and 0.2 on TDOA is observed to be optimal. When combined with TDOA, which includes information about speaker positions, a deterioration is observed when compared with the equivalent MFCC based system. The non-improvement for LFS is largely based on the deterioration on a single recording

Table 2. Results of experiments conducted on the NIST RT 05, 06, 07 and 09 datasets comparing MFCC, LFS, MFS and TDOA features and their combinations. (SER: Speaker Error Rate, R.I. : Relative Improvement). The fusion weights are mentioned in parentheses.

	Dev. set		Test set					
System/Dataset	RT05	R.I.	RT06 (SER)	R.I. (%)	RT07 (SER)	R.I. (%)	RT09 (SER)	R.I. (%)
	HMM/GMM							
MFCC	11.9	-	15.4	-	6.4	-	14.5	-
LFS	10.7	10.1	10.3	33.1	5.9	7.8	13.7	5.51
MFS	10.5	11.8	14.1	8.4	7.3	-	14.3	1.3
	IB							
MFCC	17.5	-	22.8	-	16.7	-	26.9	-
LFS	16.5	5.7	18.4	19.3	10.5	37.1	21.8	18.9
MFS	16	8.6	22.1	3.1	13	22.1	19.4	27.9
MFCC + TDOA (0.8, 0.2)	13.9	-	14.1	-	9.6	-	13.3	-
LFS + TDOA (0.8, 0.2)	16.7	-	12.8	9.2	7.2	25.0	10.9	18.0
MFS + TDOA (0.8, 0.2)	15.3	-	14.7	-	6.9	4.1	11.8	7.6
MFCC + LFS + TDOA (0.3, 0.6, 0.1)	13.6	2.1	14.1	9.2	8	16.7	10	11.2
MFCC + MFS + TDOA (0.4, 0.5, 0.1)	14.3	-	12.9	8.5	8.3	13.5	10.8	18.7
MFS + LFS + TDOA (0.4, 0.4, 0.2)	10.7	23	11.9	15.6	8.1	15.6	10.9	18.0

in the dataset (AMI_20041210-1052), while a relative improvement of 9% is observed on the rest of them. Thus, the testing on LFS is proposed to be continued. Results in [9] and [10] have shown that MFCC and MFS can complement each other. Thus, the fusion of these features along with time domain features is also studied. The results of the fusion experiments, reported as systems MFCC + LFS + TDOA, MFCC + MFS + TDOA and MFS + LFS + TDOA, are also given in Table 2. The results indicate, as expected, that fusion of MFS and LFS with MFCC has potential. In continuation with the non-improvement observed earlier when combining with TDOA, the MFCC + MFS + TDOA system performs poorer compared with the MFCC + TDOA system.

The results across the RT 06, 07 and 09 datasets presented in Table 2 indicate a clear trend - LFS and MFS based systems perform better compared with the MFCC based systems. While comparing systems based on the individual spectral features a best case relative improvement of **37.1**% is observed for LFS on the RT 07 dataset. This is **22.1**% for MFS on RT 07. For the HMM/GMM system, best case relative improvements of **33.1**% and **8.4**% are observed on RT06.

The addition of time domain features has a positive effect on the spectral features. This behavior is not surprising as LFS and MFS are linearly related to MFCC. However, the performance gains obtained on the slope based features make them better compared with the MFCC + TDOA system. Relative improvements of 9.2, 25.0 and 18.0% is observed for the LFS + TDOA when compared with the MFCC + TDOA system. In case of MFS + TDOA system, the performance on RT 06 is worse by 0.6 % absolute. In the other two datasets, improvements are observed once again.

As conventional diarization systems on meeting data tend to combine spectral features such as MFCC with the time domain features, the combination of filterbank slope features with TDOA provides can be seen to provide a much better alternative in terms of performance. This confirms our hypothesis that filterbank slope features are better suited to emphasizing speaker characteristics compared with MFCC.

The improvements observed in the systems obtained from fusion of multiple features is far superior compared to that observed in the development data set. However, as in the case of the development dataset the fusion of MFS and LFS features provide the best performance for RT 07 and RT 08 datasets while there is only a minor deterioration by 0.1% (absolute) on the RT 09 dataset. Thus, it can be concluded that combining the variants of the filterbank slope features is much more beneficial compared to combining it with MFCC. This does not however discard the possibility of combining MFCC as an additional feature.

5. CONCLUSIONS

The Filterbank slope based features, MFS and LFS, are a suitable replacement for MFCC as the primary feature in the IB based framework for speaker diarization. When tested on the benchmark NIST RT 06, 07 and 09 datasets, significant improvements are observed to corroborate this claim. In the context of speaker diarization, the improvements obtained with LFS are much better compared to that obtained with MFS. Combining them with TDOA features is observed to be helpful. The fusion of MFS and LFS provides far superior results compared to the individual fusion of these features with MFCC.

6. ACKNOWLEDGEMENTS

The authors thank the Swiss National Science Foundation for their financial support for this project through the National center of Competence in Research on "Interactive Multimodal Information Management", and more specifically through the sub-project DIMHA (Diarizing Massive Amounts of Heterogeneous Audio).

7. REFERENCES

[1] Jitendra Ajmera and Chuck Wooters, "A robust speaker clustering algorithm," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2003*. IEEE, 2003, pp. 411–416.

[2] Chuck Wooters and Marijn Huijbregts, "The icsi rt07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*. 2008, pp. 509–519, Springer.

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

[4] Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[5] Tomi Kinnunen and B Haizhou Li, "An overview of text-independent speaker recognition: from features to supervectors," January 2010, vol. 52(1), pp. 12–40, Speech Communication.

[6] Timo Becker others, "Speaker verification based on formants using Gaussian mixture models," September 2008, pp. 1505–1508, In Proc. of Interspeech.

[7] Davis, S. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition," 1980, vol. 28(4), pp. 357–366, IEEE Trans. Acoust., Speech, Signal Processing.

[8] Hema A. Murthy et al., "Robust text-independent speaker identification over telephone channels," 1999, vol. 7(5), pp. 554–568, IEEE Trans. Speech and Audio Processing.

[9] S. Madikeri and Hema A Murthy, "Mel filter bank energy-based slope feature and its application to speaker recognition," 2011, pp. 1–4, In Proc. of National Conference on Communication.

[10] Srikanth R Madikeri and Hema A Murthy, "Effect of feature warping and decorrelation on Mel filterbank slope for speaker recognition," July 2012, pp. 1–5, In Proc. of SPCOM.

[11] Xinhui Zhou, Daniel Garcia-Romero, Ramani Duraiswami, Carol Espy-Wilson, and Shihab Shamma, "Linear versus Mel frequency cepstral coefficients for speaker recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2011, pp. 559–564.

[12] S Furui, "Cepstral analysis technique for automatic speaker verification," 1981, vol. 18, pp. 859–872, Pattern Recognition Letters.

[13] Jose M Pardo, Xavier Anguera, and Chuck Wooters, "Speaker diarization for multiple distant microphone meetings: mixing acoustic features and inter-channel time differences.," in *In Proc. of INTERSPEECH*, 2006.

[14] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard, "An information theoretic combination of MFCC and TDOA features for speaker diarization," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 431–438, 2011.

[15] Deepu Vijayasenan, *An information theoretic approach to speaker diarization of meeting recordings*, Ph.D. thesis, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE, 2010.

[16] Xavier Anguera, C Woofers, and Javier Hernando, "Speaker diarization for multi-party meetings using acoustic fusion," in *IEEE Workshop onAutomatic Speech Recognition and Understanding, 2005*. IEEE, 2005, pp. 426–431.

[17] Noam Slonim and Naftali Tishby, "The power of word clusters for text classification," in *23rd European Colloquium on Information Retrieval Research*, 2001, vol. 1.

[18] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in *IEEE Workshop on Automatic Speech Recognition & Understanding, 2007*. IEEE, 2007, pp. 250–255.

[19] Thomas M Cover and Joy A Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[20] Xavier Anguera, "Beamformit (the fast and robust acoustic beamformer)," http://www.xavieranguera.com/beamformit/.