Modulation Spectrogram Features for Improved Speaker Diarization

Oriol Vinyals [1,2], Gerald Friedland [2]

[1] University of California, Berkeley, CA
[2] International Computer Science Institute, Berkeley, CA

Abstract

We propose the use of modulation spectrogram features in speaker diarization. These features carry longer term characteristics of the acoustic signals than the widely used MFCCs, thus providing potential improvement by using both features in combination. Using the state-of-the-art ICSI speaker diarization system, an improvement of 20.77% relative DER is obtained on the NIST Rich Transcription 2007 task with respect to the MFCC only system.

Index Terms: modulation spectrogram, speaker diarization

1. Introduction

A small set of standard features, such as MFCC or PLP, in different dimensionalities tends to be used for almost any speech related task. In speaker diarization, the task is to segment audio into speaker-homogeneous regions with the goal of answering the question "Who spoke when?". Current systems usually rely on the combination of Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs) [1].

In the following article, we show that despite the dominance of short-term cepstral features in speaker recognition, a range of longer term features can provide significant information for speaker discriminability. As suggested in [2], looking at patterns derived from a larger segment of speech can reveal individual characteristics of the speakers' voices as well as their speaking behavior, which cannot be captured by frame-based short-term cepstral analysis.

The modulation spectrogram (hereafter referred to as MSG) was first introduced by Kingsburey et al. [3]. These features were proposed to improve speech recognition in reverberant audio conditions. However, other authors [4] have noticed that these features also carry speaker specific information. In the task of speaker diarization we show that, due to the different characteristics that both MFCC and MSG features extract from the spectrogram, the features can help solving the diarization task.

We use the combination of these features to achieve a 21 % relative improvement of the Diarization Error Rate (DER). The results were measured on both the NIST RT06 and RT07 test and evaluation datasets and are compared to the top performing system as of the NIST RT evaluation in 2007.

The article is structured as follows. Section 2 surveys related work in speaker diarization and the use of alternative features, i.e. non short-term cepstral features, in speaker recognition. Section 3 describes the MSG features and its potential application to Speaker Diarization. Section 4 presents our baseline, namely the ICSI speaker diarization system. Section 5 discusses the actual integration of the features into the ICSI speaker diarization system and presents the experimental results on different NIST benchmarks. Section 6 summarizes the article and presents future work.
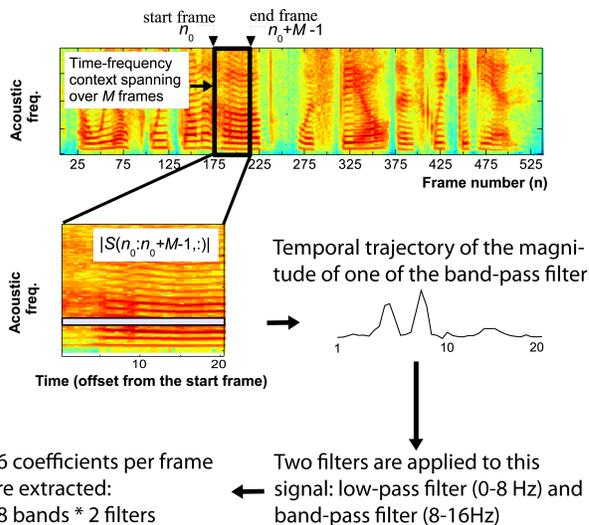
Figure 1: A diagram showing the process of MSG extraction. After computing the spectrogram, a window of 21 frames is used. For each of the 18 frequency bands, the signal is filtered using a low-pass and a band-pass filter, and two values for each band is obtained (to sum a total number of features of 36). Diagram modified from [4].

2. Related work

Most state-of-the-art speaker diarization systems, including the ICSI Speaker Diarization engine, use a one stage approach, i.e. the combination of agglomerative clustering with Bayesian Information Criterion (BIC) [5] and Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs) [1] (see Section 4). While many different machine-learning strategies have been explored around this basic idea, exploration of features for use in this approach has been limited to varying the dimensionality of the cepstral features. [6] proposed a framework for combining MFCC features with PLP. In [7], the authors introduce the use of delay features —that is, the delay between signals between different microphones in an array— to improve the DER. These kinds of features can only be extracted in the multiple distant microphone (MDM) condition, where several far-field signals are available. However, this paper concentrates on improving the single distant microphone (SDM) case using acoustic features only. The use of MSG features has been explored in the related task of speaker identification, where these features have been used in isolation and together with MFCCs to model speakers [4].
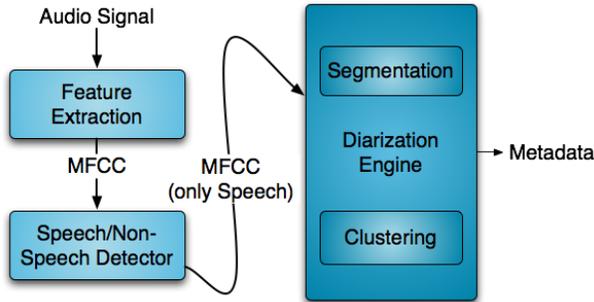
Figure 2: A diagram illustrating the baseline ICSI Speaker Diarization Engine which is described in Section 4. The audio signal, given as cepstral features (MFCC) undergoes a two stage process: Speech/Non-Speech filtering, and one-step segmentation and clustering.

3. The modulation spectrogram

The modulation spectrogram provides an alternative and complementary representation of the speech signal with a focus on temporal structure. Developed by Kingsbury et al. and detailed in [3], the modulation spectrogram represents a filtered version of the spectrogram of a speech signal. The spectrogram of the signal is computed using an FFT with step size of 10 ms and an analysis window of 25 ms. In contrast to MFCC features, where for each frame the DCT coefficients of the Mel log-FFT amplitudes are computed, the MSG analyzes the spectrogram using 18 bands from 0 to 8 KHz, filtering the resulting 18 temporal signals with two different filters: a 0-8 Hz filter and an 8-16 Hz filter. For each frame, the MSG features capture the low-pass and band-pass behavior of the spectrogram of the signal within each of the 18 subbands, resulting in a total of 36 features per frame.

As we stated, in contrast to MFCCs, the modulation spectrogram provides information about longer temporal phenomena as it uses 0.21 seconds of analysis to extract the features. Thus, we expect that, jointly with MFCCs, this representation of the spectrum of the signal will be richer and perform better in the task of speaker diarization.

4. Baseline ICSI speaker diarization engine

As explained in Section 1, the goal of speaker diarization is to segment audio into speaker-homogeneous regions with the ultimate goal of answering the question "Who spoke when?" [1]. In contrast to speaker recognition or identification, speaker diarization attempts to use no prior knowledge. This means, usually, no specific speaker models are trained for the speakers that are to be identified in the recording. A speaker diarization system conceptually performs three tasks: First, discriminate between speech and non-speech regions, second, detect speaker changes to segment the audio data, third, group the segmented regions together into speaker-homogeneous clusters. Some systems unify the two last steps into a single one, i.e. segmentation and clustering is performed in one step. Over the years, many different algorithms have been developed in the speech community. A summary can be found in [8].

The speaker diarization engine developed at ICSI uses an agglomerative clustering approach to perform both the segmentation of the audio track into speaker-homogeneous time segments and the grouping of these segments into speaker-homogeneous clusters in one step.
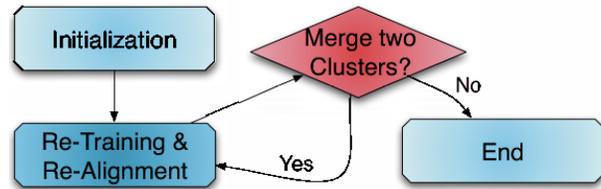

Figure 3: The agglomerative clustering approach of the ICSI Speaker Diarization Engine as explained in Section 4. Retraining and re-segmentation ends when no more models can be merged as of the BIC score. At the end, the number of clusters is hoped to be equal to the number of speakers.

A speech/non-speech detector is used to filter out regions that do not contain speech. This is usually either done using threshold-based heuristics (for example, using a pitch detector and only considering voiced-regions of the speech) or using a trained approach (for example by training Gaussian Mixture Models on speech and noise, respectively).

The audio track is usually processed as 19th-order MFCC features using a frame size of 10 ms. The non-speech regions are excluded from the agglomerative clustering. Figure 2 illustrates the big picture of the algorithm.

The algorithm is initialized using a much higher number of clusters than speakers assumed in the audio track. Let this number be k. An initial segmentation is generated by uniformly partitioning the audio track into k segments of the same length. Using the initial segmentation, k Gaussian Mixture Models are trained. As classifications based on 10 ms frames are very noisy, a minimum duration of 2.5 seconds is assumed for each speech segment, and then Viterbi alignment is performed. The algorithm then performs the following loop:

- Re-Segmentation: Run the Viterbi alignment to find the optimal path of frames and models, with a minimum duration of 2.5 seconds.

- Re-Training: Given the new segmentation of the audio track, compute new Gaussian Mixture Models for each of them.

- Cluster Merging: Given the new Gaussian Mixture Models, try to find the two models that most likely represent the same speaker. This is done by computing the BIC score (Bayesian Information Criterion) of each of the models and the BIC score of a new GMM trained on the merged segments for two clusters. If the BIC score of the merged Gaussian Mixture Model is smaller than or equal to the sum of the individual BIC scores, the two models are merged and the algorithm loops at the re-segmentation using the merged Gaussian Mixture Model. If no pair is found, the algorithm stops.

Figure 3 illustrates the steps of the algorithm. A more detailed description can be found in [9, 10, 11].

The output consists of meta-data describing speech segments in terms of starting time, ending time, and speaker cluster name. This output is usually evaluated against manually annotated ground truth segments. A dynamic programming procedure is used to find the optimal one-to-one mapping between the hypothesis and the ground truth segments so that the total overlap between the reference speaker and the corresponding mapped hypothesized speaker cluster is maximized. The difference is expressed as Diarization Error Rate which is defined by

NIST[1]. The Diarization Error Rate (DER) can be decomposed into three components: misses (speaker in reference, but not in hypothesis), false alarms (speaker in hypothesis, but not in reference), and speaker-errors (mapped reference is not the same as hypothesized speaker).

The ICSI speaker diarization system has competed in the NIST evaluations of the past several years and established itself well among state-of-the-art systems[2].

The current official score is 21.74 % DER for the single-microphone case (RT07 evaluation set). This error can be decomposed in 6.8 % speech/non speech error and 14.9 % speaker clustering error. The speaker error includes all incorrectly classified segments, including overlapped speech and very short segments.

5. Integration into the ICSI diarization system and experimental results

In this section, we describe several techniques using the MSG features on the ICSI Speaker Diarization system, as well as an extensive error analysis on where and why the MSG features helped in diarization.

The test setup is as follows: MSG features are extracted every 10 ms. We extract 18 subbands of the spectrum and use two filters (0-8Hz and 8-16Hz) to perform the modulation of each band. Thus, the dimensionality per frame is 36. In all our experiments the speech/non-speech detection was the same as in the RT07 submission. The segmenter performs iterative training and re-segmentation of the audio into three classes: speech, silence, and audible nonspeech. To bootstrap the process, an initial segmentation is created with an HMM trained on broadcast news data. A detailed description can be found in [11].

To perform an analysis of the performance and improvement of the MSG features, all the following experiments were performed on the NIST Rich Transcription 07 meeting data (for the single distant microphone condition). It contains eight meetings recorded in several geographic locations with differing numbers of people (this set is named hereafter as Eval07). Even though the diarization task is unsupervised, there are some parameters like the amount of gaussians to start with or the weight for several feature stream that should be learned. Another set of 21 meetings, based on NIST meeting data of previous years, is used for parameter selection (named hereafter as Dev07).

The approach we propose for combining several features is similar to the one in [7]. In particular, the diarization engine performs a maximization of an objective function based on the likelihood of the observed data given the model (in our case, the model is an ergodic HMM). We can then define the combined likelihood for the emission probabilities as:

$$p(x_{MFCC}, x_{MSG}|\theta_i) =$$
$$= p(x_{MFCC}|\theta_{i1})p(x_{MSG}|\theta_{i2})^{\alpha}$$

where θ_{i1} represent the parameters of cluster i using the MFCC observed data and θ_{i2} represents the parameters using the MSG features. The model we use for the emission probabilities are GMMs where the number of components varies for each feature stream. Note that there is an assumption of independence between the two set of features. Finally, as we observed that MFCC features tend to perform better than MSG features, we

[1] http://nist.gov/speech/tests/rt/rt2004/fall

[2] Unfortunately, we are not allowed to present any ranking. Please refer to the NIST website for further information: http://www.nist.gov/speech/tests/rt/rt2007/

Features used	DER Dev07	DER Eval07
MFCC	17.57	21.74
MFCC + MSG	13.26	17.28

Table 1: DER on the Dev07 and Eval07 using ICSI diarization with MFCC and MFCC+MSG.

used the α parameter to modify the confidence given to each feature stream. As α is decreased, the likelihoods of the MSG features given each class become more similar in values (the extreme case where $\alpha = 0$ maps all the likelihoods to 1). Hence, the effect of this parameter is to give different confidence value to each feature stream.

The Dev07 set of meetings is used to find the optimal value of α. The initial number of gaussians of the MSG features is set to 1. The rest of the parameters of the system are the same as the ones used in the RT07 evaluation (16 initial clusters and 5 gaussians per cluster for the MFCC feature vector). Table 1 shows the results on the development set with the optimal value of $\alpha = 0.1$. The use of the MSG features resulted in a 24.53 % relative improvement of the DER on the Dev07 and 20.77 % on the Eval07 set. The results are compared to the system that competed in the 2007 NIST RT evaluations.

5.1. Further analysis

Figure 4 shows the DER evolution per each algorithm stage of the baseline system vs. our combined approach. As can be seen, the MSG features contribute especially in the last stages of the agglomerative clustering approach. Since the α value found using the development set was low, the effect of the MSG features on the first iterations will be unnoticed by the algorithm: the MFCCs alone are able to refine the segments and merge clusters that belong to the same speaker.

As the clusters are merged, the average length increase and thus the long-term dependencies that the MSG features extract are more robust. Moreover, in the last stages of the algorithm the clusters are more pure (each cluster contains speech from only one person), and, as a consequence, the discriminative power that the MSG features have is amplified by the fact that the clusters represent speech from mostly one person. If we observe the tail of Figure 4, it is clear that the information provided by the MSG features is quite useful in the last stages, where otherwise the MFCCs were not able to correct some errors, and thus providing the 24.53 % final relative improvement on the Dev07 set.

6. Conclusions and future work

In this paper we presented the use of the modulation spectrogram as an additional stream of features to improve speaker diariziation. In combination with the commonly used MFCCs, we observed a significant improvement of our system with respect to the official submission on the last NIST RT 2007 evaluation task. This result was also verified on a larger set of meetings, which we call Dev07, which contains 21 meetings from previous evaluations. Error analysis is performed to confirm where and why the usage of this complementary features, which capture longer term dependencies than MFCCs, affect our current system.

In the future, we would like to explore other methods to combine the features. In particular, it seems reasonable to set the α value dynamically per iteration instead of statically as it is now: as seen in the error analysis, it is in the later stages of the
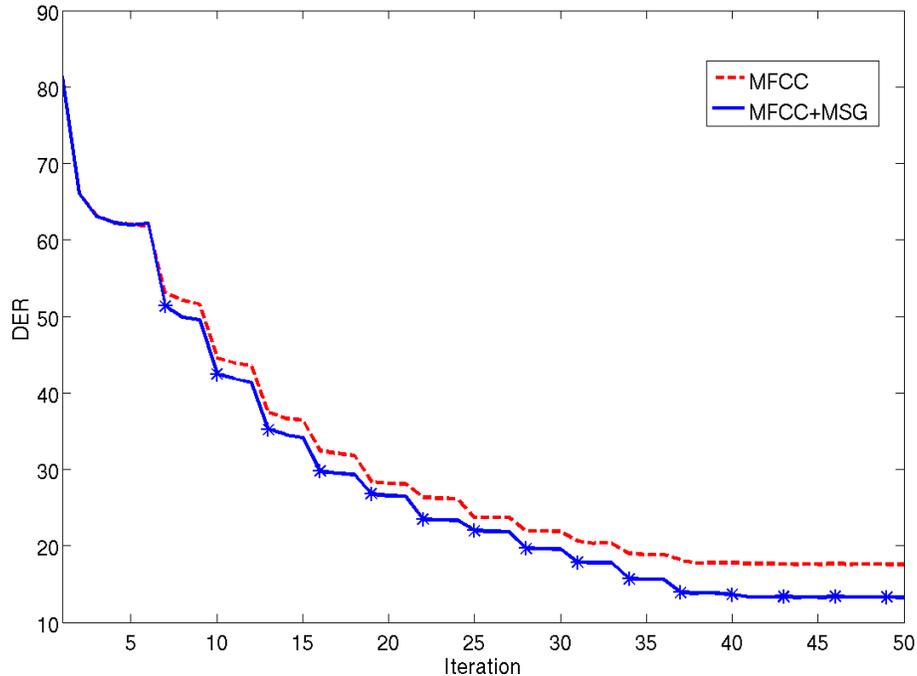
Figure 4: The average DER per iteration of the ICSI diarization engine across the Dev07 meetings. There are 16 clusters and so the potential number of cluster merging is 15. The asterisk denotes a merging of two clusters while other iterations are re-alignment of the models and the data.

agglomerative clustering approach that the MSG features help most, and so setting the weight higher might provide in further improvement.

7. Acknowledgements

8. References

[1] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings of the IEEE ICASSP*, 2005.

[2] E. Shriberg, "Higher-level features in speaker recognition." in *Speaker Classification (1)*, ser. Lecture Notes in Computer Science, C. Mller, Ed., vol. 4343. Springer, 2007, pp. 241–259.

[3] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, pp. 117–132, August 1998.

[4] T. Kinnunen, K. Lee, and H. Li, "Dimension Reduction of the Modulation Spectrogram for Speaker Verification," in *Proceedings of Speaker and Language Recognition Workshop, IEEE Odyssey*, 2008.

[5] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proceedings of DARPA speech recognition workshop*, 1998.

[6] A. Gallardo-Antolin, X. Anguera, and C. Wooters, "Multi-Stream Speaker Diarization Systems for the Meetings Domain," in *Proceedings of Interspeech*, 2006.

[7] ——, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Transactions on Computers*, vol. 56, pp. 1212–1224, September 2007.

[8] X. Anguera, "Robust speaker diarization for meetings," Ph.D. dissertation, Technical University of Catalonia, Barcelona, Spain, December 2006.

[9] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proceedings of the IEEE Automatic Speech Recognition Understanding Workshop*, 2003.

[10] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *Proceeding of the NIST MLMI Meeting Recognition Workshop, Edinburgh*, 2005.

[11] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Proceedings of the RT07 Meeting Recognition Evaluation Workshop*, 2007.