An Auditory Based Modulation Spectral Feature
for Reverberant Speech Recognition

HariKrishna Maganti, Marco Matassoni

Fondazione Bruno Kessler - Center for Information Technology - IRST
via Sommarive 18, 38050 Povo, Trento, Italy
{maganti, matasso}@fbk.eu

Abstract

In this paper, an auditory based modulation spectral feature is presented to improve automatic speech recognition performance in presence of room reverberation. The solution is based on extracting features from auditory processing characteristics, specifically gammatone filtering based long-term modulation spectral features to reduce sensitivity to environmental noise and further preserve the important speech intelligibility information in the speech signal essential for ASR. Experiments are performed on Aurora-5 meeting recorder digit task recorded with four different microphones in hands-free mode at a real meeting room. For comparison purposes the recognition results obtained using standard ETSI basic and advanced front-ends and conventional features with standard feature compensation are tested. The experimental results reveal that the proposed features provide reliable and considerable improvements with respect to the state of the art feature extraction techniques.

Index Terms: auditory processing, gammatone filtering, modulation spectrum, reverberation, speech recognition

1. Introduction

A challenging issue in robust hands-free automatic speech recognition is reliable performance of the system under adverse acoustic conditions. Apart from the additive background noise, another important source of degradation is caused by reverberation produced in the acoustic environment. The speech signal acquired in a reverberant room can be modeled as the convolution of the speech signal with the room impulse response,

\[ x(k) = s(k) * h(k) \]  \hspace{1cm} (1)

where \( x(k) \) is the degraded speech signal, \( s(k) \) represents the clean signal, \( h(k) \) is the impulse response of the room. The impulse response depends upon the distance between the speaker and the microphone, and room conditions, such as movement of people in the room, clapping, opening or closing doors, etc. Thus extracting robust features which can handle various room impulse responses is a complex and challenging task.

Current approaches for improving robustness of noisy speech can be performed at signal, feature or model level. Speech enhancement techniques aim at improving the quality of speech signal captured through single microphone or microphone array [1, 2]. Robust acoustic features attempt to represent parameters less sensitive to noise by modifying the extracted features. Common techniques include cepstral mean normalization (CMN) and cepstral mean subtraction and variance normalization (CMSVN) and relative spectral (RASTA) filtering [3, 4]. Model adaptation approach modify the acoustic model parameters to fit better with the observed speech features [1, 5].

Alternately, various auditory processing based approaches are used to improve robustness of noisy speech, which include features based on gammatone filtering and modulation spectrum [6, 7, 8, 9]. Gammatone filter bank processing is designed to model human cochlear filtering and is shown to provide robustness in adverse noise conditions for speech recognition tasks [6, 7]. In [6], gammatone based auditory front-end exhibited robust performance compared to traditional front-ends based on MFCC, PLP and standard ETSI frontend. For large vocabulary speech recognition tasks, the performance of these features have been competitive with standard features like MFCC and PLP [7]. Modulation spectrum based features computed over longer windows have been effective in measuring speech intelligibility in reverberant environments [8]. The slow variations in spectrum, computed over the large frame size have shown to improve recognition accuracy for moderate reverberant conditions [9].

In this work, an alternate approach for feature extraction combining both gammatone filtering and modulation spectrum, is presented. Gammatone frequency resolution reduces the ASR system sensitivity to environmental reverberant signal attributes and improve the speech signal characteristics. Further, long-term modulation preserves the linguistic information in the speech signal, improving the accuracy of the system. The features derived from the combination are used to provide robustness, particularly in the context of mismatch between training and testing reverberant environments. The studied features are shown to be reliable and robust to the effects of the hands-free recordings in the reverberant meeting room. The effectiveness of the proposed features is demonstrated with experiments which use real-time reverberant speech acquired through four different microphones. For comparison purposes the recognition results obtained using standard ETSI basic and advanced front-ends [11, 12] and conventional features with standard feature compensation techniques applied are tested, and usage of the proposed features proved to be efficient.

The paper is organized as follows: Section 2 gives an overview of the auditory inspired features, including gammatone filter bank processing and modulation spectrum processing. Section 3 describes the feature extraction. Section 4 presents database description, experiments and results. Finally, Section 5 concludes the paper.

2. Auditory Based Features

In this section, a general overview of auditory features based on gammatone filter bank and modulation spectrum is presented.
2.0.1. Gammatone Filter Bank

Gammatone filters are linear approximation of physiologically motivated processing performed by the cochlea [13]. It is commonly used in modeling human auditory system and consists of a series of bandpass filters. In time domain, the filter is defined by the following impulse response

\[ g(t) = at^{n-1} \cos(2\pi f_c t + \phi)e^{-2\pi bt} \]  

(2)

where \( n \) is the order of the filter, \( b \) is the bandwidth of the filter, \( a \) is the amplitude, \( f_c \) is the filter center frequency and \( \phi \) is the phase.

The frequency center frequencies and bandwidths are derived from the filter’s Equivalent Rectangular Bandwidth (ERB) as detailed in [13]. In [14], Glasberg and Moore relate center frequency and the ERB of an auditory filter as

\[ \text{ERB}(f_c) = 24.7 \left( \frac{4.37 f_c}{1000} + 1 \right) \]  

(3)

The filter output of the \( m \)th gammatone filter, \( X_m \), can be expressed by

\[ X_m(n) = x(n) \ast h_m(n) \]  

(4)

where \( h_m(n) \) is the impulse response of the filter.

The frequency response of the 32-channel gammatone filterbank is as shown in Fig. 1.

Figure 1: Frequency response for the 32-channel gammatone filterbank.

The long-term modulations examine the slow temporal evolution of the speech energy with time windows in the range of 160 - 800 ms, contrary to the conventional short-term modulations studied with time windows of 10 - 30 ms which capture rapid changes of the speech signals. Generally, the modulation spectrum is computed as following: Speech signal \( X(k) \) is segmented into frames by a window function \( w(k, t) \), where \( t \) is frame number. Short-time Fourier transform of the windowed speech signal \( X(t, f) \) is calculated as

\[ Y(t, f) = \sum_{i=-\infty}^{\infty} X(f - i)W(i, t) \]  

(5)

The modulation spectrum \( Y_{m}(f, g) \) is obtained by applying Fourier transform on the running spectra, obtained by taking absolute values \( |Y(t, f)| \) at each frequency, expressed as

\[ Y_m(f, g) = FT[|Y(t, f)|]_{t=1...T} \]  

(6)

where \( T \) is the total number of frames and \( g \) is the modulation frequency. The relative prominence of slow temporal modulations is different at various frequencies, similar to perceptual ability of human auditory system. Most of the useful linguistic information is in the modulation frequency components from the range between 2 and 16 Hz, with dominant component at around 4 Hz [15, 10]. In [10], it has been shown that for noisy environments, the components of the modulation spectrum below 2 Hz and above 10 Hz are less important for speech intelligibility, particularly the band below 1 Hz contains mostly information about the environment. Therefore the recognition performance can be improved by suppressing this band in the feature extraction.

The comparative waveforms, gammatonegrams and modulation spectrum plots of the clean and noisy versions of the same speech utterance are as shown in Fig. 3. From Fig 3.f, some of the important characteristics of the modulation spectrum can be observed. The important information of speech is concentrated in the area from 2 Hz and 16 Hz, particularly 2 Hz and 4 Hz contain crucial information related to the variation of phonemes.

3. Feature Extraction

The block schematic for the gammatone modulation spectrum based feature extraction technique is shown in Fig. 2. The speech signal first undergoes pre-emphasis, which flatten the frequency characteristics of the speech signal. The signal is then processed by a gammatone filterbank which uses 32 frequency channels equally spaced on the equivalent ERB scale as shown in Fig. 1. The computationally effective gammatone filter bank implementation as described in [16] is used. The gammatone filter bank transform is computed over \( L \) ms and the segment is shifted by \( n \) ms. The log magnitude resulting coefficients are then decorrelated by applying a discrete cosine transform (DCT). The computations are made over all the incoming signal, resulting in a sequence of energy magnitudes for each band sampled at 1/n Hz. Then, frame by frame analysis is performed and a \( N \)-dimensional parameter is obtained for

Figure 2: Processing stages of the gammatone modulation spectrum feature. DCT and FFT represent discrete cosine transform and Fourier transform. \( f \) indicate coefficient.
each frame. The modulation spectrum of each coefficient which is defined as the Fourier transform of its temporal evolution is computed. In each band, the modulations of the signal are analyzed by computing FFT over the 10 ms Hamming window and the segment is shifted by 10 ms. The energies for the frequencies between the 2 - 16 Hz, which represent the important components for the speech signal are computed.

For example, if the given signal \( x(t) \) is sampled at 8 kHz, a first-order high pass pre-emphasis filter is applied and short segments of speech are extracted with a 25 ms rectangular window. The window is shifted by 10 ms which corresponds to a frame rate of 100 Hz. Each speech frame is then processed by a 32-channel gammatone filterbank. The 32 logarithmic gammatone spectral values are transformed to the cepstral domain by means of a DCT. Thirteen cepstral coefficients \( C0 \) to \( C12 \) are calculated. \( C0 \) is replaced by logarithm of the energy computed from the speech samples. The modulation spectrum of each coefficient, (sampled at 100Hz) is calculated with a 160 ms window, shifted by 10 ms. Thirteen coefficients \( C13 \) to \( C26 \) are further extracted.

### 4. Experiments and Results

The experiments are conducted on a subset of the Aurora-5 corpus - meeting recorder digits. The data comprise real recordings in a meeting room, recorded in a hands-free mode at the International Computer Science Institute in Berkeley. The dataset consists of 2400 utterances from 24 speakers, with 7800 digits in total. The speech was captured with four different microphones, placed at the middle of the table in the meeting room. The recordings contain only a small amount of background noise, but have the effects of hands-free recording in the reverberant room. There are four different versions of all utterances recorded with four different microphones, with recording levels kept low.

To evaluate the performance, a full HTK based recognition system is used. The HMM-based recognizer architecture specified for use with the Aurora 5 database is used [18]. The training data is downsampled version of clean TIDigits at a sampling frequency of 8 kHz, with 8623 utterances. There are eleven whole word HMMs each with 16 states and with each state having four Gaussian mixtures. The \( \text{sil} \) model has three states and each state has four mixtures.

Table 1 shows the results of baselines in % word accuracies for meeting recording digits recorded with four different microphones, labeled as 6, 7, E, and F. The average performance of four microphones for different features is shown at the last column of the table. ETSI-2 and ETSI-1 correspond to the standard advanced and basic front-ends as described in [11, 12]. MFCC and PLP are the standard 39-dimensional Mel frequency and Perceptual linear prediction features along with their delta and acceleration derivatives. A PLP is considered instead of the RASTA-PLP because the performance was similar in both the cases. Also, spectral subtraction pre-processing [2] with MFCC and PLP, did not show any improvement, and hence are not indicated in the table.

From Table 1, It is evident that the advanced ETSI front-end has highest error rates compared to the basic ETSI front-end, and MFCC and PLP. This clearly demonstrates that for reverberant environments the advanced ETSI front-end is not effective as compared to its performance in the presence of additive background noise. It can be inferred that the techniques applicable for additive background noise removal are not suitable to handle reverberant conditions. The basic ETSI front-end and MFCC have similar performance, owing to the similarity in extraction of their features. Among the baselines, PLP has the best performance indicating the robustness of the features, and is used for further comparisons.

Table 2 shows the performance of PLP along with Mean Variance ARMA (MVA) processing, indicated as PLP-MVA [17]. The filter of order 3 is used for ARMA filtering. GFCC indicate 39-dimensional Gammatone Frequency Cepstral Coefficients (GFCC) along with their delta and acceleration derivatives. The GFCC features are extracted in a similar way as reported in Section 3 with \( C0 \) to \( C12 \) being the corresponding cepstral coefficients. GFMC-26 indicate Gammatone Fre-
quency Modulation Spectral based Cepstral (GFMC) features derived in a same way as GFCC but appended with modulation spectral features corresponding to C13 to C26 as discussed in Section 3. GFMC are the extension of GFMC-26, which also include derivatives of the modulation spectral features.

From Tables 1 and 2, it can be observed that the MVA processing was beneficial for PLP. It can be seen that the GFCC features were effective, performing better than any of the baseline systems, with a performance similar to that of PLP with additional MVA processing. This is consistent with earlier studies which have shown that Gammatone based features exhibit robust performance compared to MFCC, PLP features and ETSI frontends [6, 7]. It can also be seen that the performance of GFMC-26 with reduced feature set (26) is comparable to PLP-MVA and GFCC features with standard feature set (39).

It can also be observed that the performance of GFMC is the best among all the baselines and features compared, and consistent across all the channels. Fig. 4 shows performance of GFMC computed with various combinations of modulation frequencies. We can observe that the best performance can be obtained with combinations below 16 Hz and performance degrades with combinations above 16 Hz, which is also consistent with results in [10]. The results indicate that the gammatone frequency resolution was effective in reducing system sensitivity to reverberation and improved the speech signal characteristics. The emphasis on slow temporal changes in the spectral structure of long-term modulations preserved the required speech intelligibility information in the signal which further improved the accuracy of the system.

![Figure 4: Performance of various modulation frequencies.](image)

## 5. Conclusions

The paper has presented auditory inspired modulation spectral features for improving ASR performance in presence of room reverberation. The features were evaluated on Aurora-5 meeting recorder digit task recorded with four different microphones in hands-free mode at a real meeting room. Results were compared with standard ETSI basic and advanced front-ends and conventional features with feature compensation. The results show that the proposed features perform consistently better both in terms of robustness and reliability.

Our study also raised number of issues, including improvement of standard additive noise removal techniques to deal with reverberation condition. The present work was limited to handle reverberant conditions, without considering additive noise or transmission channel effects on speech. For the future, we like to investigate these issues to efficiently deal with real world noisy speech, and evaluate performance of proposed features on large vocabulary tasks and also include appropriate model adaptation techniques.

## 6. References


### Table 2: Word recognition accuracies (%) for different feature extraction techniques on four different microphones.

<table>
<thead>
<tr>
<th>Channel</th>
<th>6</th>
<th>7</th>
<th>E</th>
<th>F</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP-MVA</td>
<td>83.3</td>
<td>76.3</td>
<td>75.9</td>
<td>80.7</td>
<td>79.0</td>
</tr>
<tr>
<td>GFCC</td>
<td>83.5</td>
<td>76.5</td>
<td>76.4</td>
<td>82.3</td>
<td>79.1</td>
</tr>
<tr>
<td>GFMC-26</td>
<td>80.8</td>
<td>75.9</td>
<td>76.2</td>
<td>79.7</td>
<td>78.1</td>
</tr>
<tr>
<td>GFMC</td>
<td>88.6</td>
<td>83.7</td>
<td>83.5</td>
<td>87.2</td>
<td>85.6</td>
</tr>
</tbody>
</table>