Auditory Processing Inspired Robust Feature Enhancement for Speech Recognition

Hari Krishna Maganti and Marco Matassoni

Fondazione Bruno Kessler, Center for Information Technology - IRST
via Sommarive 18, 38123 Povo, Trento, Italy

Abstract. The performance of Mel-frequency cepstrum based automatic speech recognition system significantly degrade in noisy environments. In this article, the feasibility of utilizing the bio-inspired auditory features to improve noise robustness is investigated. The features are based on auditory characteristics, which include gammatone filtering and modulation spectral processing to emulate the mechanisms performed in the cochlea and middle ear aimed to improve robustness in human ear. The robust noise resistant features that emulate cochlea frequency resolution are extracted by gammatone filtering. And then a long-term modulation spectral processing, which preserves speech intelligibility in the signal is performed. Compared and discussed are the features based on the performance on Aurora5 database, comprising the meeting recorder digit task recorded with four different microphones in a hands-free mode at a real meeting room and living room and office room simulated data corrupted with different levels of additive noises. The performance of these features is also investigated for CHiME challenge, aiming at speech separation and recognition in noise background that has been collected from a real family room using binaural microphones. The experimental results show that the proposed features provide considerable improvement with respect to the standard feature extraction techniques for both the versions of the database.

1 Introduction

A significant trend in ubiquitous computing is to facilitate the user to communicate and interact naturally with concerned applications. Speech is an appealing mode of communication for such applications. The human-machine interaction using automatic speech processing technologies is a diversified research area, which has been investigated actively [1,2,3].

Speech acquisition, processing and recognition in a non-ideal acoustic environments are complex tasks due to presence of unknown additive noise, reverberation and interfering speakers. Additive noise from interfering noise sources, and convolutive noise arising from acoustic environment and transmission channel characteristics contribute to a degradation of performance in speech recognition systems. This article addresses the problem of robustness of automatic speech recognition (ASR) systems due to convolutive noise by modeling techniques performed by cochlea in human auditory processing system.

The influence of additive background noise on the speech signal can be expressed as

$$y(t) = x(t) + n(t) \tag{1}$$

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 205–218, 2013.

where $y(t)$ is the degraded speech signal, $x(t)$ represents the clean signal, $n(t)$ is the additive noise, which is uncorrelated with the speech signal and unknown. Different techniques have been proposed based on voice activity detection based noise estimation, minimum statistics noise estimation, histogram and quantile based methods, and estimation of the posteriori and a priori signal-to-noise ratio [4]. In Ephraim and Cohen [5], various approaches to speech enhancement based on noise estimation and spectral subtraction are discussed.

Apart from the stationary background noise, another important source of degradation is caused by reverberation produced in acoustic environment. The speech signal acquired in a reverberant room can be modeled as convolution of the speech signal with the room impulse response,

$$y(t) = x(t) * h(t) \tag{2}$$

where $y(t)$ is the degraded speech signal, $x(t)$ represents the clean signal, $h(t)$ 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. A variant of spectral subtraction has been proposed in [6] to enhance speech degraded by reverberation.

In general to improve robustness of the noisy speech, processing 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 [7,8]. 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 [2,9]. Model adaptation approach modify the acoustic model parameters to fit better with the observed speech features [7,10].

Performance of the human auditory system is more adept at noisy speech recognition. Auditory modeling, which simulates some properties of the human auditory system have been applied to speech recognition system to enhance its robustness. The information coded in auditory spike trains and the information transfer processing principles found in the auditory pathway are used in [11,12]. The neural synchrony is used for creating noise-robust representations of speech [12]. The model parameters are fine-tuned to conform to the population discharge patterns in the auditory nerve which are then used to derive estimates of the spectrum on a frame-by-frame basis. This was extremely effective in noise and improved performance of the ASR dramatically. Various auditory processing based approaches were proposed to improve robustness [13,14,15] and in particular, the works described in [12,16] were focused to address the additive noise problem. Further, in [17] a model of auditory perception (PEMO) developed by Dau et al. [15] is used as a front end for ASR, which performed better than the standard Mel-frequency based cepstral coefficients (MFCC) for an isolated word recognition task. Principles and models relating to auditory processing, which attempt to model human hearing to some extent have been applied for speech recognition in [9,18].

The important aspect in a speech recognition system is to have abstract representation of highly redundant speech signal, which is achieved by frequency analysis. The cochlea and hair cells of the inner ear perform spectrum analysis to extract relevant features. The models for auditory spectrum analysis are based on filterbank design, which are usually characterized by non-uniform frequency resolution and non-uniform bandwidth on linear scale. Examples include popular speech analysis techniques, namely Mel frequency cepstrum and perceptual linear prediction which try to emulate human auditory perception. Other important processing is based upon Gammatone filter bank, which is designed to model human cochlear filtering and is shown to provide robustness in adverse noise conditions for speech recognition tasks [16,19]. In [16], 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 [19]. Another important psychoacoustic property is modulation spectrum of speech, which is important for speech intelligibility. The relative prominence of slow temporal modulations is different at various frequencies, similar to perceptual ability of human auditory system. Particularly, 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 [20,21,18]. Modulation spectrum based features computed over longer windows have been effective in measuring speech intelligibility in noisy environments and speech detection [22,23,24].

In this work, an alternate approach based on psychoacoustic properties combining gammatone filtering and modulation spectrum of speech, to preserve both *quality* and *intelligibility* for feature extraction 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 conventional features 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 methodology for feature extraction. Section 4 presents database description, experiments and results. Section 5 discusses the results. Finally, Section 6 concludes the paper.

2 Feature Description

In this section, a brief introduction and general overview of auditory features based on gammatone filter bank and modulation spectrum is presented.

2.1 Gammatone Filter Bank

The gammatone filter was first conceptualized by Flanagan as a model of the basilar membrane displacement in the human ear [25]. Johannesma used it to approximate responses recorded from the cochlear nucleus in the cat [26]. de Boer and de Jongh used a gammatone function to model impulse responses from auditory nerve fiber recordings, which have been estimated using a linear reverse-correlation technique [27]. Patterson et al. showed that the gammatone filter also delineates psychoacoustically determined auditory filters in humans [28].

Gammatone filters are linear approximation of physiologically motivated processing performed by the cochlea[29], comprise series of bandpass filters, whose impulse response is defined by:

$$g(t) = at^{n-1}cos(2\pi f_c t + \phi)e^{-2\pi bt} \tag{3}$$

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 ϕ is the phase.

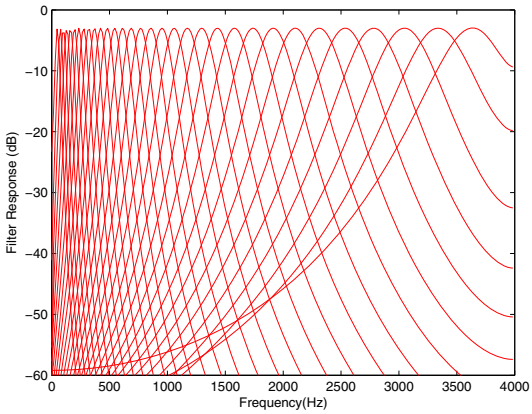

Fig. 1. Frequency response for the 32-channel gammatone filterbank

The filter center frequencies and bandwidths are derived from the filter's Equivalent Rectangular Bandwidth (ERB) as detailed in [29]. In [30], Glasberg and Moore relate center frequency and the ERB of an auditory filter as

$$ERB(f_c) = 24.7(\frac{4.37f_c}{1000} + 1) \tag{4}$$

The filter output of the m^{th} gammatone filter , X_m can be expressed by

$$X_m(t) = x(t) * h_m(t) \tag{5}$$

where $h_m(t)$ is the impulse response of the filter.

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

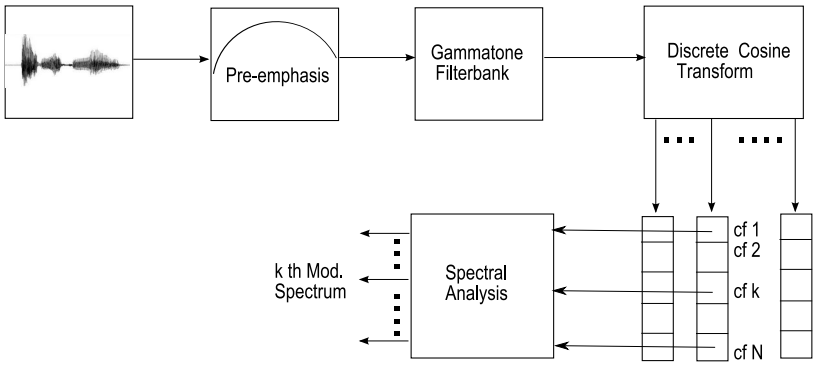

Fig. 2. Processing stages of the gammatone modulation spectral feature

2.2 Modulation Spectrum

The temporal evolution of speech spectral parameters, which describe slow variation in energy represent important information associated with phonetic segments [31]. The low-frequency modulations encode information pertaining to syllables, by virtue of variation in the modulation pattern across the acoustic spectrum. Dudley showed that essential information in speech is embedded in modulation patterns lower than 25 Hz distributed over a few as 10 discrete spectral channels [32]. Further, studies by Drullman et al. confirmed the importance of amplitude modulation frequencies on speech intelligibility, particularly modulation frequencies below 16Hz contributing to speech intelligibility [33]. Houtgast and Steeneeken demonstrated that modulation frequencies between 2 and 10 Hz can be used as an objective measure of speech intelligibility, for assessing quality of speech over wide range of acoustic environments [22].

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

$$Y(t,f) = \sum_{i=-\infty}^{\infty} X(f-i)W(i,t) \tag{6}$$

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

$$Y_m(f,g) = FT[|Y(t,f)|]|_{t=1,\ldots T} \tag{7}$$

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

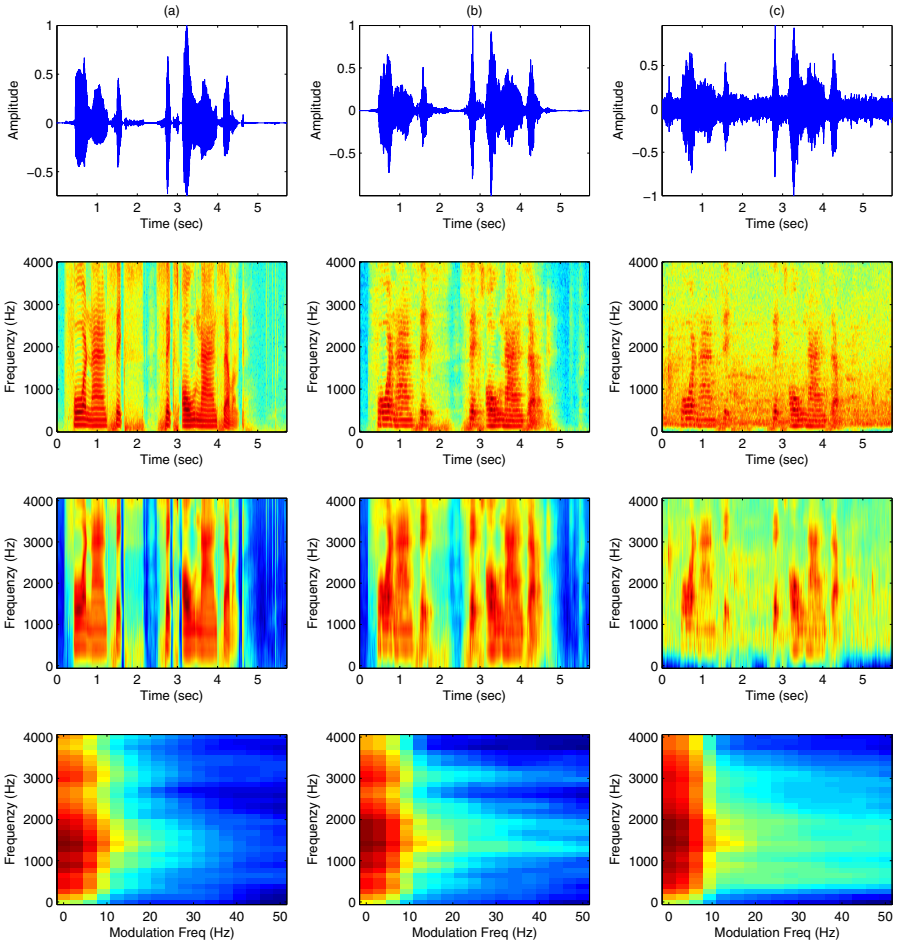

Fig. 3. Waveform, spectrogram, gammatonegram, and modulation spectrum density plots for the (a)clean, (b)reverberant and (c)additive noise corrupted speech

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 [33,21]. In [21], 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, spectrograms, gammatonegrams and modulation spectrum density plots of the clean and noisy versions corrupted with convolutive and additive noises of the same speech utterance are as shown in Fig. 3. The example is from Aurora 5 and as following:

(a) clean sentence "4966o97" from TI-DIGITS
(b) sentence in reverberant environment (living room, T60 appr. 0.5s)
(c) sentence in reverberant environment (living room, T60 appr. 0.5s) + additive noise (interior noise at 10dB).

From modulation spectrum density plots, 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 Methodology

The block schematic for the gammatone modulation spectrum based feature extraction technique is shown in Fig. 5. 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 impulse responses of the gammatone filterbank are similar to the impulse responses of the auditory system found in physiological measurements [27]. The filterbank is linear and does not consider nonlinear effects such as level-dependent upward spread of masking and combination tones. The computationally effective gammatone filter bank implementation as described in [34] 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 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 P ms Hamming window and the segment is shifted by p 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 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. The modulation spectrum of each coefficient, (sampled at 100Hz) is calculated with a 160 ms window, shifted by 10 ms. Thirteen coefficients $C13$ to $C26$ which are first-order derivatives are further extracted. The features are named gammatone filterbank modulation cepstral (GFMC) features.

The same processing is also performed by replacing gammatone filterbank with Mel filterbank in the Figure 5 resulting in Mel-frequency modulation cepstral (MFMC) features. The performance of these features in comparison to GFMC features are discussed in Section 4.

4 Experiments and Results

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 [35]. 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 sil model has three states and each state has four mixtures.

4.1 Convolutive Noise

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 additive 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.

Table 1 shows the results 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 correspond to the standard advanced front-end as described in [35]. PLP and MFCC are the standard 39-dimensional Perceptual linear prediction and Mel frequency features along with their delta and acceleration derivatives. MFMC indicate Mel Frequency Modulation Spectral based Cepstral (MFMC) features where the first thirteen features are extracted in a traditional way, and the rest are the modulation features (13) and their derivatives (13) derived as discussed in Section 3, except for Gammatone filterbank being replaced with Mel filterbank. The GFCC features are extracted in a similar way as reported in Section 3 with $C0$ to $C12$ being the corresponding cepstral coefficients. GFMC indicate Gammatone Frequency Modulation Spectral based Cepstral (GFMC) features derived in a same way as GFCC but appended with modulation spectral features corresponding to $C13$ to $C26$ and their corresponding derivatives as discussed in Section 3.

Table 1. Word recognition accuracies (%) for different feature extraction techniques on four different microphones

Channel	6	7	E	F	Average
ETSI-2	64.3	47.6	58.1	62.7	58.1
PLP	73.8	63.8	68.1	71.4	69.2
MFCC	75.8	64.7	67.3	75.9	70.9
MFMC	75.6	61.0	70.8	77.9	71.3
GFCC	86.0	79.0	78.3	84.2	81.9
GFMC	87.8	82.7	82.2	86.9	84.9

From Table 1, it is evident that the advanced ETSI front-end has highest error rates compared to the MFCC and PLP. This 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 MFMC features have better performance than MFCC, which in turn had better performance than PLP. It can also be seen that the GFCC features were effective, performing better than any of the baseline systems (ETSI-2, PLP, MFCC). This is consistent with the earlier studies which have shown that gammatone based features exhibit robust performance compared to MFCC, PLP features and ETSI frontend [16,19].

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. However, the combination of Mel filtering and modulation spectral features is not as beneficial as gammatone filtering with modulation spectral features. This clearly demonstrates the efficiency of this combination of these features in reverberation conditions.

4.2 Additive Noise

Further, to test the efficiency for practical conditions which contain additive noises along with reverberation effects, experiments are conducted on hands free office and hands free livingroom simulated data with clean and 15 dB, 10 dB, 5 dB, 0 dB SNR additive noise corrupted signals. The data is from Aurora-5 database, where condition is simulated as combination of additive noise and reverberation[35]. Aurora-5 covers all effects of noises as they occur in realistic application scenarios. In this experiments, hands free speech input in a office and in a living room is considered. The reverberation time for the office room was randomly varied in the range of 0.3 to 0.4 s and for the living room was in the range of 0.4 to 0.5 s. In Table 2, HFOffice, HFLroom, - represents hands free office, hands free living room and no additive noise respectively.

Table 2. Word recognition accuracies (%) for clean, hands free office and hands free living room conditions

Feature	HFOffice					HFLroom				
	-	15dB	10dB	5dB	0dB	-	15dB	10dB	5dB	0dB
PLP	88.6	65.1	40.4	21.1	6.7	74.3	46.9	27.7	14.2	3.7
MFCC	90.1	61.6	41.5	18.0	4.0	75.8	40.2	23.8	8.6	1.9
MFMC	94.5	68.1	40.5	18.5	8.7	85.2	52	28.2	13.5	7.8
GFCC	89.1	65.6	39.5	18.4	8.3	73.8	48.9	29.1	13.7	6.2
GFMC	92.2	73.3	44.8	20.7	10.1	78.6	57.4	34.1	15.6	8.2

From Table 2, it can be observed that for all the features the performance degrades significantly in additive noise compared to no additive noise case. Also, it can be seen that GFCC has better performance than MFCC and PLP. It can also be observed that MFMC has better performance than GFCC, showing that the combination of modulation features in mel domain is beneficial for this task. It can be observed that GFMC has better performance than GFCC, MFCC, MFMC and PLP indicating efficiency of this features in additive noise and reverberant conditions.

4.3 CHiME Challenge

The task of Computational Hearing in Multisource Environments (CHiME) challenge is to separate the speech and recognise the commands being spoken using systems that have been trained on noise-free commands and room noise recordings [36]. The CHiME background noise is recorded separately from the target speech. The target speech is subsequently artificially added but in a manner that closely simulates the effect of the speech being present in the room. This controls the target speech SNR, target talker location, talker characteristics etc. For the background noise, a domestic environment was considered, such as would be encountered in a home automation application. It provides rich mix of sound sources, some of which may be easy to model (e.g. a washing machine that remains in a fixed position and runs a predictable program) and some which are not (e.g. children running around while talking, screaming and laughing).

The ASR task is speaker dependent and employs a small but phonetically confusable vocabulary. The recordings are from a genuine living room (in a house with two small children) measured over a period of several weeks and the SNRs employed in the challenge range between 18 dB down to -6 dB.

Figure 4 shows comparison of GFMC features and MFCC before separation of the speech signals. It can be observed that GFMC is much effective than MFCC for all ranges of SNRs, particularly at very low SNRs.

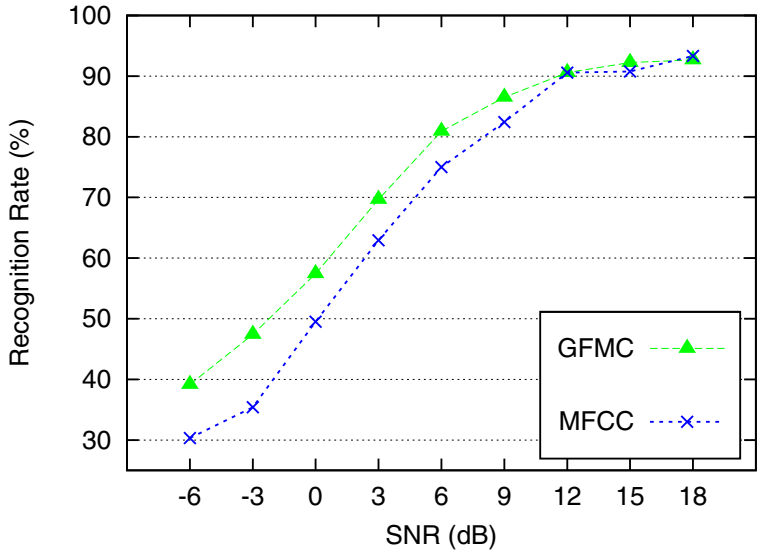

Fig. 4. Comparison of the proposed GFMC features and MFCC before separation of the speech signals

Figure 5 shows comparison of GFMC features and MFCC after separation of the speech signals which is achieved through semi-blind source separation (SBSS). The BSS algorithm is a modified Recursively Regularized Independent Component Analysis [37] according to a semi-blind structure as in [38], where the used prior is the mixing

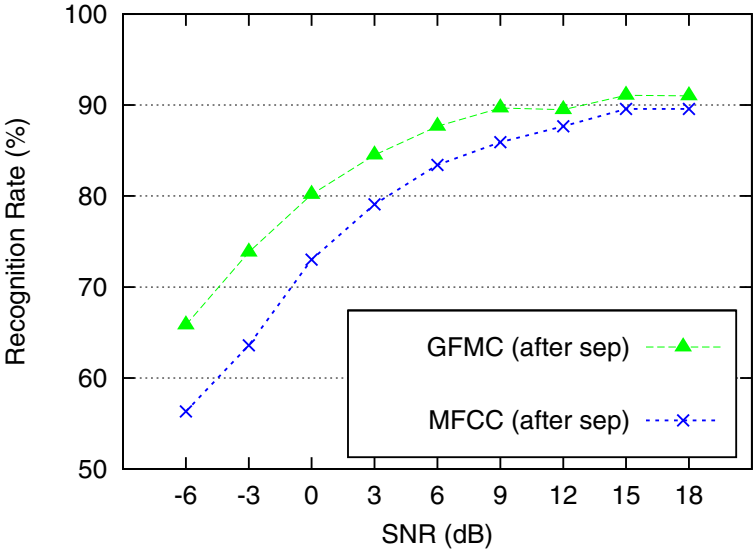

Fig. 5. Comparison of the proposed GFMC features and MFCC after separation of the speech signals achieved through semi-blind source separation

parameters of the target source (estimated beforehand). It can be observed that GFMC is much efficient than MFCC for all ranges of SNRs, with significant performance at very low SNRs.

5 Discussion

The results from both Table 1 and 2 and Figures 4 and 5 indicate that the gammatone frequency resolution was effective in reducing system sensitivity to reverberation and additive noise, and improved the speech signal characteristics. It can also be observed from Table 1, that the combination of gammatone filtering with modulation spectral features is beneficial than the combination of Mel filtering and modulation spectral features. 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. Thus, by extracting features that model human hearing to some extent mimicking the processing performed by cochlea, particularly emulating cochlea frequency resolution was beneficial for speech feature enhancement.

6 Conclusions

The paper has presented auditory inspired modulation spectral features for improving ASR performance in presence of room reverberation. The proposed features were derived from features based on emulating the processing performed by cochlea to improve the robustness, specifically gammatone frequency filtering and long-term modulations of the speech signal. The features were evaluated on Aurora-5 database, meeting

recorder digit task and living room and office room simulated data corrupted with different levels of additive noises. Results were compared with standard ETSI advanced front-end and conventional features. The results show that the proposed features perform consistently better both in terms of robustness and reliability. The work presented results in both additive noise and reverberant scenario where the speech signal was corrupted with 15 dB, 10dB, 5 dB and 0dB SNR noise, simulated with hands-free office and hands-free living room conditions. The work also presented performance of these features on CHiME challenge before and after separation of acoustic sources. The results are promising, performing better than the conventional features, indicating the efficiency of this features in practical scenarios.

Our study raised number of issues, including study of auditory inspired techniques for improvement of standard additive noise removal techniques to deal with reverberation condition. The gammatone filter implemented in this work is linear and does not consider nonlinear effects such as level-dependent upward spread of masking and combination tones. For the future, we like to investigate these issues to efficiently deal with real world noisy speech, and evaluate these features on large vocabulary tasks.

References

1. Kellermann, W.: Some current challenges in multichannel acoustic signal processing. The Journal of the Acoustical Society of America 120, 3177–3178 (2006)
2. Droppo, J., Acero, A.: Environmental Robustness. In: Handbook of Speech Processing, pp. 653–679. Springer, Heidelberg (2008)
3. Maganti, H.K., Member, S., Gatica-perez, D., Mccowan, I.: Speech enhancement and recognition in meetings with an audio-visual sensor array. In: IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne, EPFL (2006)
4. Woelfel, J., McDonough, J.: Distant Speech Recognition, 1st edn. John Wiley (2009)
5. Ephraim, Y., Cohen, I.: Recent Advances in Speech Enhancement. CRC Press (2006)
6. Habets, E.A.P.: Single-channel speech dereverberation based on spectral subtraction. In: PRORISC, Veldhoven, The Netherlands, pp. 250–254 (2004)
7. Omologo, M., Svaizer, P., Matassoni, M.: Environmental conditions and acoustic transduction in hands-free speech recognition. Speech Communication 25, 75–95 (1998)
8. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech and Audio Processing 9, 504–512 (2001)
9. Hermansky, H., Morgan, N.: Rasta processing of speech. IEEE Transactions on Speech and Audio Processing 2, 578–589 (1994)
10. Gales, M., Young, S.: A fast and flexible implementation of parallel model combination. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1995, vol. 1, pp. 133–136 (1995)
11. Holmberg, M., Gelbart, D., Ramacher, U., Hemmert, W.: Automatic Speech Recognition with Neural Spike Trains. In: INTERSPEECH (2005)
12. Deng, L., Sheikhzadeh, H.: Use of Temporal Codes Computed From a Cochlear Model for Speech Recognition. Psychology Press (2006)
13. Ghitza, O.: Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. Journal of Phonetics (1988)
14. Seneff, S.: A joint synchrony/mean-rate model of auditory speech processing. Journal of Phonetics 16, 55–76 (1988)

15. Dau, T., Pueschel, D., Kohlrausch, A.: A quantitative model of the effective signal processing in the auditory system. The Journal of the Acoustical Society of America 99, 3615–3622 (1996)
16. Flynn, R., Jones, E.: A comparative study of auditory-based front-ends for robust speech recognition using the aurora 2 database. In: Irish Signals and Systems Conference, 2006, pp. 111–116. IET (2006)
17. Kleinschmidt, M., Tchorz, J., Kollmeier, B.: Combining speech enhancement and auditory feature extraction for robust speech recognition. Speech Commun. 34, 75–91 (2000)
18. Hermansky, H.: Auditory modeling in automatic recognition of speech. ECSAP (1996)
19. Schluter, R., Bezrukov, L., Wagner, H., Ney, H.: Gammatone features and feature combination for large vocabulary speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, vol. 4, pp. IV-649–IV-652 (2007)
20. Drullman, R., Festen, J.M., Plomp, R.: Effect of reducing slow temporal modulations on speech reception. The Journal of the Acoustical Society of America 95, 2670–2680 (1994)
21. Kanedera, N., Arai, T., Hermansky, H., Pavel, M.: On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communication 28, 43–55 (1999)
22. Houtgast, T., Steeneken, H.J.M., Plomp, R.: Predicting speech intelligibility in rooms from the modulation transfer function. Acustica 46, 60–72 (1980)
23. Kingsbury, B.: Perceptually Inspired Signal-processing Strategies for Robust Speech Recognition in Reverberant Environments. PhD thesis, Michigan State University (1998)
24. Maganti, H.K., Motlicek, P., Gatica-Perez, D.: Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP (2007)
25. Flanagan, J.L.: Models for approximating basilar membrane displacement. Journal of the Acoustical Society of America 32 (1960)
26. Johannesma, P.I.: The pre-response stimulus ensemble of neurons in the cochlear nucleus. In: Symposium on Hearing Theory (Institute for Perception Research), Eindhoven, Holland, pp. 58–69 (1972)
27. Boer, E.D.: On the principle of specific coding. Journal of Dynamic Systems, Measurement, and Control 95, 265–273 (1973)
28. Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., Rice, P.: An efficient auditory filterbank based on the gammatone function. In: Meeting of the IOC Speech Group on Auditory Modelling at RSRE (1987)
29. Slaney, M.: An efficient implementation of the patterson holdsworth auditory filterbank. Technical report, Apple Computers, Perception Group (1993)
30. Glasberg, B.R., Moore, B.C.J.: Derivation of auditory filter shapes from notched-noise data. Hearing Research 47, 103–138 (1990)
31. Greenberg, S.: On the origins of speech intelligibility in the real world. In: ESCA Workshop on Robust Speech Recognition for Unkown Communication Channels, pp. 23–32 (1997)
32. Dudley, H.: Remarking speech. The Journal of the Acoustical Society of America 11, 169–177 (1939)
33. Drullman, R., Festen, J.M., Plomp, R.: Effect of temporal envelope smearing on speech reception. Journal of The Acoustical Society of America 95 (1994)
34. Ellis, D.: Gammatone-like spectrograms (2010), `http://www.ee.columbia.edu/~dpwe/resources/matlab/gammatonegram`
35. Hirsch, H.: Aurora-5 experimental framework for the performance evaluation of speech recognition in case of a hands-free speech input in noisy environments (2007), `http://aurora.hsnr.de/aurora-5/reports.html`

36. Christensen, H., Baker, J., Ma, N., Green, P.: The chime corpus: a resource and a challenge for computational hearing in multisource environments. In: Interspeech 2010 (2010)
37. Nesta, F., Wada, T., Juang, B.H.: Batch-online semi-blind source separation applied to multi-channel acoustic echo cancellation. IEEE Transactions on Audio, Speech, and Language Processing 19, 583–599 (2011)
38. Nesta, F., Svaizer, P., Omologo, M.: Convolutive bss of short mixtures by ica recursively regularized across frequencies. IEEE Transactions on Audio, Speech, and Language Processing 19, 624–639 (2011)