A Level-dependent Auditory Filter-bank for Speech Recognition in Reverberant Environments

HariKrishna Maganti, Marco Matassoni

Fondazione Bruno Kessler - Center for Information Technology - IRST
via Sommarive 18, 38123 Povo, Trento, Italy
{maganti, matasso}@fbk.eu

Abstract

Distortions due to reverberation have detrimental effect on the performance of automatic speech recognition (ASR). In this work, an auditory filter-bank based feature is presented to improve the ASR in reverberant conditions. The proposed technique is based on the gammachirp filter bank which provides level dependent frequency response to emulate mechanisms performed in the human auditory system, particularly basilar membrane filtering aimed to improve robustness of the ear. The low frequency tail of gammachirp filter which is unaffected by bandwidth parameters due to level dependency frequency resolution is effective in reducing the reverberation distortions. Experiments are performed on the Aurora-5 meeting recorder digit task recorded with four different microphones in hands-free mode at a real meeting room. The ASR experiments using the proposed gammachirp based features show reliable and consistent improvements when compared to other conventional feature extraction techniques.

Index Terms: auditory processing, gammachirp filtering, asymmetric and level dependency frequency analysis, speech recognition

1. Introduction

In spite of a host of recent advances in speech recognition technology, the accuracy of recognition continues to be highly influenced by speaker-microphone distance. The performance degrades with increasing distance due acoustic environmental noise and reverberation. It restricts the usage of various speech recognition applications including teleconferences, human-computer dialogue systems, dictation and navigation systems.

In the context of reverberation environments, the discrepancy between different training and testing conditions is the root cause for performance degradation. Typically, training data is recorded in clean and non-reverberant conditions. Speech signal enhancement, feature normalization and model parameterization techniques are applied to improve robustness against convolutive distortion caused by reverberation at various levels of processing [1, 2, 3].

Human auditory perception system is more conducive in providing robustness against noise. Hence techniques exhibiting few characteristics of auditory system have been used in speech recognition to improve the robustness [4, 5]. An important aspect in a speech recognition system is to have abstract representation of highly redundant speech signal, which is achieved by frequency analysis. In auditory system, the cochlea and hair cells of the inner ear perform spectrum analysis to extract relevant features. The cochlea, as a filterbank exhibits non-uniform frequency resolution, asymmetric and level-dependent frequency response of an individual filters. Examples of non-uniform frequency resolution in popular speech analysis techniques include Mel frequency based features and perceptual linear prediction which attempt to emulate human auditory perception. An other popular filterbank inspired by auditory system which has non-uniform bandwidths and non-uniform spacing of center frequencies is gammatone filter. The use of gammatone filterbank provided robustness in adverse noise conditions for speech recognition tasks in [6, 7, 8] when compared to traditional front-ends based on MFCC, PLP and standard ETSI on isolated word and large vocabulary speech recognition tasks. Another important psychoacoustic property is modulation spectrum of speech, which represent low temporal modulation components which are important for speech intelligibility [9, 10]. The relative prominence of slow temporal modulations is different at various frequencies, similar to perceptual ability of human auditory system. The modulation spectral features derived from the gammatone filterbanks have been shown to improve the robustness for far-field speaker identification [11]. Also, in our earlier work [12], an auditory based modulation spectral feature is presented which was a combination of gammatone filtering and modulation spectral features.

The gammatone filter is a linear filter and its frequency response is symmetric about center frequency and does not model level dependent properties. The gammachirp filter, derived by Irino and Patterson, is a modification of gammatone filter by adding frequency modulation term. The gammachirp auditory filter is real part of the analytic gammachirp function which has been shown to be an excellent function for the asymmetric, level dependency observed in basilar membrane filtering [13]. The gammachirp is characterized with asymmetry in the low frequency tail of auditory filter response and models level dependent properties such as decrease in gain and a shift in center frequency of the filter with increase in level. In this work, this property of gammachirp filter is utilized to minimize the tailing effect of decaying sound to reduce reverberation distortions for speech recognition. And long term modulation preserved the speech intelligibility further improving the recognition accuracy. The studied features are shown to be reliable and robust to the effects of 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.

The paper is organized as follows: Section 2 provides theory and implementation details of the gammachirp filter and

modulation spectral features. In Section 3, the proposed feature extraction is presented. Section 4 presents database description, experiments and results. Finally, Section 5 concludes the paper.

2. Theoretical Background

2.1. Gammachirp Filter Bank

The impulse response of a gammachirp filter is defined by

$$g_c(t) = g_t(t) \, . \, e^{\,j\,C\,ln\,t} \tag{1}$$

where $g_t(t)$ is a gamma envelope modulated by a tone carrier defined by

$$g_t(t) = a\,t^{n-1} e^{-2\pi\,b\,ERB(f_r)\,t} e^{\,j\pi\,(f_r)\,t + j\,\phi} \tag{2}$$

where $t \geq 0$, C is the chirp factor which defines frequency modulation to produce an asymmetric amplitude spectrum, a is the amplitude, n and b are parameters defining the distribution, f_r is asymptotic frequency and ϕ is initial phase. $ERB(f_r)$ is the equivalent rectangular bandwidth of auditory filter at f_r, and at moderate levels, $ERB(f_r) = 24.7 + 0.0108(f_r)$. When $C = 0$, $g_c(t)$ becomes the impulse response of a gammatone filter.

The amplitude spectrum of gammachirp can be expressed in terms of gammatone as

$$|G_c(f)| = \frac{|a\Gamma(n + jC)|}{|2\pi bERB(f_r) + j2\pi(f - f_r)|^n} . \, e^{C\theta}, \tag{3}$$

$$\theta = \tan^{-1}\left(\frac{f - f_r}{ERB(f_r)}\right) \tag{4}$$

n and b define envelope of gamma function and when C = 0, the numerator in equation 3 represents amplitude of gammatone since $e^{C\theta}$ = 1, and peak frequency is obtained as

$$f_{peak} = f_r + CbERB(f_r)/n \tag{5}$$

$e^{C\theta}$ produces shift in the peak frequency according to equation 4 and introduces asymmetry into the amplitude spectrum. With amplitude normalized, equation 3 can be expressed as

$$|G_c(f)| = |G_T(f)| \, . |H_A(f)| \tag{6}$$

where $|G_T(f)|$ is the amplitude spectrum of the gammatone, which is level-independent and invariant. Hence, gammachirp can be represented by two cascaded filters, the first one is invariant gammatone filter $|G_T(f)|$ and other $|H_A(f)|$ is asymmetric level-dependent filter.

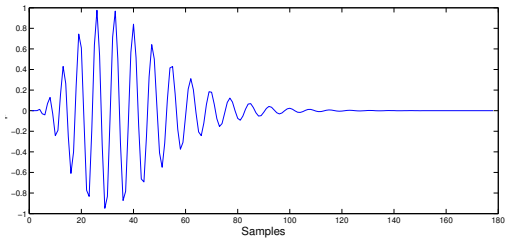

Figure 1: *Impulse response of a gammachirp filter with f_r = 1000 Hz, n= 4, b = 1.0 and C = 3.4.*

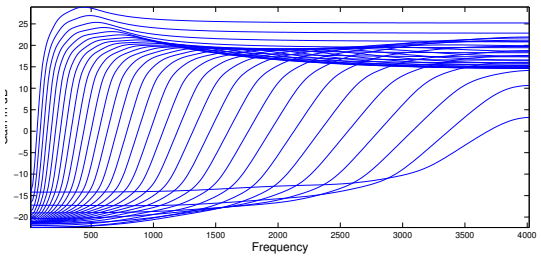

Figure 2: *The plots of asymmetry function for center frequencies ranging from 50 Hz - 3655 Hz, C = 3.4, b =1.0 and n = 4.*

2.1.1. Implementation

The asymmetric function is realized using IIR asymmetric compensation filter. The minimum-phase IIR filter $H_c(z)$ is constructed by a cascade of 4 second order filters $H_{ck}(z)$ as

$$H_c(z) = \prod_{k=1}^{4} H_{ck}(z) \tag{7}$$

$H_{ck}(z)$ and other required filter parameters are described in [14].

As mentioned, the gammachirp filter provides asymmetric component based on input level. In this work, passive mode of the filter is considered by taking f_r equal to center frequency. The amplitude spectra of asymmetric compensation IIR filter for the gammachirp filter for center frequencies which range from 50 Hz to 3655 Hz is as shown in Figure 2. It can be clearly seen that the low-frequency tail of filter is unaffected by the bandwidth parameters. By changing the center frequency of asymmetry, the gain is made level-dependent in a way agreeable to psychophysical data [14].

Figure 3 shows the amplitude spectra of asymmetric level-dependent filter $|H_A(f)|$ for different values of chirp factor C. It can be observed that $|H_A(f)|$ is an all-pass filter when $C = 0$, a high-pass filter when $C > 0$ and a low-pass filter when $C < 0$. The slope and the range of amplitude increase when the absolute value of C increases. The overall resulting spectrum is asymmetric and exhibits a sharper drop off, on high frequency side of center frequency. With asymmetry associated to the stimulus level making it level-dependent, the gammachirp provides a good fit to human masking[13].

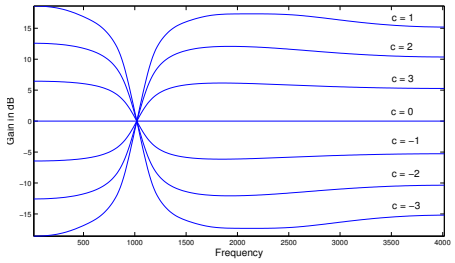

Figure 3: *The plots of the asymmetry function for f_r = 1000 Hz, b = 1.0 and C = {3,2,1,0,-1,-2,-3}.*

2.2. Modulation Features

The temporal evolution of speech spectral parameters, which describe slow variation in energy represent important information associated with phonetic segments. The low-frequency modulations encode information pertaining to syllables, by virtue of variation in the modulation pattern across the acoustic spectrum. The essential information in speech is embedded in modulation patterns lower than 25 Hz distributed over a few discrete spectral channels [9, 10].

The modulation spectral features are derived exactly the same way as described in [12].

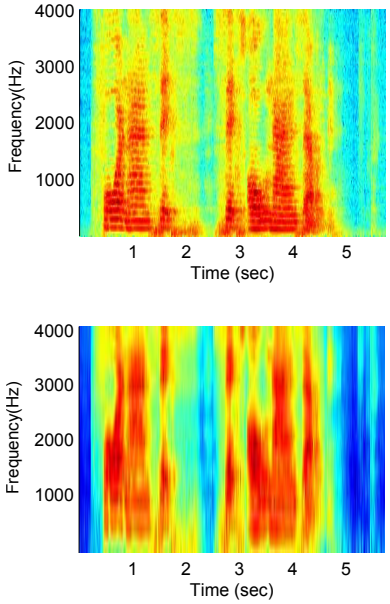

Figure 4: *Spectrogram and gammachirpgram for reverberated utterance from TIDIGITS.*

3. Feature Extraction

The proposed feature extraction methodology is shown in Figure 5. First, the 8kHz speech signal undergoes pre-emphasis and short segments of speech are extracted with a 25 ms rectangular window, shifted by 10 ms. This is then filtered by a bank of 32 critical-band gammatone filters whose filter center frequencies range from 50 Hz to 3655 Hz (half of sampling rate). On each critical band, the asymmetric function is applied which is based on center frequencies as described in Section 2. The computed chirp components are combined with the corresponding gammatone filter outputs to form outputs of a gammachirp filterbank. The computationally effective gammatone filter bank implementation as described in [15] is used. The asymmetric chirp components and gammachirp filter bank is implemented as described in [14]. The 32 logarithmic gammachirp spectral values are transformed to the cepstral domain by means of 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. The cummulated energies for the frequencies between the 2 - 16 Hz, which represent the important components for the speech signal are computed as $C13$ to $C26$. The two important parameters of the gammachirp filter $b = 1.68$ and $c = 2.5$ are derived by fitting the frequency curves to notched noise masking data [14]. The spectrogram and gammachirpgram for reverberated utterance "4966o97" from TIDIGITS with T60 = 0.5s is shown in Figure 4.

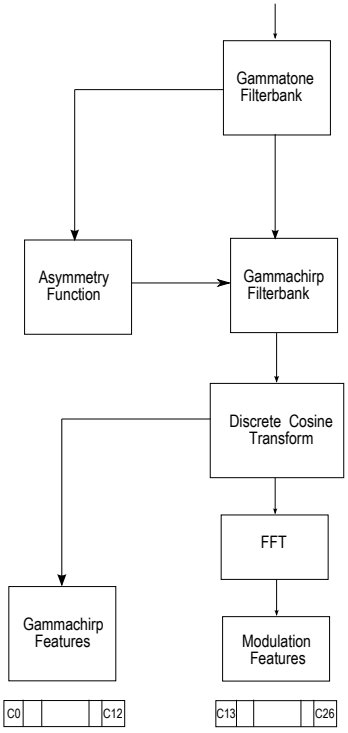

Figure 5: *Processing stages of the gammachirp features.* $C0$ - $C12$ *are the gammachirp cepstral coefficients and* $C13$ - $C26$ *are the gammachirp modulation coefficients.*

4. Experiments and Results

Experiments are performed 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. All data was recorded in the same (roughly, 13 x 25 foot) instrumented meeting room. The room contains a central conference table almost completely filling the room, and can seat up to about 15 people. The reverberation time T60 is in the range of 0.2 s to 0.3 s.

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 [16]. The dataset consists of 2400 utterances from 24 speakers, with 7800 digits in total. The speech was captured with four different microphones, placed at middle of the table in the meeting room. The recordings have the effects of hands-free recording in the reverberant room. There are four different versions of all utterances recorded with four different microphones.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.

Table 1 shows the results in % word error rates for different features 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.

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

Channel	6	7	E	F	Average
MFCC	75.8	64.7	67.3	75.9	70.9
PLP-MVA	83.3	76.3	75.9	80.7	79.0
GTCC	83.7	76.5	76.4	82.3	79.7
GTMC	88.6	83.7	83.5	87.2	85.6
GCCC	83.4	77.2	77.7	83.8	80.5
GCMC	89.7	84.6	85.9	88.3	87.1

For comparison purposes, the standard 39-dimensional Mel-frequency (MFCC) and perceptual linear prediction (PLP) features along with their delta and acceleration derivatives are used. Cepstral mean normalization and Mean Variance ARMA (MVA) model adaptation are performed for MFCC and PLP to improve the performance of these features. The filter of order 3 is used for ARMA filtering for PLP [17]. The GTCC are gammatone frequency cepstral coefficient features along with their delta and acceleration derivatives. GTMC are the gammatone frequency based modulation cepstral features, extracted as reported in Section 3 corresponding to $C0$ to $C26$ and derivatives of $C13$ to $C26$, but considering the gammatone filterbank instead of gammachirp. GCCC and GCMS are the gammachirp and gammachirp modulation features.

It can be seen from Table 1 that gammatone based features perform better than MFCC and PLP-MVA which is consistent with earlier studies in [6, 7, 8] and modulation spectral features further improved the performance by preserving the speech intelligibility in the signal. It can also be seen that the proposed gammachirp has better performance than gammatone alone and gammachirp modulation features have best performance among all the features compared. It can also be observed that the performance of gammachirp based modulation features is consistent across all the four different channels. Apart from consistent and robust performance, the level-dependency based on input stimulus is simple to be incorporated into existing speech recognition applications. The results show that the proposed level-dependent frequency response was effective in curtailing reverberation distortions, thereby improving the accuracy of the system.

5. Conclusions

In this paper, an asymmetric based level-dependent auditory feature is proposed to improve speech recognition performance in reverberant environments. The proposed features were derived from auditory characteristics, which include gammatone filtering, asymmetric and level-dependent frequency response and modulation spectral processing emulating cochlear and middle ear to improve robustness. The features were evaluated on the Aurora-5 meeting recorder digit task recorded with four different microphones in hands-free mode at a real meeting room. The level-dependency improved performance for both the cases of gammatone and gammatone based modulation filtering. The proposed features perform consistently better in terms of robustness for all the channels. This study opens up a new avenue for incorporating level-dependency based frequency response into the existing speech recognition technologies making them more suitable to psychophysical data and good fit for human masking, thereby improving the robustness of system.

The present work was limited to handle reverberant conditions, without considering other noise effects on speech which will be investigated in the future work. In this work the filter is designed to be in passive mode by considering the level dependency based on center frequencies. The active mode of the filter will also be studied in future by incorporating dynamic and compressive characteristics to the auditory filter.

6. Acknowledgements

The authors would like to thank Dr. Toshio Irino, Wakayam University, Japan for kindly providing the software for gammachirp filters.

7. References

[1] Droppo, J.and Acero, A.,"Environmental Robustness", in Springer Handbook of Speech Processing, Benesty, J.; Sondhi, M. M. and Huang, Y.[Eds], 653-679,Springer, 2008

[2] Rosenberg A. E., Lee C. H., and Soong F.K. "Channel Normalization Techniques for HMM-Based Speaker Verification", ICSLP, 1835 -1838, 1994.

[3] Gales M. J. F.,"Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition", Computer Speech and Language, vol.12, no.2, pp.75 -98,1998.

[4] Kim C., " Signal Processing for Robust Speech Recognition Motivated by Auditory Processing", Ph. D Thesis, CMU, 2010.

[5] Brown G.J., and Palomaki K. J., "A Reverberation-robust Automatic Speech Recognition System Based on Temporal Masking ", Journal of Acoustical Society of America, 123(5), 2978, 2008.

[6] Flynn R. and Jones E., "A Comparative Study of Auditory-based Front-ends for Robust Speech Recognition using the Aurora 2 Database", IET Irish Sig. and Sys. Conf.,(ISSC), 111–116, 2006.

[7] Schluter R.,Bezrukov I., Wagner H., Ney H. "Gammatone Features and Feature Combination for Large Vocabulary Speech Recognition", IEEE International Conf. on Acoustics, Speech, and Signal Proc. (ICASSP), Hawai, U.S.A, 1:4-7, 2007.

[8] Shao Y., Jin Z., Wang D.L., and Srinivasan S. "An Auditory-based Feature for Robust Speech Recognition" Proceedings of IEEE International Conf. on Acoustics, Speech, and Signal Proc. (ICASSP), Taipei, Taiwan, pp. 4625-4628, 2009.

[9] Drullman R., Festen J., and Plomp R. "Effect of Reducing Slow Temporal Modulations on Speech Reception", Journal of Acoustical Society of America,pp. 2670-2680, 1994.

[10] Kanedera N., Arai T., Hermansky H. and Pavel M., "On the Importance of Various Modulation Frequencies for Speech Recognition" Proceedings of Eurospeech, Rhodes, 1997.

[11] Falk T. H., and Chan W. Y., "Modulation Spectral Features for Robust Far-Field Speaker Identification", IEEE Trans. Speech and Audio Proc., 18(1):90–100, 2010.

[12] Maganti H. K., and Matassoni M., "An Auditory Based Modulation Spectral Feature for Reverberant Speech Recognition", Proc. Interspeech, Makuhari, Japan, 2010 pp. 570 -573.

[13] Irino T., and Patterson R. D., "A Time-domain, Level-dependent Auditory filter: The Gammachirp", Journal of Acoustical Society of America, pp. 412-419, 1997.

[14] Irino T., and Unoki M. "An Analysis Auditory Filterbank Based on an IIR Implementation of the Gammachirp", Journal of Acoustical Society of America, pp. 397-406, 1999.

[15] Ellis D.P.W. "Gammatone-like Spectrograms" Online: http://www.ee.columbia.edu/ dpwe/resources/matlab/gammatonegram/.

[16] Hirsch H.G. "Aurora-5 Experimental Framework for the Performance Evaluation of Speech Recognition in Case of a Hands-free Speech Input in Noisy Environments" Online: http://aurora.hsnr.de/aurora-5/reports.html.

[17] Chen C. and Bilmes J.A. "MVA Processing of Speech Features", IEEE Trans. on Speech and Audio Proc., 15(1):257–270, 2007.