Gammatone Wavelet Cepstral Coefficients for Robust Speech Recognition

Aniruddha Adiga
Department of Electrical Engineering
Indian Institute of Science
Bangalore-560 012, Karnataka, India
Email: aniruddha@ee.iisc.ernet.in

Mathew Magimai.-Doss
Idiap Research Institute
CH-1920, Martigny, Switzerland
Email: mathew@idiap.ch

Chandra Sekhar Seelamantula
Department of Electrical Engineering
Indian Institute of Science
Bangalore-560 012, Karnataka, India
Email: chandra.sekhar@ieee.org

Abstract*—We develop noise robust features using *Gammatone wavelets* derived from the popular *Gammatone functions*. These wavelets incorporate the characteristics of human peripheral auditory systems, in particular the frequency response of the *basilar membrane*. We refer to the new features as** *Gammatone Wavelet Cepstral Coefficients* **(GWCC). The procedure involved in extracting GWCC from a speech signal is similar to that of the conventional Mel-Frequency Cepstral Coefficients (MFCC) technique, with the difference being in the type of filterbank used. We replace the conventional mel filterbank in MFCC with a** *Gammatone wavelet filterbank, which we construct using Gammatone wavelets. We also explore the effect of Gammatone filterbank based features (Gammatone Cepstral Coefficients (GCC)). On AURORA 2 database, the comparison of GWCCs or GCCs with MFCCs shows that Gammatone based features yield a better recognition performance at low SNRs.**

Keywords*—*Gammatone wavelets, Auditory modeling, Cepstral coefficients, Speech recognition.

I. Introduction

The goal of feature extraction in speech recognition is to provide a stable and robust representation for speech signal. The two widely used features are MFCC [1] and Perceptual Linear Prediction (PLP). Both methods involve extracting cepstral vector that is derived from a filterbank, which is designed according to a particular human auditory model [2]. In case of MFCC, filterbank construction is based on pitch perception, while PLP uses a filterbank that approximates the critical band masking property of the peripheral auditory system.

Another aspect of the auditory system on which the construction of the filterbank can be based is the frequency response of the *basilar membrane*. The basilar membrane in the inner ear aids in frequency resolution and many models have been proposed to mimic this function. Most models are based on representing the basilar membrane as a constant-Q filterbank. Patterson and Smith have compared various models and show that the Gammatone functions provide a good fit to the experimentally determined auditory response [3]. Based on these functions a construction of Gammatone filterbank has been presented in [4].

An important characteristic of basilar membrane as suggested by Yang et al. [5] is that its functioning can be viewed as an affine wavelet transform. Since wavelets are zero-average (bandpass) functions and exhibit constant-Q behaviour [6], wavelet transformation can be represented by a bandpass constant-Q filterbank. The generation of Continuous Wavelet Transform (CWT) requires infinite translations and dilations of the wavelet function [6] and in practice, we can only approximate the transform with a finite set of filters. Solbach et al. give a method of generating CWT from the Gammatone functions [7]. This construction yields an approximate wavelet transform as Gammatone functions do not have a zero at $\omega = 0$. We overcome this problem by using Gammatone wavelets (derived from Gammatone functions) [8] and constructing a constant-Q filterbank from these wavelets. Apart from modeling the basilar membrane, another motivation for using wavelets comes from the fact that wavelet transform based features can be used in obtaining Vocal Tract Length Invariant (VTLI) features, as shown by Mertins and Rademacher [9]. The wavelets considered have similar asymmetry characteristcs as the popular Gammatone functions.

In this paper we construct Gammatone wavelet filterbank using these wavelets. We replace the mel-filterbank in the MFCC feature extraction process with the Gammatone wavelet filterbank and investigate the effect of changing filterbank on the performance of the Automatic Speech Recognition (ASR) system. We also experiment with Gammatone filterbank and refer to the features as Gammatone Cepstral Coefficients (GCC). A similar feature extraction approach has been presented in [10] wherein auditory transform (consisting of filters based on Gammatone functions) coupled with hair cell model are used in obtaining cochlear filter cepstral coefficents. Even in this case a near wavelet is considered.

The organization of the paper is as follows. In section II, we give a brief introduction of the Gammatone functions and construction of Gammatone wavelets. In Section III we provide a method for Gammatone wavelet filterbank construction. In Section IV, we discuss the feature extraction process. In Section V we describe the experimental setup and analyse the results of the experiment. In Section VI, we discuss previous works that have incorporated Gammatone filterbank in feature extraction and summarize the work presented in this paper.

II. Gammatone wavelets

The basilar membrane within the cochlea in the inner ear has high frequency selectivity. The frequencies are resolved tonotopically, that is, different points on the basilar membrane resonate at different frequencies. Experimental studies have shown that the frequency response at any point on the membrane is asymmetric. In the modeling of this response, the asymmetric Gammatone functions provide a good fit to the experimentally determined response [3].

The Gammatone function is a tone (sinusoid) modulated by a gamma distribution function and is expressed as $g(t) = t^{(N-1)}e^{-\alpha t}e^{\mathrm{j}\omega_c t}u(t)$, where t (in seconds) denotes time, α is the bandwidth parameter and determines the effective duration of $g(t)$; the center frequency is ω_c (in radians/second), $u(t)$ is the unit step function and N is the order, which controls the rise and decay of the function. For N in the range 3–5, the Gammatone function provides a good approximation to the basilar membrane responses [3]. We use $N = 4$ in our construction. Though in Patterson and Smith's model real Gammatone function was considered, we use the complex version. The reason for using such functions is that in feature extraction process we only require one-sided spectra and complex functions help in achieving that goal. The Fourier transform of $g(t)$ is $\hat{g}(\omega) = \frac{(N-1)!}{(\alpha+\mathrm{j}(\omega-\omega_c))^N}$, where ω (in radians/second) is the angular frequency.

A. Gammatone wavelets

Taking the derivative of the Gammatone function introduces a zero at $\omega = 0$. The derivative of the Gammatone has a straightforward Fourier transform expression given by

$$\widehat{\psi}(\omega) = \mathrm{j}\omega \cdot \hat{g}(\omega) = \frac{\mathrm{j}\omega \cdot (N-1)!}{(\alpha + \mathrm{j}(\omega - \omega_c))^N}. \qquad (1)$$

It is easy to verify that $\hat{\psi}(0) = 0$. The corresponding function in the time domain is

$$\begin{aligned} \psi(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left\{t^{(N-1)}e^{-\alpha t}e^{\mathrm{j}\,\omega_c t}u(t)\right\}, \\ &= \left((N-1)t^{(N-2)} + \beta t^{(N-1)}\right)e^{\beta t}u(t), \qquad (2) \end{aligned}$$

where $\beta = -\alpha + \mathrm{j}\omega_c$. Higher-order derivatives of the Gammatone function also qualify as wavelets, provided that the order of the derivative does not exceed the order of the denominator in (1). A detailed description about the properties of the wavelets is provided in [8].

III. Gammatone wavelet filterbank

We develop the filterbank along the same lines as Slaney's implementation of Gammatone filterbank [4] and incorporate our filterbank in the auditory tool box [11].

In Slaney's construction of the Gammatone filterbank, bandwidth of each filter is described as an Equivalent Rectangular Bandwidth (ERB). The expression chosen for calculating ERB (in Hz) at any frequency f (in Hz) is

$$ERB(f) = \frac{f}{9.26} + 24.7 \qquad (3)$$

The calculation of center frequency f_c for a channel k, is based on the following expression,

$$f_c(k) = -C + e^{k\log\left(\frac{f_{min}+C}{f_{max}+C)}\right)/K} \cdot (f_{max} + C), \qquad (4)$$

where $1 \leq k \leq K$, K is the total number of filters, $C = 228.83$, f_{min}(in Hz) and f_{max} (in Hz) are lowest and highest cutoff frequencies of the filterbank. In our construction $f_{min} = 133\,\mathrm{Hz}$ and $f_{max} = 4\,\mathrm{kHz}$ (half the sampling rate of the input data) are chosen. Substituting the center frequency and bandwidth values obtained using Equation (4) and Equation (3), in Equation (2), we generate a constant-Q bandpass filterbank. The filterbank thus obtained consists of continuous time filters and from the implementation point of view we need to derive discrete filters. The discrete filters are obtained by sampling the impulse response of Gammatone wavelet function in Equation (2) at the sampling rate of the input data.

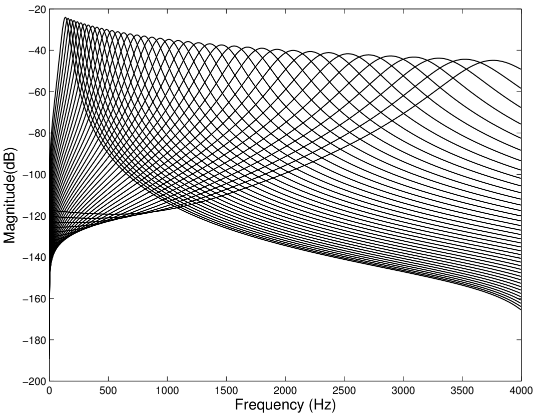

Fig. 1: *A Gammatone wavelet filterbank consisting of 40 filters with a lowest frequency of 133.33 Hz and highest frequency of 4 kHz.*

The filterbank is as shown in Figure 1 and the plot reveals the asymmetricity property of Gammatone wavelet in the frequency domain. The filterbank consists of 40 filters with the lowest frequency of 133 Hz and highest frequency of 4 kHz (half the sampling frequency of the speech data used). The choice of center frequencies and bandwidth is the same as used in Slaney's implementation of Gammatone filterbank. We observe a reduction in the height of the filters with the increase in frequency, this is a consequence of the fact that the area under the filters have been kept equal.

IV. Feature extraction using Gammatone wavelet filterbank

In the extraction of features for ASR we draw inspiration from Mel Frequency Cepstral Coefficients (MFCC). We follow the same set of steps as followed in obtaining MFCCs, but the difference lies in choice of filterbank. The Mel-filterbank in MFCC, which is made up of triangular filters equally spaced in Mel scale, is replaced by the Gammatone wavelet filterbank. We refer to the coefficients obtained using Gammatone wavelet filterbank as Gammatone Wavelet Cepstral Coefficients (GWCC). The flowchart in Figure 2 illustrates the feature extraction processes for MFCC and GWCC. In order to maintain consistency in the implementation of MFCC and Gammatone wavelet cepstral coefficients, we integrated our filterbank in the auditory toolbox's MFCC. Also the areas of Gammatone wavelet filters and Mel-filters were kept equal.

Since the construction of Gammatone wavelet filterbank is based on Gammatone filterbank, we have extracted features using this filterbank. The extraction process is similar to that of GWCC, but instead of the Gammatone wavelet filterbank, we use Gammatone filterbank. We refer to these features as Gammatone Cepstral Coefficients (GCC).

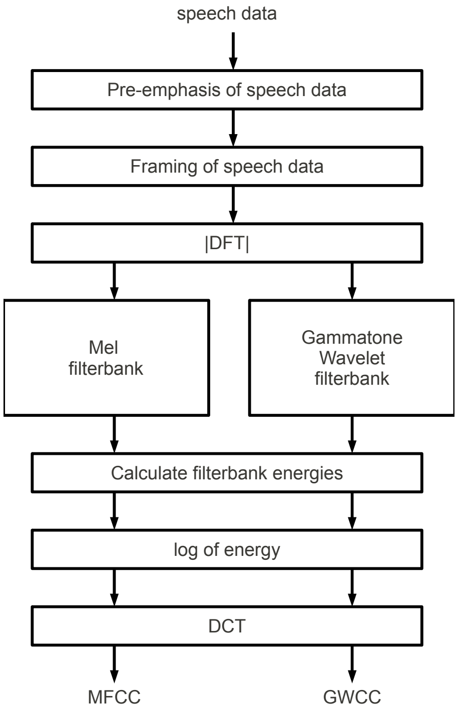
speech data
Pre-emphasis of speech data
Framing of speech data
|DFT|
Mel filterbank
Gammatone Wavelet filterbank
Calculate filterbank energies
log of energy
DCT
MFCC
GWCC

Fig. 2: *Flowchart describing the methods involved in obtaining MFCC and GWCC.*

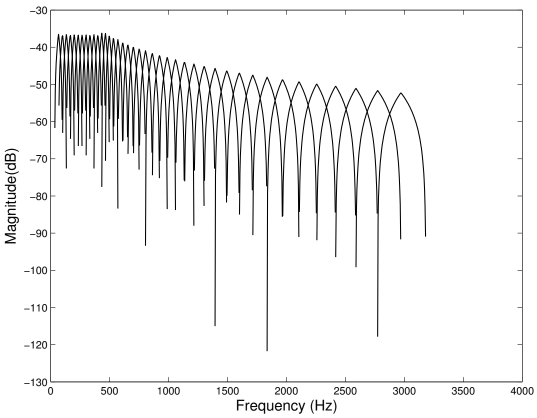

Fig. 3: *A Mel-filterbank consisting of 40 filters with a lowest frequency of 133.33 Hz and highest frequency of 4 kHz.*

The Mel-filterbank used in auditory toolbox's MFCC implementation is shown in Figure 3. The filterbank consists of 40 filters and the spacing between them is different when compared to Gammatone wavelet filterbank. The first 13 filters are spaced linearly and the rest are spaced logarithmic [1]. Comparing this filterbank with the Gammatone wavelet filterbank shown in Figure 1, we observe that the Mel-filters are strictly bandlimited and hence have a sharper rolloff, whereas every Gammatone wavelet filter responses spreads across the entire frequency band.

V. Experimental setup and results

We investigate the noise robustness of GWCC on AURORA 2 database. The recognizer is trained using *clean speech training data*, which consists of 8440 utterances. The test data consists of two sets, *test set A* and *test set B*. The *test set A* contains utterances corrupted by four noise types, suburban train, babble, car and exhibition and *test set B* contains utterances corrupted by four other noise types, restaurant, street, airport and exhibition hall. Both *test set A* and *test set B* contain 4004 utterances consisting of a sequence of one to seven digits, with 1001 utterances for each noise type. All utterances occur at seven noise levels, viz. clean, and SNR = 20, 15, 10, 5, 0, and -5 dB. The speech data is sampled at 8 kHz. More details about the database and clean condition setup can be found in [12].

We used the reference recognizer based on the HTK software package supplied with AURORA 2 database for the recognition experiments [13] [14]. The reference recognizer takes as input, static features and estimates dynamic features (Δ and $\Delta\Delta$). Then, trains an HMM for each digit. The digits are modeled using basic 16 state Gaussian Mixture Model (GMM) with 3 Gaussians per state. Silence is modeled by a 3 state GMM with 6 Gaussians per state. For each of the features, namely, MFCCs, GWCCs, GCCs extracted using Slaney's auditory toolbox, we trained and tested a system. In addition, we also trained and tested a system with MFCCs extracted using HTK software package. The word accuracy results obtained using MFCC (HTK), MFCCs, GWCCs and GCCs are tabulated in Table I and Table II. The average word accuracy for a given test set at a particular SNR was calculated by averaging the accuracies for the four different noises in that set. The average word accuracy under all SNR conditions is also provided.

It can be seen from the results that HTK based MFCCs and Slaney's auditory toolbox based MFCCs yield significantly different performance. More specifically, HTK based MFCCs, which yields a performance comparable to system reported earlier in the literature [12][14], performs worse than MFCCs extracted using auditory toolbox. This difference in the recognition performance can be attributed to the variation in filterbank implementation between Slaney's toolkit(equal area filters) and HTK toolkit (equal height filters). The consequence of equal height filters is that filters with higher bandwidth have higher energy. Typically for a speech signal low frequency regions have higher SNR when compared to high frequency regions, thus the influence of noise is more prevalent towards higher frequency regions. The construction of the filterbank is such that filters with higher center frequencies have higher bandwidths and effect of this is a larger amplification of energy in the higher frequency regions of the speech signal when compared to lower frequency regions, thus resulting in poorer performance as opposed to equal area filters.

In both *test set A* and *test set B*, we observe that under all SNR conditions, both GCC and GWCC show higher word accuracy when compared to MFCC. More importantly, we observe greater improvement at low SNRs like 10, 5 and 0 dB. The improvement in word accuracy of both GWCC and GCC can be attributed to the change of filterbank that was incorporated in the feature extraction process. Recalling Section III, the differences in the filterbanks where the shape, placement and bandwidth of the filters. Thus we may infer that

TABLE I: *Word accuracy comparison (in %) for test set A.*

SNR(dB)	MFCC(HTK)	MFCC	GWCC	GCC
clean	99.10	98.05	98.67	98.43
20	95.42	95.19	96.42	96.24
15	85.44	89.94	91.75	92.33
10	62.11	72.76	76.05	78.77
5	31.48	40.36	43.61	46.80
0	12.54	19.86	21.93	22.87
Avg (0-20dB)	64.35	69.36	71.41	72.57

TABLE II: *Word accuracy comparison (in %) for test set B.*

SNR(dB)	MFCC(HTK)	MFCC	GWCC	GCC
clean	99.10	98.05	98.67	98.43
20	94.09	95.05	95.65	95.69
15	81.97	90.32	91.86	92.38
10	57.01	76.80	79.61	81.11
5	28.32	47.33	51.76	52.72
0	12.02	23.72	25.61	26.36
Avg(0-20dB)	63.75	71.87	73.86	74.45

the asymmetry of the filter response and the placement of the filters have a bearing on the features.

It can be observed that GCC provide better results when compared to GWCC, however, there are areas in the construction of Gammatone wavelet filterbank that could yield better features. In our construction we have only considered first order derivative of the Gammatone function. Increasing the order of the derivatives give rise to wavelets with a slower decaying tail when compared to the first order. The possibility of using these wavelets in better approximation of the basilar membrane responses has to be investigated. Also the $j\omega$ term in Equation (1) indicates an implicit pre-emphasis being performed by the Gammatone wavelet. So the effect of removing the explicit pre-emphasis in GWCC on the features has to be analyzed.

VI. Conclusions

In the past Gammatone filterbanks have been explored for feature extraction process. For instance, Tchorz and Kollmeier in [15] constructed a Gammatone filterbank based auditory model front end that simulates the spectral and temporal aspects of the peripheral auditory system and showed improvement in recognition performance when compared with MFCC. Tüske et al. in [16] introduced a non-stationary signal analysis into ASR, by developing pitch-adaptive Gammatone filterbank, which typically gave noise robustness over standard features. Schlüter et al. in [17] gave a method of obtaining features from a Gammatone filterbank and also showed various techniques for combining these features with standard features like MFCC and PLP. They also investigated the effects of using 3^{rd}, 10^{th} and log amplitude compression schemes. Li and Huang in [10] developed auditory transforms (which uses cochlear filters that are similar to Gammatone functions) based cochear filter cepstral coefficients (CFCC) and show robustness of these features over MFCC. The feature extraction followed in this case is also similar to MFCC. All these works typically used to extract a feature different than standard MFCC or PLP and showed the robustness aspect of Gammatone filters. However, in our work we observe that by simply replacing Mel-filterbanks by filterbanks obtained using Gammatone wavelets or Gammatone and keeping rest of the processing same yields robustness over standard MFCC features.

Acknowledgment: The ASR part of the research work of Aniruddha Adiga was carried out at Idiap Research Institute and was funded by Indo Swiss Joint Research Programme (http://indo-swiss.epfl.ch) grant no. 39. The work of Mathew Magimai Doss was partly funded by the Swiss National Science Foundation under the National Centre of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2).

References

[1] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[2] B. Gold and M. Nelson, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley, 2000.

[3] R. Patterson and I. N. Smith, “An efficient auditory filterbank based on the gammatone function,,” *Speech-Group meeting of the Institute of Acoustics on Auditory Modelling*, vol. 54, Apr 1987.

[4] M. Slaney, “An efficient implementation of the patterson-holdsworth auditory filterbank,” *Apple computer technical report No.35*, 1993.

[5] X. Yang, K. Wang, and S. A. Shamma, “Auditory representations of acoustic signals,” *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824 –839, mar 1992.

[6] S. Mallat, *A Wavelet Tour of Signal Processing, 3rd ed.* Academic Press, Dec.

[7] L. Solbach, R. Wöhrmann, and J. Kliewer, “The complex-valued continuous wavelet transform as a preprocessor for auditory scene analysis,” in *Proceedings of IJCAI−95, Workshop on Auditory Scene Analysis, Montreal, Canada*, Aug. 1995.

[8] A. Venkitaraman, A. Adiga, and C. S. Seelamantula, “Auditory-motivated gammatone wavelet transform,” *Accepted for publication in Signal Processing*.

[9] A. Mertins and J. Rademacher, “Vocal tract length invariant features for automatic speech recognition,” in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005*, nov. 2005, pp. 308 –312.

[10] Q. Li and Y. Huang, “An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1791–1801, 2011.

[11] M. Slaney, “Auditory toolbox — version 2.0,” *Apple computer technical report No.45*, 1993.

[12] H. G. Hirsch and D. Pearce, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *Proceedings of ISCA ITRW ASR*, Aug. 2000, pp. 181–188.

[13] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[14] D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, “Evaluation of a noise-robust dsr front-end on aurora database,” in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP ’02)*, sep. 2002, pp. 17–20.

[15] J. Tchorz and B. Kollmeier, “A model of auditory perception as front end for automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 106, pp. 2040–50, Oct 1999.

[16] Z. Tuske, P. Golik, R. Schluter, and F. R. Drepper, “Non-stationary feature extraction for automatic speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 5204 – 5207.

[17] R. Schluter, L. Bezrukov, H. Wagner, and H. Ney, “Gammatone features and feature combination for large vocabulary speech recognition,” in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, april 2007, pp. IV–649 –IV–652.