ORIGINAL RESEARCH

A robust BFCC feature extraction for ASR system

Ta-Wen Kuan[1], An-Chao Tsai*[2], Po-Hsun Sung[1], Jhing-Fa Wang[1,2], Hsien-Shun Kuo[1]

[1] *National Cheng-Kung University, Tainan City, Taiwan*
[2] *Tajen University, Pingtung County, Taiwan*

Received: October 10, 2015 **Accepted:** December 9, 2015 **Online Published:** January 5, 2016
DOI: 10.5430/air.v5n2p14 **URL:** http://dx.doi.org/10.5430/air.v5n2p14

ABSTRACT

An auditory-based feature extraction algorithm naming the Basilar-membrane Frequency-band Cepstral Coefficient (BFCC) is proposed to increase the robustness for automatic speech recognition. Compared to Fourier spectrogram based of the Mel-Frequency Cepstral Coefficient (MFCC) method, the proposed BFCC method engages an auditory spectrogram based on a gammachirp wavelet transform to simulate the auditory response of human inner ear to improve the noise immunity. In addition, the Hidden Markov Model (HMM) is used for evaluating the proposed BFCC in phases of training and testing purposes conducted by AURORA-2 corpus with different Signal-to-Noise Ratios (SNRs) degrees of datasets. The experimental results indicate the proposed BFCC, compared with MFCC, Gammatone Wavelet Cepstral Coefficient (GWCC), and Gammatone Frequency Cepstral Coefficient (GFCC), improves the speech recognition rate by 13%, 17%, and 0.5% respectively, on average given speech samples with SNRs ranging from -5 to 20 dB.

Key Words: Gammachirp filterbank, Speech recognition, Cepstral coefficients, Auditory modeling, Basilar-membrane Frequency-band Cepstral Coefficient, Mel-Frequency Cepstral Coefficient, Noise robustness

1. INTRODUCTION

The automatic speech recognition (ASR) system nowadays sophistically and widely deployed in various and numerous of smart devices and electronics for many applications, for example, google voice search on laptop or smartphone are popularly used for information retrieval, navigation, internet shopping, entertainment *etc.*, the smart home appliances controlled by voice and the identity authentication for internet bank or smart home by voice-print recognition *etc.*, however, to against noise interruption in ASR still being drawn much attention for investigators to improve ASR robustness.

In this paper, a new feature extraction based on a gammachirp filterbank is proposed to improve the noise robustness for ASR, wherein Discrete Cosine Transform (DCT) operation, mel-filter bank and energy calculation in Mel-Frequency Cepstral Coefficient (MFCC)[1] are replaced by the procedures of the gammachirp filterbank generation, the cochlear wavelet transformation and the auditory spectrogram[2] in the proposed namely Basilar-membrane Frequency-band Cepstral Coefficient (BFCC) method, which is thereafter used to evaluate the ASR accuracy performance in the cases of the noise environments, for example, people who uses ASR system on smart phone for information retrieval while being interrupted by the chatting passengers surrounding in a car or a subway system, or people who is in an museum hall or a concert hall to search the corresponding information regarding the seeing object or listening music by voice or music search at smart phone. Accordingly, four kinds of noises including suburban train, babble, car, and exhibition hall, from the AURORA-2 database are used for evaluating the proposed BFCC.

* **Correspondence:** An-Chao Tsai; Email: actsai@tajen.edu.tw; Address: Department of Multimedia Design, Tajen University, No.20, Weixin Rd., Yanpu Township, Pingtung County 90741, Taiwan

Compared to the traditional auditory filter, the grmmachirp[3,4] filter was demonstrated for performing the smart candidate for an asymmetric and layer dependency filterbank in simulating the auditory processing, particularly to the noise speech data. For the grmmachirp filter provided the excellent characteristics of notched-noise masking data to improve the noise robustness in an ASR system. Although MFCC was investigated the sophisticated performance in speech recognition system, the properties of Fourier transform and the triangular mel-filterbank in MFCC were shown unlikely the sound wave sensitivity at basilar-membrane in human auditory system, and gave the less robustness in the presence of additive noise. In other works, suppression of the additive noise interruption for ASR by using spectral subtraction methods[5] and by Wiener filters[6] were ever well-investigated.

Many models were examined the frequency resolution of the basilar membrane functions within the cochlea of human inner ear. Most of these models used a constant-Q filterbank to represent the membrane. Among them, the gammatone function[7,8] was indicated the better approximation of the experimentally determined auditory response. Accordingly, Adiga *et al.*[8] proposed a robust speech recognition system based on Gammatone Wavelet filterbank Cepstral Coefficients (GWCC), of which the steps in the feature extraction processes are mostly the same as MFCC. Similarly, Shao *et al.*[9] proposed a robust speech identification system against noise based on Gammatone Frequency Cepstral Coefficients (GFCC). Yang *et al.*[10] inspected that the functions of the basilar membrane in the human ear that can be viewed as an affine wavelet transform. Therefore, the filterbank used in auditory-based speech recognition systems[11–13] should also possess a wavelet property. However, in general, to construct the Continuous Wavelet Transform (CWT) that requires an infinite number of translations of the wavelet function; in this case, this proposed work only uses a finite set of filters for transformation. In Ref.,[14] the auditory-based transformation approach realized a robust feature extraction algorithm for speaker identification under mismatched conditions that achieved the greater identification accuracy than MFCC.

In this work, an auditory-based feature extraction method naming BFCC based on the basilar membrane functions within the cochlea of human inner ear is proposed. The AURORA-2 database is conducted in the experiments, and HTK (HMM tool kit) is applied for model training and testing. Four kinds of noise datasets including, suburban train, babble, car, and exhibition hall, mixed in different SNRs degree from -5 dB to 20 dB step 5 dB that are used for system training and testing. The experimental results indicate the average word accuracy of the proposed BFCC method is higher than that of MFCC about 13.6%, and also higher than that of GWCC and GFCC about 17.3% and 0.5%, respectively.

The rest of this work is organized as follows. Section 2 described the detailed procedures and comparison among the BFCC framework with the related works. Experiments setting and results with relating comparison and analysis are discussed in Section 3. Conclusion is drawn in Section 4.

2. BASILAR-MEMBRANE FREQUENCY-BAND CEPSTRAL COEFFICIENT

The flowchart of the proposed BFCC method which is interrupted by noise surrounding environment for ASR system is shown in Figure 1. Initially, in training phase, the input speech with additional noise is transformed to the auditory spectrogram[15,16] and calculated by the proposed BFCC method, next to be trained by HMM through HTK toolkit, whereas in the recognition phase, the input speech is collected and transformed to auditory spectrogram, and the extracted feature is then yielded through BFCC prior to Viterbi algorithm for speech recognition. In order to outperform the accuracy and robustness of the proposed BFCC feature extraction compared to the other related approaches, the following sections elucidate the proposed BFCC, GWCC, CFCC and MFCC, with the comparison conducted and discussed in the experimental results in Section 3.

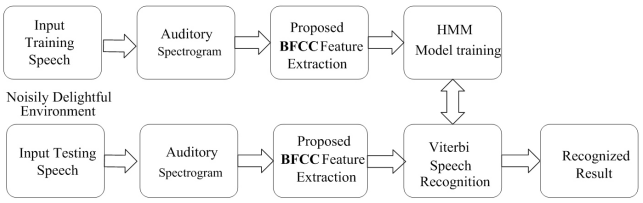

Figure 1. Overview of BFCC feature extraction processes for ASR system

2.1 BFCC feature extraction method

The basic notion for basilar-membrane function in cochlea of human inner ears is the traveling wave of motion performed in basilar membrane, the characteristic frequency at a specific point of membrane parameters is determined along its length, in which the widest and least stiff at the apex of the cochlea senses the high frequency, on the contrary, the narrowest and most stiff location perceives the low frequency. Such characteristics give the motivation to propose the BFCC feature extraction to improve the robustness in an ASR. The fundamental flowchart for the proposed BFCC is shown in Figure 2, wherein the human inner ear can be modeled by a gammachirp filter[17] for the high frequency selectivity performance. Initially, the proposed BFCC normalizes the speech signal in accordance with (1),

$$s[n] = (x[n] - x_{mean})/x_{\max} \tag{1}$$

where n is the sample index, $x[n]$ is the speech signal, x_{mean} is the mean intensity of the speech signal, and $x_{\max}$ is the maximum intensity of the speech signal.

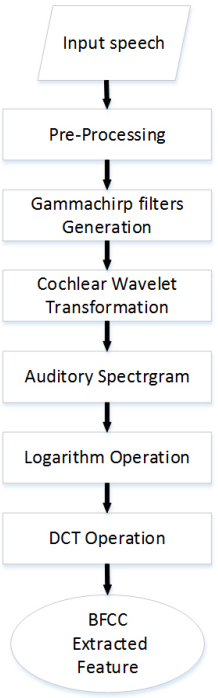

Figure 2. Flowchart of the proposed BFCC feature extraction Method

Having normalized the speech signal, the gammachirp wavelet kernel is then built. The gammachirp function is derived from the gammatone function[18] which is the impulse responses in cochlear[19] simulated by the reverse-correlation approach,[3,20] the magnitude response of gammatone filter masked the data to acquire the magnitude response in the auditory filter psychophysically. Patterson *et al.*[20] investigated that the gammatone filter performed nearly the same characteristics as the rounded exponential auditory filter[21,22] in magnitude response. Above examinations indicated that the gammatone auditory filter was expected to simulate the time-domain features as human cochlear filtering. Beside, an asymmetric magnitude response in a filter gave a "chirp" in a carrier term known as the response in the basilar membrane examined by Irino *et al.*,[17] the carrier term in gammatone function was then yielded a gammachrip auditory filter to simulate the level-dependent asymmetry.[4]

In the proposed BFCC, the bandwidth of each filter is designed as the corresponding Equivalent Rectangular Bandwidth (ERB) measured by psychoacoustics to approximate the bandwidths of filter as human auditory. Note that the ERB at any frequency f (in Hz) can be calculated, and the ERB is specifically chosen here since it provides a close approximation of the filter bandwidths as the human auditory system. The function for gammachirp filterbank thus can be represented as (2),

$$g_c[i,n] = \left(\frac{j-n}{b_i}\right)^{(\sigma-1)} \exp\left[-j2\pi\lambda \cdot ERB(f_c)\cdot\left(\frac{j-n}{b_i}\right)\right] \cdot\exp\left[j2\pi f_c\left(\frac{j-n}{b_i}\right) + j\varepsilon\ln\left(\frac{j-n}{b_i}\right) + \phi\right] \tag{2}$$

where σ and λ are parameters corresponding to the envelope in the gamma distribution, f_c is the central frequency of each filter, ε is the chirp term, ϕ is the phase, n is the sample index, j is the sample location variable, and b_i is the scale factor of each frequency bin i. When $\varepsilon = 0$, the gammatone function is generated as (3),

$$g_t[i,n] = \left(\frac{j-n}{b_i}\right)^{(\sigma-1)} \exp\left[-j2\pi\lambda \cdot ERB(f_c)\cdot\left(\frac{j-n}{b_i}\right)\right] \cdot\exp\left[j2\pi f_c\left(\frac{j-n}{b_i}\right) + \phi\right] \tag{3}$$

The simulation paradigms of the impulse response in gammatone filter and gammachirp filter are shown in Figures 3 and 4, respectively, with eight channels of filter.

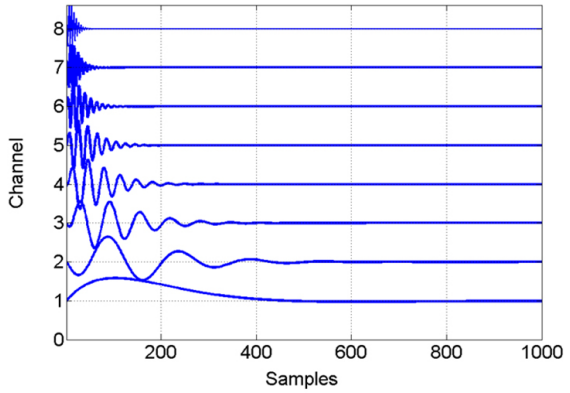

Figure 3. Impulse response of the gammatone filter

After the signal $s[n]$ is normalized, then the gammachirp filterbank $g_c[i,n]$ is built and used in the proposed BFCC method followed by the cochlear wavelet transform (CWT) as (4),

$$CW[i,n] = \sum_{i=1}^{I} \frac{1}{\sqrt{b_i}} g_c[i,n] * s[n] \tag{4}$$

where $CW[i,n]$ is the complex-valued output of the discrete CWT, i is the bin index of frequency, I is the summed frequency of bins and n is the signal sample. The absolute values $|CW[i,n]|$ of the outcome is derived from (4) for all i collected from the auditory spectrogram. The spectrogram is partitioned as a method used in MFCC to the windowing process. Thereafter, the partitioned spectrogram next input into a logarithm function and DCT operation to extract BFCC features. In this work, $g(l)$ in DCT operation is defined as (5),

$$g(l)=\sum_{k=0}^{N-1}c(k)\sqrt{\frac{2}{N}}\cos\left[\frac{(2l+1)k\pi}{2N}\right],l=0,1,\cdots,N-1 \tag{5}$$

where $c(k)$ is shown as (6),

$$c(k)=\begin{cases}\sqrt{1/2} & for\ k=0\\ 1 & for\ k=1,2,\cdots,N-1\end{cases} \tag{6}$$

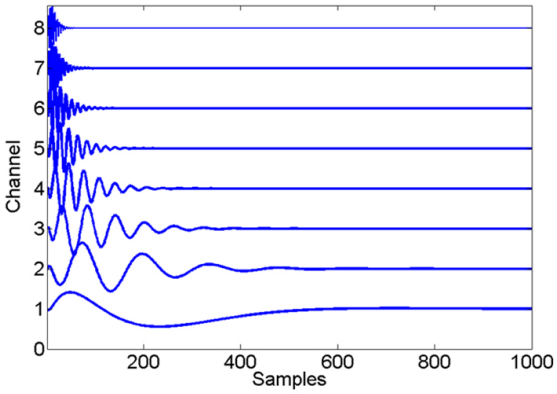

Figure 4. Impulse response of the gammachirp filter

After DCT operation finished, the BFCC features are acquired.

For $CW[i,n]$ is the complex-valued output in the discrete CWT, the spatial information can be included in the phase part, such a case is very useful for collecting the sound data in the case of multi-microphones. The differences of feature extraction procedures between MFCC and the proposed BFCC are shown in Figure 5.

2.2 Gammatone wavelet cepstral coefficients

The gammatone wavelet cepstral coefficient abbreviated as GWCC[8] was proposed by Adiga *et al.* that treated as a feature extraction method based on human auditory perception, the gammatone wavelet used in GWCC is derived from the popular gammatone functions to develop a robust feature extraction algorithm against noise. The wavelets combined the characteristics of human peripheral auditory system, particularly in the spatially-varying frequency response of the basilar membrane. The steps involved in GWCC extraction method for input speech are similar to MFCC technique having the different filterbank used, wherein GWCC replaced the mel-filter bank in MFCC by using a gammatone-wavelet filterbank.

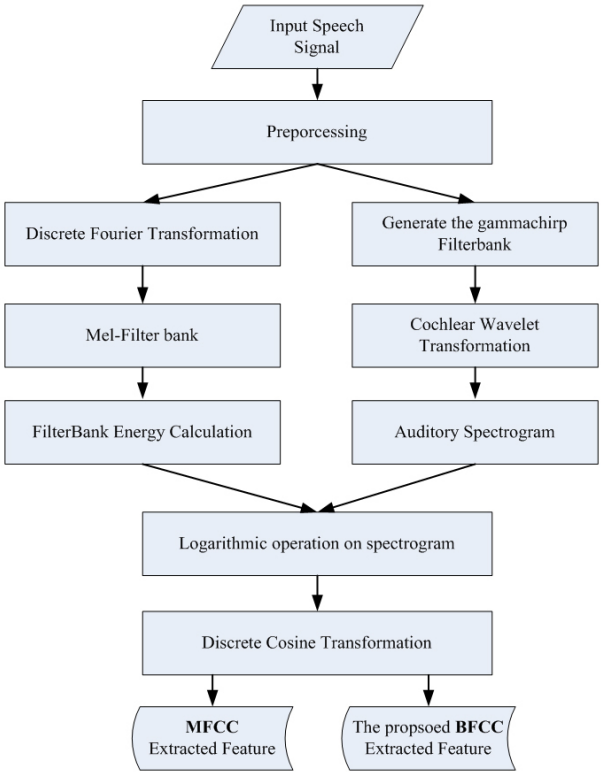

Figure 5. Comparison of feature extraction methods between MFCC and the proposed BFCC

The gammatone function shown in (7) is a sinusoid tone modulated by a gamma distribution function,

$$g(t)=t^{(N-1)}e^{-\alpha t}e^{j\omega_c t}u(t) \tag{7}$$

where t (in seconds) denotes time, w_c is the center frequency (in radians/second), α is the bandwidth parameter which determines the effective duration of $g(t)$, $u(t)$ is the unit step function, and N is the order controls the rise and decay of the function. If N is parameterized in the range 3 to 5, the gammatone function provided a great similarity as the basilar membrane responses. The gammatone function could not be seen as a wavelet because it does not satisfy an important gauge for wavelet analysis, namely the admissibility condition that FT of the wavelet should vanish at zero frequency. The Fourier transform of $g(t)$ is as (8),

$$\hat{g}(\omega)=\frac{(N-1)!}{\left(\alpha+j(\omega-\omega_c)\right)^N} \tag{8}$$

where w (in radians/second) is the angular frequency.

The derivative of the gammatone is a straightforward Fourier transform expression given by (9)

$$\hat{\psi}(\omega)=j\omega\cdot\hat{g}(\omega)=\frac{j\omega\cdot(N-1)!}{\left(\alpha+j(\omega-\omega_c)\right)^N} \tag{9}$$

In order to stabilize the resulting wavelets to the existing and corresponding WT, n cannot exceed the order parameter N of the gammatone function.[20] Therefore, the gammatone wavelets function in the time domain could be presented as (10):

$$\begin{aligned}\psi(t)&=\frac{d}{dt}\left\{t^{(N-1)}e^{-\alpha t}e^{j\omega_c t}u(t)\right\}\\&=\left((N-1)t^{(N-2)}+\beta t^{(N-1)}\right)e^{\beta t}u(t)\end{aligned} \tag{10}$$

where $\beta=-\alpha+e^{jwc}$.

In the construction of the gammatone filterbank, bandwidth of each filter is taken as the corresponding ERB. The equation chosen for calculating ERB (in Hz) at any frequency f (in Hz) could be expressed as (11):

$$ERB(f)=\frac{f}{9.26}+24.7 \tag{11}$$

The calculation of center frequency f_c for a channel k is based on (12):

$$f_c(k)=-C+e^{k\log\left(\frac{f_{\min}+C}{f_{\max}+C}\right)/K}\cdot(f_{\max}+C) \tag{12}$$

where $1\le k\le K$, K is the total number of filters, $C=228.83$, $f_{\min}$ and $f_{\max}$ are the lowest and highest cutoff frequencies of the filterbank.

As mentioned above, the feature extraction step is similar to MFCC, however, only the applied filterbank is different, wherein the gammatone wavelets are applied in GFCC rather than the wavelet transformation used in MFCC to extract the input speech into the features. The flowchart of GWCC compared to MFCC and the proposed BFCC is shown in Figure 6.

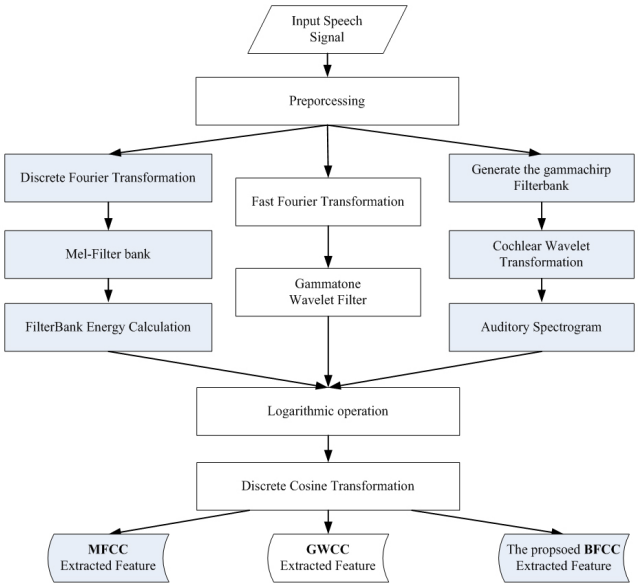

Figure 6. Flow chart of GWCC compared to MFCC and the proposed BFCC

2.3 Cochlear frequency cepstral coefficients

The cochlear filter cepstral coefficients named CFCC[14] was proposed by Q. Li *et al.*, to improve the robustness for the speaker recognition. An auditory transformation was applied to the modules imitating the signal processing function in human cochlea; CFCC improved the robustness to noise with the benefits of only using the real number for calculation and separating the frequency at any linear or nonlinear scale.

Given $f(t)$ as the input speech signal, a cochlear filter $\varphi(t)$ regarding the basilar membrane impulse response for convolution with $f(t)$ transformation is displayed in (13):

$$\begin{aligned}T(a,b)&=f(t)*\frac{1}{\sqrt{|a|}}\varphi\left(\frac{t-b}{a}\right)dt \quad \text{or}\\T(a,b)&=f(t)*\varphi_{a,b}(t)\,dt\end{aligned} \tag{13}$$

where a and b are real numbers, and $T(a,b)$ is the output relating the decomposed signal in basilar membrane, and $\varphi_{a,b}(t)$ is in (14)

$$\varphi_{a,b}(t)=\frac{1}{\sqrt{|a|}}\varphi\left(\frac{t-b}{a}\right) \tag{14}$$

In (14), the factor a is a scale parameter, where b is a time shift argument, and the critical potion in the transformation is the cochlear filter displayed as (15).

$$\Re_{a,b}(t)=\frac{1}{\sqrt{|a|}}\left(\frac{t-b}{a}\right)^{\alpha}\bullet\exp\left[-2\pi f_L\beta\left(\frac{t-b}{a}\right)\right]$$
$$\times\cos\left[2\pi f_L\left(\frac{t-b}{a}\right)+\theta\right]h(t-b) \quad (15)$$

where α and β indicate the shape and width of the cochlear filter in the frequency domain, $h(t)$ is the unit step function. After the cochlear filter done, the input speech is processed by the hair cell in (16),

$$w(a,b)=T(a,b)^2;\forall T(a,b) \quad (16)$$

Next, the duration of the count relating the current band central frequency given by (17):

$$S(i,j)=\frac{1}{d}\sum_{b=l}^{l+d-1}w(i,b),\ l=1,L,2L,\ldots;\forall i,j \quad (17)$$

where d is the window length, after the hair cell processed, the logarithm operation is replaced by the cubic root and output for DCT operation to generate the CFCC. The difference of flowcharts among CFCC, MFCC, the proposed BFCC and GWCC are shown in Figure 7. However, CFCC without using the filterbank thus excluded in the experiments, for CFCC performed the totally different processes among the participated feature extraction methods.

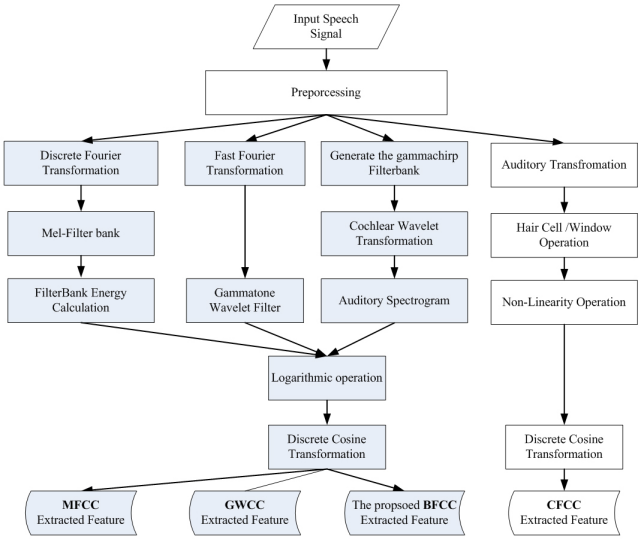

Figure 7. The flowchart of CFCC compared to MFCC, the proposed BFCC and GWCC, however CFCC is not participated the experiments for the totally different processing flowchart among three approaches

3. Experimental results and comparison

3.1 Corpus setup

To evaluate the robustness and performance of the proposed BFCC method in two cases of clean and noisy conditions, further to compare with the multiple feature extraction methods, such as MFCC, GWCC and CFCC, the AURORA-2 database is chosen for the experiments. The speech model is trained by different SNR degrees (-5 to 20 dB step -5 dB) of mixed dataset comprising clean and noisy data. To generate the noisy data, the clean dataset is further mixed with four types of noise data, including (1) suburban train, (2) babble, (3) car, and (4) exhibition hall. Totally, 8,440 utterances are yielded for training purposes and split equally into 20 subsets, each SNR subset contains 422 utterances and the sampling rate is 8 kHz, and each subset including one clean data with five SNR types of noisy data, that is, 5 dB, 10 dB, 15 dB, and 20 dB, respectively. In the test part, four types of noisy data with different SNRs on -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB are also built. Each SNR subset contains 1,001 utterances and totally 24,024 utterances are generated for testing. The entire mixed training and testing data are presented in a short sentence. Each sentence is composed of one to seven words. The words are zero to nine digits and an alphabet "O". The detailed corpus setup with data type, categories and characteristics are shown in Table 1.

Table 1. Corpus Setup

Data type	Categories	Characteristics
Training Dataset	Data type	Clean and Noisy
	Sampling rate	8 kHz
	Number of Speakers	55 males and 55 females
	Noise types	Suburban train, babble, car, exhibition hall
	SNR Types	Clean, 5 dB, 10 dB, 15 dB, 20 dB
	Utterances Number/SNR	422
	Total utterances	8,440
Testing Dataset	Data type	Noisy
	Sampling rate	8 kHz
	Number of Speakers	55 males and 55 females
	Noise types	Suburban train, babble, car, exhibition hall
	SNR Types	-5 to 20 dB step 5 dB
	Utterances Number/SNR	1,001
	Total utterances	24,024

The speech recognizer used for evaluating multiple feature extraction methods are based on HTK for training and recognition as in Figure 8. Particularly, GFCC herein is involved

for the comparison with other three methods, due to GFCC almost performed the same processing as BFCC, wherein the gammatone filter in GFCC are replaced by gammachirp filter in the proposed BFCC. In addition, the frequency band is divided into 64 sub-bands named the cochlear frequency instantaneous frequency (CFIF) thus having 64 dimensions and plus the Cochlear Frequency Spectrogram Energy (CFSE), totally 104 dimensions are summed for testing namely BFCC+CIF+CFSE.

To train HMM models, the lexicon, syllable labels and training corpus are in need of well preparation prior to deal with the static features with two estimated dynamic features (Δ and $\Delta\Delta$) for each digit. Note that a whole-word model is used to train for each digit, and silence as well as space models are also trained in the process either. Thus, a total of 13 word models are involved. The recognition performance of the proposed scheme is evaluated in terms of the word accuracy and compared with that of the related methods for benchmarking purposes.

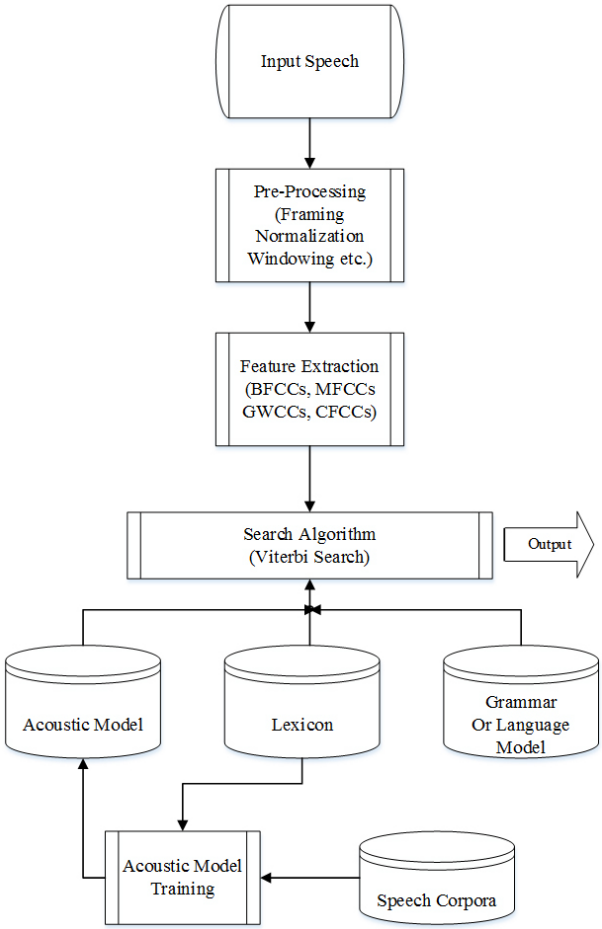

Figure 8. Training/recognition flowchart of speech recognizer based on HTK

3.2 Experimental results and comparison

In the experiments, the examples based on mono AURORA-2 of digit words are shown in Table 2 for evaluating the performance and robustness among MFCC, GWCC, GFCC and the proposed BFCC under different SNRs degrees.

Table 2. Corpus types and pronunciation in AURORA-2

Index	Contents
Content of Digit	One to nine and zero
Content of Letter	O
Example1	2203 (Pronunciation)
Example2	3467701 (Pronunciation)

Figure 9 is shown the case of word accuracy under suburban-train noisy data among five feature extraction methods, that is, MFCC, BFCC, GFCC, GWCC and BFCC+CFSE+CFIF. The observation is indicated that the proposed BFCC and GFCC both are almost performed up to 90% of accuracy than other three methods, wherein MFCC is shown the worst case of accuracy in any SNRs. Compared with MFCC, the proposed BFCC improved up to 23% of word accuracy, and 18% of higher accuracy than GWCC.

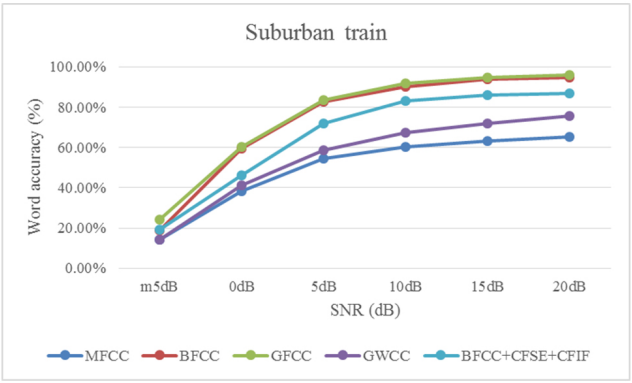

Figure 9. AURORA-2 word accuracy of suburban train noisy data

In the case of babble noisy data in Figure 10, the experimental result indicates MFCC which performed the better accuracy than other methods between -5 dB and 0 dB, however, when SNR arises more than 10 dB, GWCC performs the better one than other four methods. The proposed BFCC herein is shown the similar accuracy performance as MFCC from -5 to 0 dB, and gives the equivalent performance as GWCC from 10 to 20 dB. This implies the proposed BFCC having more robustness in babble noisy case than other four methods in any SNRs, and all methods' accuracy tend to statures at 90% when SNRs are arising over than 10dB.

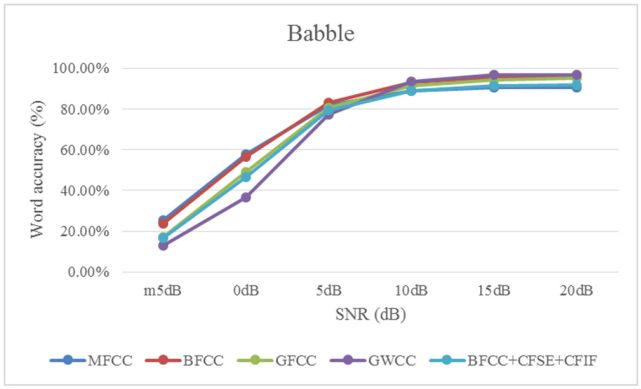

Figure 10. AURORA-2 word accuracy of babble noisy data

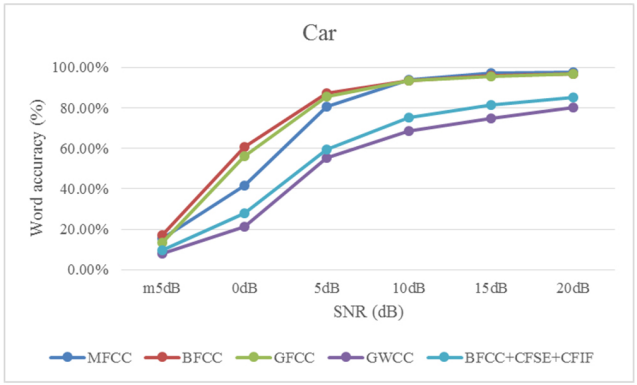

Figure 11. AURORA-2 word accuracy of car noisy data

Figure 11 indicates the case of car noisy data, wherein the proposed BFCC performs the better recognition accuracy in the cases of 15 dB, 0 dB, 5 dB and 10 dB, and MFCC and GFCC reach the same accuracy as the proposed BFCC at 15 dB and 20 dB, respectively. Figure 12 is shown the case of accuracy performance in exhibition noisy data, in which the proposed BFCC gives the nearest same accuracy as GFCC in any SNR conditions, such an accuracy distribution looks like the case of babble noisy data in Figure 10 among four approaches, for example, MFCC is shown the worst case in any SNR conditions and the accuracy of BFCC+CFSE+CFIF performs the middle accuracy among four participated methods as case of suburban train noisy data in Figure 9. This can be explained that the noise effects of suburban train and exhibition hall on feature extraction having the same influence on accuracy performance.

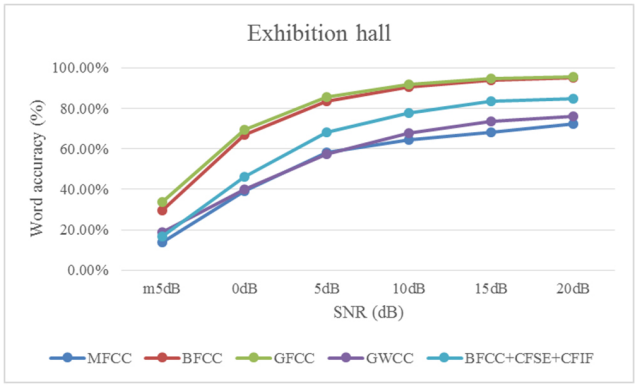

Figure 12. AURORA-2 word accuracy of exhibition hall

Table 3. AURORA-2 average word accuracy of four types of noisy data

Noise types & SNRs Methods	**Four types of noisy data including car, babble, exhibition hall and suburb train**					
	SNR-5 dB	**SNR-0 dB**	**SNR-5 dB**	**SNR-10 dB**	**SNR-15 dB**	**SNR-20 dB**
	Average Word Accuracy					
MFCC	17.31%	44.22%	68.72%	76.84%	79.79%	81.48%
BFCC	22.44%	60.93%	84.15%	91.81%	94.81%	95.62%
GFCC	22.20%	58.76%	83.79%	92.11%	94.80%	95.82%
GWCC	13.45%	34.73%	62.12%	74.33%	79.27%	82.10%
BFCC+CFSE+CFIF	15.58%	41.67%	69.65%	81.15%	85.54%	87.05%

In light of above experimental results, the average word accuracy can be observed in Figure 13 and Table 3 among all participated methods as well as the multiple noisy datasets in different SNRs condition, accordingly, the results indicate the proposed BFCC gives the approximate accuracy performance as GFCC, and GWCC performs the worst case averagely, all the accuracies tend to saturate when SNR up to 15 dB. Compared to MFCC, the proposed BFCC improves the speech recognition rate by 13% averagely in cases of SNRs ranging from -5 to 20 dB. Furthermore, compared to the GWCC, the proposed scheme remarkably improves the speech recognition rate by 17% averagely, with SNRs ranging from -5 to 20 dB.

3.3 Analysis and discussion

In the experimental results, the word accuracy of BFCC is almost equivalent to that of MFCC in the case of babble noise at the SNRs lower than 10 dB in Figure 9 and at the

case of car noise when SNRs higher than 10 dB in Figure 10, however, the difference between two methods are only about 1%. Moreover, the word accuracy is observed in babble noise case that BFCC is higher than that of MFCC about 5% with SNRs ranging from 5 to 20 dB in Figure 10. Particularly, in the case of car noise, the word accuracy of BFCC is remarkably superior to MFCC over than 19% at 0 dB SNR, and over 20% of higher word accuracy in the cases of suburban train and exhibition hall noise types at any SNRs. The poorer performance of MFCC method in two cases of the suburban train and exhibition hall noise types is due to the frequent erroneous recognition of the silence part of utterances regarded as a digit.

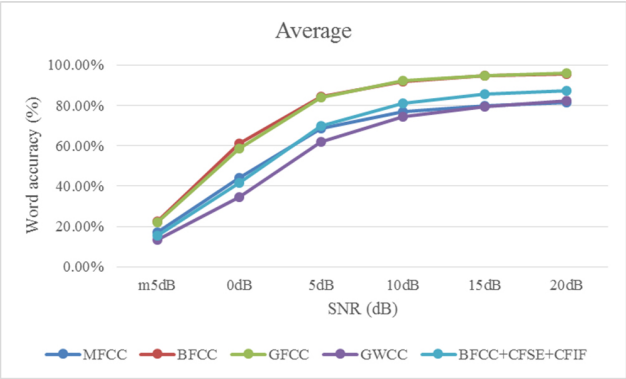

Figure 13. AURORA-2 average word accuracy of all four type noisy data

In addition, Fast Fourier Transform (FFT) in MFCC is computationally efficient. However, the signal characteristic is very different from that of the human auditory system. For instance, the use of fixed-length windows causes the entire speech band in the pitch harmonics. Furthermore, the frequency bands are also linearly distributed in FFT, whereas the bands are non-linearly distributed in the human cochlea. Compared to FFT, the wavelet transformation approach used in BFCC led to a sharper time resolution which contains the larger ratio of high frequency components than the low frequency components,[23] and therefore generates an auditory spectrogram to closely resemble the human auditory system.

4. Conclusion

In this paper, a BFCC feature extraction algorithm based on a gammachirp filter and the cochlear wavelet transformation is proposed for the robustness improvement in ASR system. The experimental results indicate that the word accuracy of the proposed BFCC method is averagely higher than that of MFCC and GWCC about 13% and 17% respectively, through the testing over 24,024 samples in a different SNRs dataset ranging from -5 to 20 dB step 5 dB. The superior performance of the proposed BFCC method is analytically attributed to the characteristics of the gammachirp filterbank, of which the functions closely resemble the cochlear filterbank in the human ear, and the wavelet transformation approach nearly simulates the human ear in splitting the frequency band thus performing a greater sensitivity to the low frequency components.

Acknowledgements

This work was supported by the MOST, Taiwan under project number D104-36A06.

References

[1] Han W, Chan CF, Choy CS, *et al*. An efficient MFCC extraction method in speech recognition. IEEE International Symposium on Circuits and Systems (ISCAS 2006) p. 4.

[2] Solbach L, Wohrmann R, Kliewer J. The complex-valued continuous wavelet transform as a preprocessor for auditory scene analysis. Readings in Computational Auditory Scene Analysis. Erlbaum Publishers; 1998. p.273-92.

[3] Irino T, Patterson RD. A time-domain, level-dependent auditory filter: The gammachirp. The Journal of the Acoustical Society of America. 1997; 101(1): 412-9. http://dx.doi.org/10.1121/1.417975

[4] Lutfi RA, Patterson RD. On the growth of masking asymmetry with stimulus intensity. The Journal of the Acoustical Society of America. 1984; 76(3): 739-45. PMid:6491046. http://dx.doi.org/10.1121/1.391260

[5] Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions Acoustics, Speech and Signal Processing. 1979; 27(2): 113-20. http://dx.doi.org/10.1109/TASSP.1979.1163209

[6] Goldstein JS, Reed SI, Scharf LL. A multistage representation of the Wiener filter based on orthogonal projections. IEEE Transactions on Information Theory. 1998; 44(7): 2943-59. http://dx.doi.org/10.1109/18.737524

[7] Patterson R, Smith IN. An efficient auditory filterbank based on the gammatone function. Speech-Group meeting of the Institute of Acoustics on Auditory Modeling; 1987. p.54.

[8] Adiga A, Magimai M, Seelamantula CS. Gammatone wavelet Cepstral Coefficients for robust speech recognition. IEEE Region 10 Conference on TENCON; 2013. p. 1-4; 22-25.

[9] Shao Y, Srinivasan S, Wang D. Incorporating Auditory Feature Uncertainties in Robust Speaker Identification. Acoustics, Speech and Signal Processing. International Conference on Acoustics, Speech and Signal Processing. ICASSP. Hawaii. 2007; 4: 277-80.

[10] Yang X, Wang K, Shamma SA. Auditory representations of acoustic signals. IEEE Transactions on Information Theory. 1992; 38(2): 824-39. http://dx.doi.org/10.1109/18.119739

[11] Li Q. An auditory-based transform for audio signal processing. IEEE Workshop on Applications of Signal Processing to Audio and Acous-

tics. WASPAA; 2009. p. 181-4.

[12] Mertins A, Rademacher J. Vocal tract length invariant features for automatic speech recognition. IEEE Workshop on Automatic Speech Recognition and Understanding. 2005: 308-12.

[13] Hanson B, Applebaum TH. Robust speaker- independent word recognition using static dynamic and acceleration features: experiments with Lombard and noisy speech. International conference on acoustics, speech and signal processing. ICASSP; 1990. p. 857-60.

[14] Li Q, Huang Y. An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. IEEE Transactions on Audio, Speech, and Language Processing. 2011; 19(6): 1791-801. http://dx.doi.org/10.1109/TASL.2010.2101594

[15] Schofield D. Visualizations of speech based on a model of the peripheral auditory system. NASA STI/Recon Technical Report. 1985; 86: 17593.

[16] Patterson RD. Auditory filter shapes derived with noise stimuli. The Journal of the Acoustical Society of America. 1976; 59(3): 640-54. http://dx.doi.org/10.1121/1.380914

[17] Irino T, Patterson RD. A compressive gammachirp auditory filter for both physiological and psychophysical data. The Journal of the Acoustical Society of America. 2001; 109(5): 2008-22. http://dx.doi.org/10.1121/1.1367253

[18] Venkitaraman A, Adiga A, Seelamantula CS. Auditory-motivated gammatone wavelet transform. Signal Processing. 2014; 94: 608-19. http://dx.doi.org/10.1016/j.sigpro.2013.07.029

[19] Carney LH, McDuffy MJ, Shekhter I. Frequency glides in the impulse responses of auditory-nerve fibers. The Journal of the Acoustical Society of America. 1990; 105(4): 2384-91. http://dx.doi.org/10.1121/1.426843

[20] Patterson RD, *et al*. An efficient auditory filterbank based on the gammatone function. Meeting of the IOC Speech Group on Auditory Modeling at RSRE. 1987; 2(7).

[21] Moore BCJ, Peters RW, Glasberg BR. Auditory filter shapes at low center frequencies. The Journal of the Acoustical Society of America. 1990; 88(1): 132-40. PMid:2380441. http://dx.doi.org/10.1121/1.399960

[22] Rosen S, Baker RJ. Characterizing auditory filter nonlinearity. Hearing research. 1994; 73(2): 231-43. http://dx.doi.org/10.1016/0378-5955(94)90239-9

[23] Daubechies I. The wavelet transform, time-frequency localization and signal analysis. IEEE Transactions on Information Theory. 1990; 36(5): 961-1005. http://dx.doi.org/10.1109/18.57199