An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification under Mismatched Conditions

Qi Li, *Senior Member, IEEE* and Yan Huang, *Member, IEEE*

***Abstract*— An auditory-based feature extraction algorithm is presented. We name the new features as cochlear filter cepstral coefficients (CFCC) which are defined based on a recently developed auditory transform (AT) plus a set of modules to emulate the signal processing functions in the cochlea. The CFCC features are applied to a speaker identification task to address the acoustic mismatch problem between training and testing environments. Usually, the performance of acoustic models trained in clean speech drops significantly when tested in noisy speech. The CFCC features have shown strong robustness in this kind of situation. In our experiments, the CFCC features consistently perform better than the baseline MFCC features under all three mismatched testing conditions – white noise, car noise, and babble noise. For example, in clean conditions, both MFCC and CFCC features perform similarly, over 96%, but when the SNR of the input signal is 6 dB, the accuracy of the MFCC features drops to 41.2%, while the CFCC features still achieve an accuracy of 88.3%. The proposed CFCC features also compare favorably to PLP and RASTA-PLP features. The CFCC features consistently perform much better than PLP. Under white noise, the CFCC features are significantly better than RASTA-PLP, while under car and babble noise, the CFCC features provide similar performances to RASTA-PLP.**

***Index Terms*— Feature extraction algorithm, auditory-based features, automatic speaker recognition, robust speaker recognition, speaker identification, cochlea.**

I. Introduction

FEATURE extraction is the first crucial component in automatic speech processing. Generally speaking, successful front-end features should carry enough discriminative information for classification or recognition, fit well with the back-end modeling, and be robust with respect to the changes of acoustic environments. To the best of our knowledge, obtaining a satisfactory system performance under various operating modes still remains problematic, especially when acoustic training and testing environments are mismatched. Since the human hearing system is robust to the mismatched conditions, we propose an auditory-based feature extraction algorithm that is modeled on the basic signal processing functions in the ear. The proposed algorithm is also based on our recently developed auditory-based time-frequency transform named Auditory Transform (AT) [1], [2]. The features generated from the proposed algorithm are named cochlear filter cepstral coefficients (CFCC).

This work was supported by the US AFRL under the contract number FA8750-08-C0028.

Q. Li is with Li Creative Technologies, Inc., 25 B Hanover Road, Suite 140, Florham Park, NJ 07932, USA. Tel: (973) 822-0048, Fax: (973) 822-0399 (Email: qili@ieee.org).

Y. Huang was with Li Creative Technologies, Inc., 25 B Hanover Road, Suite 140, Florham Park, NJ 07932, USA. She is now with Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA (Email: yanhuang@ieee.org).

A. Traditional Speech Feature Extraction and the Fourier Analysis

At a high level, most speech feature extraction methods fall into the following two categories: modeling the human voice production system or modeling the peripheral auditory system.

For the first approach, one of the most popular features is a group of cepstral coefficients derived from linear prediction known as the linear prediction cepstral coefficients (LPCC) [3], [4]. The LPCC feature extraction utilizes an all-pole filter to model the human vocal tract with speech formants captured by the poles of the all-pole filter. The narrow band (e.g., up to 4 KHz) LPCC features work well in a clean environment. However, in our previous experiments, the linear predictive spectral envelope shows large spectral distortion in noisy environments [5], [6]. This results in significant performance degradation.

For the second approach, there are two groups of features, based on either Fourier transforms (FT) or auditory-based transforms. Representative for the first group are the MFCCs (Mel frequency cepstral coefficients), where a fast Fourier transform (FFT) is applied to generate the spectrum in the linear scale, and then a bank of band-pass filters is placed along a Mel frequency scale on top of the FFT output [7]. Alternatively, the FFT output is warped to a Mel or Bark scale and then a bank of band-pass filters is placed linearly on top of the warped FFT output [5], [6]. The proposed algorithm in this paper belongs to the second group, where the auditory-based transform is defined as an invertible, time-frequency transform. The output from this kind of transform can be in any kind of frequency scale (e.g., linear, Bark, ERB, etc). Therefore, there is no need to place the band-pass filter in a Mel scale as in the MFCC or warp the frequency distributions as in [5], [6].

The MFCC features [7] in the first group are one of the most popular features for speech and speaker recognition. Like the LPCC features, the MFCC features perform well in clean environments but not in adverse environments or mismatched training and testing conditions. Perceptual linear predictive (PLP) analysis is another peripheral auditory-based approach.

Based on the FFT output, it uses several perceptually motivated transforms, including Bark frequency, equal-loudness preemphasis, and cubic-root amplitude compression [8]. The relative spectra, known as RASTA, is further developed to filter the time trajectory to suppress constant factors in the spectral component [9]. It is often cascaded with the PLP feature extraction to form the RASTA-PLP features. Comparisons between MFCC and RASTA-PLP have been reported in [10]. Further comparisons with the proposed CFCC features in experiments will be given at the end of this paper.

Both MFCC and RASTA-PLP features are based on the Fourier transform (FT). As mentioned above, the FT has a fixed time-frequency resolution and a well-defined inverse transform. Fast algorithms exist for both the forward transform and the inverse transform. Despite its simplicity and efficient computation algorithms, we believe that when applied to speech processing the time-frequency decomposition mechanism of the FT is different than the mechanism in the hearing system. First, it uses fixed-length windows, which generate pitch harmonics over the entire speech bands. Secondly, its individual frequency bands are distributed linearly, which is different from the distribution in the human cochlea. Further wrapping is needed to convert to the Bark, MEL, or other scales. Finally, in our recent study in [1], [2], we observed that the FFT spectrogram has more noise distortion and computation noise than an auditory-based transform which we recently developed. One of the examples is shown in Fig. 4. Thus, we find it necessary to develop a new feature extraction algorithm based on the new auditory-based, time-frequency transform [2] to replace the FT in speech feature extraction.

B. Auditory-Based Time-Frequency Analysis

The traveling wave of the basilar membrane (BM) in the cochlea and its impulse response have been observed and reported in the literature, such as [11], [12], [13], [14], [15], [16], [17]. Moreover, the BM tuning and auditory filters have also been studied in the literature [18], [19], [20], [21], [22], [23]. Many electronic and mathematic models have been defined to simulate the traveling wave, the auditory filters, and the frequency responses of the BM [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34].

Based on the study of the human hearing system, Li proposed an auditory-based, time-frequency transform (AT) in [1], [2]. The new transform is comprised of a pairing of a forward transform and an inverse transform. Through the forward transform, the speech signal can be decomposed into a number of frequency bands using a bank of cochlear filters. The frequency distribution of the cochlear filters is similar to the one in the cochlea and the impulse response of the filters is similar to that of the travelling wave. Through the inverse transform, the original speech signal can be reconstructed from the decomposed band-pass signals. In [2], Li has presented the proof of the inverse transform of the AT and validated the inverse AT in experiments.

Compared to the FFT, the AT has flexible time-frequency resolution and its frequency distribution can take on any linear or nonlinear form. Therefore, it is easy to define the distribution to be similar to that of the Bark, Mel, or ERB scale, which is similar to the frequency distribution function of the Basilar membrane. Most importantly, the proposed transform has significant advantages in noise robustness and can be free of the pitch harmonic distortion as plotted in [2] and Fig. 4. Therefore, the AT provides a new platform for feature extraction research. It forms the foundation for our robust feature extraction algorithm.

In summary, the ultimate goal of this study is to develop a practical, front-end speech feature extraction algorithm that conceptually emulates the human peripheral hearing system and uses the concept to achieve an improved noise robustness performance under mismatched training and testing conditions.

The remainder of this paper is organized as follows: Section II demonstrates the proposed auditory feature extraction algorithm and provides an analytic study and discussion; Section III studies the feature parameters using a development dataset and presents the experimental results of the proposed CFCC in comparison to other front-end features in a testing dataset; finally, Section IV concludes the paper.

II. Proposed Auditory-Based Feature Extraction Algorithm

This section describes the structure of the proposed auditory-based feature extraction algorithm and provides details of its computation. Although we would like to emulate the human peripheral hearing system, the computational aspects must meet the requirements of real-time applications; therefore, we will simulate only the most important features of the human peripheral hearing system.

An illustrative block diagram of the proposed algorithm is shown in Fig. 1. The proposed algorithm is intended to conceptually replicate the hearing system at a high level and consists of the following modules: auditory transform implemented by a cochlear filter bank, hair-cell function with windowing, cubic-root nonlinearity, and discrete cosine transform (DCT). A detailed description of each module follows.

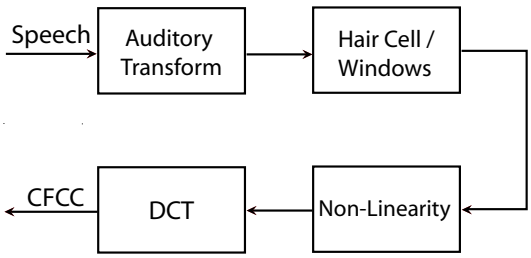

Fig. 1. Schematic diagram of the proposed auditory-based feature extraction algorithm named cochlear filter cepstral coefficients (CFCC).

A. The Auditory Transform

The auditory transform in Fig. 1 is the forward transform of a pair of invertible auditory-based transforms, as defined and described by Li in [2]. It can be implemented as a filter bank. As the foundation of the auditory-based feature extraction algorithm, we use the forward auditory transform to replace the Fast Fourier transform used in many other

features. The auditory transform models the traveling wave in the cochlea where the sound waveform is decomposed into a set of subband signals.

Let $f(t)$ be a speech signal. A transform of $f(t)$ with respect to a cochlear filter $\psi(t)$, representing the basilar membrane (BM) impulse response in the cochlea, is defined as:

$$T(a,b) = f(t) * \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right)\ dt, \tag{1}$$

where $*$ denotes the convolution operation, a and b are real, both $f(t)$ and $\psi(t)$ belong to $\mathbf{L}^2(R)$, and $T(a,b)$ representing the traveling waves in the BM is the decomposed signal and filter output. The above equation can also be written as:

$$T(a,b) = f(t) * \psi_{a,b}(t)\ dt, \tag{2}$$

where

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right). \tag{3}$$

Like in the wavelet transform, the factor a is a scale or dilation variable. By changing a, we can shift the central frequency of ψ to receive a band of decomposed signals. Factor b is a time shift or translation variable. For a given value of a, factor b shifts the function $\psi_{a,0}(t)$ by an amount b along the time axis.

Note that $1/\sqrt{|a|}$ is an energy normalizing factor. It ensures that the energy stays the same for all a and b; therefore, we have:

$$\int_{-\infty}^{\infty} |\psi_{a,b}(t)|^2\ dt = \int_{-\infty}^{\infty} |\psi(t)|^2\ dt. \tag{4}$$

The cochlear filter, as the most important part of the transform, is defined as:

$$\begin{aligned}\psi_{a,b}(t) &= \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right)\\ &= \frac{1}{\sqrt{|a|}}\left(\frac{t-b}{a}\right)^{\alpha}\exp\left[-2\pi f_L\beta\left(\frac{t-b}{a}\right)\right]\\ &\quad \cos\left[2\pi f_L\left(\frac{t-b}{a}\right)+\theta\right]u(t-b),\end{aligned} \tag{5}$$

where $\alpha > 0$ and $\beta > 0$, $u(t)$ is the unit step function; i.e. $u(t) = 1$ for $t \geq 0$ and 0 otherwise. Parameters α and β determine the shape and width of the cochlear filter in the frequency domain. They can be empirically optimized as shown in our experiments in Section III. The value of θ should be selected such that (6) is satisfied:

$$\int_{-\infty}^{\infty} \psi(t)\ dt = 0. \tag{6}$$

This is required by the transform theory to ensure no information is lost during the transform [2]. The value of a can be determined by the current filter central frequency, f_c, and the lowest central frequency, f_L, in the cochlear filter bank:

$$a = f_L/f_c. \tag{7}$$

Since we contract $\psi_{a,b}(t)$ with the lowest frequency along the time axis, the value of a is in the range $0 < a \leq 1$. If we stretch ψ, the value of a would be constrained to $a > 1$. The frequency distribution of the cochlear filter can be in the form of linear or nonlinear scales such as ERB (equivalent rectangular bandwidth) [26], Bark [35], Mel scale [7], or log. For a particular band number i, the corresponding value of a is represented as a_i, which needs to be pre-calculated for the required central frequency of the cochlear filters at band number i.

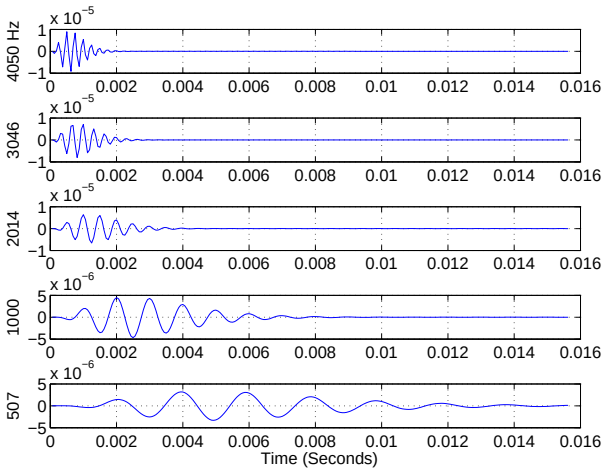

Fig. 2. Impulse responses of the BM in the auditory transform (AT) when $\alpha = 3$ and $\beta = 0.2$, plotted by (5). The labels on the far left of each subplot represent the central frequency of the plotted impulse response. They are very similar to psychological measurements, such as the figures in [11], [12], [36] (Fig. 1.12), [13], etc.

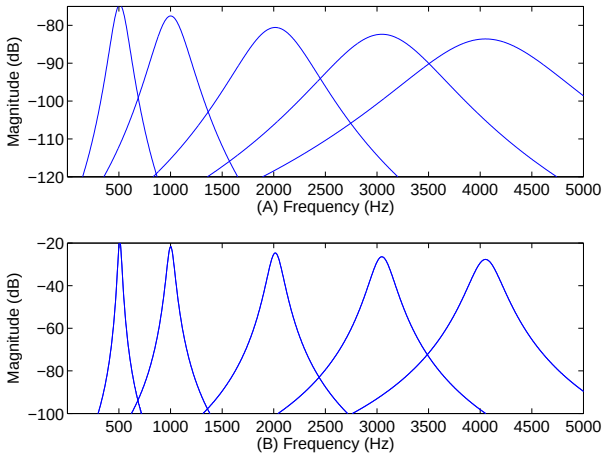

Fig. 3. The frequency responses of the cochlear filters when $\alpha = 3$: (A) $\beta = 0.2$; and (B) $\beta = 0.035$. The filter band width can be adjusted by β for different applications.

Fig. 2 shows the impulse responses for five cochlear filters plotted using (5), which are similar to the psychoacoustic experiment results, such as the impulse responses plotted in [11], [13]. Fig. 3 shows the corresponding frequency responses. Normally, we use $\alpha = 3$. The value of β controls the filter band width; i.e. the Q-factor. This makes our auditory transform (AT) different than the Gammatone function [37] in which the Q-factor is fixed.

We note that the inverse transform of the above transform exists. It has been proven mathematically and validated experimentally [2]. This property ensures that the forward transform implemented by the cochlear filter bank can avoid any information loss and thus qualifies as a platform for feature

extraction.

B. Other Operations

The cochlear filter bank is intended to emulate the impulse response in the cochlea. However, there are other operations in the ear. The inner hair cells act as a transducer for mechanical movements of the BM into neural activities. When the BM moves up and down, a shearing motion is created between the BM and the tectorial membrane [36]. It causes the displacement of the uppermost hair cells which generates the neural signals. However, the hair cells only generate the neural signals in one direction of the BM movement. When the BM moves in the opposite direction, there is neither excitation nor neuron output. We studied different implementations of the hair cell function. The following function of the hair cell output provides the best performance in our evaluated task:

$$h(a,b) = T(a,b)^2; \quad \forall\ T(a,b), \tag{8}$$

where $T(a,b)$ is the filter-bank output from (1). Here, we assume that all other detailed functions in the outer ear, middle ear, and the control of the neural system to the cochlea have been ignored or have been included in the auditory filter responses.

In the next step, the hair cell output for each band is converted into a representation of nerve spike count density. The duration of the count can be associated with the current band central frequency. We use the following equation to mimic the concept:

$$S(i,j) = \frac{1}{d}\sum_{b=\ell}^{\ell+d-1} h(i,b), \quad \ell = 1, L, 2L, \cdots; \quad \forall\ i,j, \tag{9}$$

where $d = \max\{3.5\tau_i, 20\text{ms}\}$ is the window length, τ_i is the period of the central frequency of the ith band, and $L = 10$ ms is the window shift duration. We empirically set the system parameters, but they may need to be adjusted for different datasets. Instead of using a fixed length window as in the FFT, we are using a variable length window for different frequency bands. The higher the frequency, the shorter the window. This prevents the high-frequency information from being smoothed out by a long window duration. The output of the above equation and the spectrogram of the cochlear filter bank can be used for both feature extraction and analysis.

Furthermore, we apply the scales of loudness function suggested by Stevens [38], [39] to the hair cell output as:

$$y(i,j) = S(i,j)^{1/3}. \tag{10}$$

In the last step, the discrete cosine transform (DCT) is applied to decorrelate the feature dimensions and to generate the cochlear filter cepstral coefficients (CFCCs), so the features can work with the existing back-end.

C. Analysis and Comparison

We provide a comparative analysis of the auditory transform (AT) and the well-known Fourier transform (FT), and then extend the comparison to the features derived from the AT, such as the CFCCs, and from the FT, such as the MFCCs.

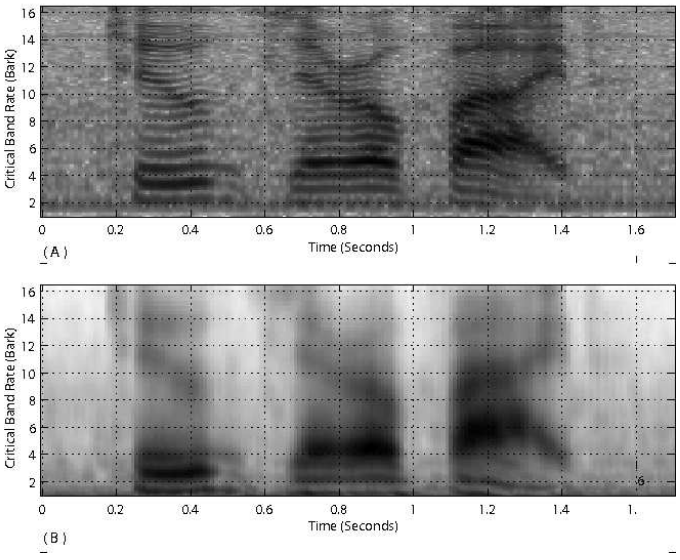

Fig. 4. Comparison of FT and AT spectrums: (A) The FFT spectrogram of a male voice "2 0 5", warped into the Bark scale from 0 to 6.4 Barks (0 to 3500 KHz). (B) The spectrogram from the cochlear filter output for the same male voice. The proposed AT is harmonic free and has less noise.

The analysis and discussion are intended to help the reader understand the CFCCs. Further comparisons will be made in the next section.

1) Comparison between AT and FT: The fast Fourier transform (FFT) is the major tool for the time-frequency transform used in speech signal processing. We use Fig. 4 to illustrate the differences between the spectrograms generated from the Fourier transform and our auditory transform [2]. The original speech wave file is recorded from a male voice. We then calculated the FFT spectrograms as shown in Fig. 4 (A), with 30 ms Hamming window shifting every 10 ms. To facilitate the comparison, we then warped the frequency distribution from linear scale to the Bark scale using the method in [6].

The spectrogram of our auditory transform is shown in Fig. 4 (B). It was generated from the output of the cochlear filter bank as defined in (5) and uses a window of fixed duration to compute the average densities for each band. In comparing the two spectrograms in Fig. 4, we can observe that there are no pitch harmonics and there is less computational noise in the spectrums generated from the auditory transform. In addition, all formant information has been kept. This is due to the variable length of cochlear filters and the selection of parameter β in (5). The harmonics in FFT spectrogram are due to the fixed window length for all frequency bands.

Furthermore, we compared the spectrums shown in Fig. 5. A male voice was recorded in a moving car using two different microphones. A close-talking microphone was placed on the speaker's lapel, and a hands-free microphone was placed on the car visor. Fig. 5 is one of the spectrums from Fig. 4 at 1.15 second time frame. The solid line represents speech recorded by the close-talking microphone, the dashed line corresponds to speech recorded by the hands-free microphone. Fig. 5 (top) is the spectrum from our auditory-based transform [2] and Fig. 5 (bottom) is from the Fourier transform. From Fig. 5, we can

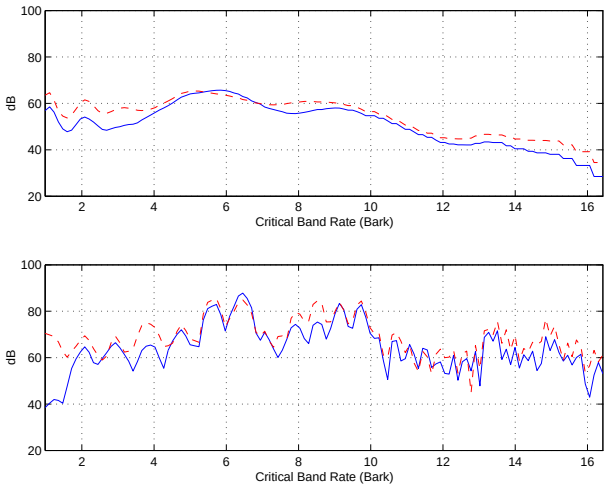

Fig. 5. Comparison of AT (top) and FFT (bottom) spectrums at the 1.15 second time frame for robustness: The solid line represents the speech from a close-talking microphone. The dashed line represents a hands-free microphone mounted on the visor of a moving car. Both speech files were recorded simultaneously. The FFT spectrum shows 30 dB distortion at low-frequency bands due to background noise compared to the AT.

observe the following in the FFT spectrum, which are not as significant in the AT spectrum:

- Distortion from background noise: The FFT spectrums show a 30 dB distortion at low-frequency bands due to the car background noise.
- Pitch harmonics: The FFT spectrums show significant pitch harmonics, which is due to the fixed length of the FFT window for all frequency bands. In AT computation, the length of the impose response of the band-pass filters is variable. It is shorter for high frequency and longer for low frequency.
- Computation noise: The noise displayed as "snow" in Fig. 4 (A) was generated by the FFT computation.

For robust speaker identification, we do need a more robust time-frequency transform as the foundation for feature extraction. The transform should generate less distortion from background noise and less computation noise from selected algorithms, such as pitch harmonics, while also retaining the useful information. Here, the auditory transform provides a robust solution to replace the Fourier transform.

2) Comparison between CFCCs and MFCCs: Since the MFCC features are popular features in both speaker and speech recognition, we compare the proposed CFCCs with the MFCCs as follows:

It is understood that the MFCC features use the FFT to convert the time domain speech signal to the frequency domain spectrum. The power spectrum is calculated and then triangle filters are applied to produce filter bank energy estimates. The triangle filters are distributed in the Mel scale. In contrast, the proposed CFCC features use a bank of cochlear filters to decompose the speech signal into multiple bands. The frequency response of a cochlear filter has a bell-like shape rather than a triangle shape. The shape and width (the Q-factor) of the filter in the frequency domain can be adjusted by parameters α and β from (5). In each of the bands, the decomposed signal is still in the time domain, represented by real numbers. The central frequencies of the cochlear filters can be arranged in any distribution, including Mel, ERB, Bark, or log.

When using the FFT to compute a spectrogram, the window size must be fixed to all frequency bands, due to the fixed point FFT. When we compute a spectrogram from the decomposed signals generated by the cochlear filters, the window size can be different for different frequency bands. For example, we use a longer window for a lower frequency band to average out the background noise and a shorter window for a higher frequency band to protect high-frequency information. Furthermore, the MFCCs use a logarithm as the nonlinearity while the CFCCs use a cubic root.

3) Comparison between CFCCs and Gammatone-Based Feature: The Gammatone frequency cepstral coefficients (GFCC) are also auditory-based speech features [40]. We introduce it briefly, so we can compare it in our experiments later. The GFCC features use a Gammatone filter bank to replace the Fourier analysis and includes down sampling, cubic root, and DCT operations.

An exact implementation following the description in [40] did not give us reasonable experimental results. To remedy the outcome, we then replaced the "downsampling" procedure in [40] by computing an average of the absolute values on the Gammatone filter-bank output using a 20 ms window shift every 10 ms, followed by a cubic root function and DCT. This procedure gave us the best results in our experiments, but because it is different from the original GFCCs, we have named it *modified GFCC* (MGFCC) features. Since this paper presents the concept of using an auditory-based filter bank as an alternative to the FFT, we consider MGFCCs to be an additional result to support the concept, and as such we will report our experimental results in subsequent sections.

We note that the Gammatone function in [41] is different than the AT cochlear filter in (5). The Gammatone filter bandwidth, (the Q-factor), is locked in to its central frequency and cannot be adjusted, while the filter bandwidth in the AT (5) can be influenced by parameter β. As shown in our experiments, the speaker identification performance can be changed when adjusting the parameter β. Also, unlike the proposed AT, there is no proof of the existence of an inverse transform of the Gammatone filter bank to ensure that there is no information loss in the forward transform.

III. Experimental Evaluation

This section presents the experimental evaluation of the CFCC features for text-independent speaker identification using a Gaussian mixture model (GMM) back-end. The CFCC/GMM system was evaluated in a task where the acoustic conditions of training and testing are mismatched, i.e. the training data set was recorded under a clean condition while the testing data sets were mixed with different types of background noise at various noise levels.

The experimental study has four tasks. We first establish the baseline system which represents the current MFCC/GMM system performance. Then a series of analytic studies on

each component of the CFCCs is conducted to optimize the CFCC feature extraction using a development dataset. Next, the CFCC features are evaluated on the test dataset and compared to the MFCC and MGFCC features. Furthermore, we also compare the CFCC features with the PLP/RASPA-PLP features on the same task.

A. Experimental Datasets

Our speaker recognition experiments started from the NTIMIT database [42]. We used a subset of 38 speakers as a development dataset and another subset of 460 speakers as a testing dataset. There was no overlap between these two datasets. Each speaker has eight utterances for training and two utterances for testing. We developed the CFCC feature extraction algorithm and determined the feature parameters from the development dataset. We then applied the developed CFCCs to the testing dataset. We achieved 3.47% relative improvement over the baseline MFCC features under matched conditions on the testing dataset. However, the NTIMIT database cannot show the performance on the mismatched conditions. We then use the Speech Separation Challenge (SSC) database [43] to report our research results because the database has several mismatched conditions. Also, this allows us to compare our results with other reported results on the same database.

For a fair comparison, we adjusted the MFCC parameters on the development datasets for both databases to the best performance. The adjusted parameters include the number of cepstral coefficients and whether or not to use the cepstral energy term and cepstral mean subtraction. While we adjusted the CFCC parameters slightly, the feature extraction structure and the procedure of the feature extraction computation remains the same from the NTIMIT to the SSC databases. Actually, as readers can find in the following report, compared to the difference caused by the feature extraction structure and algorithm, the improvement from the parameter adjustment on CFCCs, such as β and window size, is very limited.

The Speech Separation Challenge database contains speech recorded from a closed-set of 34 speakers (18 male and 16 female speakers). All speech files are single-channel data sampled at 25 kHz and all material is end-pointed (i.e. there is little or no initial or final silence) [43]. The training data was recorded under clean conditions. The testing sets were obtained by mixing clean testing utterances with white noise at different SNR levels; in total there are five testing conditions provided in the database (i.e. noisy speech at -12 dB, -6 dB, 0 dB, and 6 dB SNR, and clean speech). We find this database ideal for the study of noise robustness when training and testing conditions do not match. In particular, since all the noisy testing data is generated from the same speech with only the noise level changing, this largely reduces the performance fluctuations due to variations other than noise types and mixing levels.

In our experiments speaker models were first trained using the clean training set and then tested on noisy speech at four SNR levels. We created three disjoint subsets from the database as the training set, development set, and testing set. Each set has 34 speakers and there is no overlap of speakers across the training, development, and testing sets. We note that since our feature parameters had been turned on the NTIMIT database, the main purpose of the development dataset is to show the effects of each of the adjustable parameters on the overall system performance.

The training set has 20 utterances per speaker and 680 utterances in total. The average duration of training data per speaker is 36.8 seconds of speech. The development set has 1700 utterances in total. There are five testing conditions (i.e. noisy speech at -12 dB, -6 dB, 0 dB, and 6 dB SNR, and clean speech). Each condition has 10 utterances per speaker. The average duration of each utterance is 1.8 seconds. The development set is only with white noise. The testing set has the same five testing conditions. Each condition has 10 to 20 utterances per speaker. The duration of each testing utterance is about 2 to 3 seconds of speech. The testing set has about 2500 utterances for each noise type. For three types of noises, white, car, and babble, we have about 7500 utterances in total for testing.

Note that the training set consists of only clean speech, while both the development set and the testing set consist of clean speech and noisy speech at five different SNR levels. We mainly focused on 0 dB and 6 dB SNR conditions in our feature analysis and comparisons because when conditions are under -6 dB SNR the performance of all features is close to random.

We note that in addition to white noise testing conditions provided in the Speech Challenge database, we also generated two more sets of testing conditions with car noise or babble noise at -6 dB, 0 dB, and 6 dB SNR. The car noise and babble noise were recorded under real-world conditions, and mixed with the clean test speech from the Speech Separation Challenge database. These test sets were used as additional material to further test the robustness of the proposed auditory features. The testing set sizes, with different types of noise, are the same.

B. The Baseline System

Our baseline system uses the standard MFCC front-end features and Gaussian Mixture Models (GMMs). Twenty-dimensional MFCC features (c1 $\sim$ c20) were extracted from the speech audio based on a 25 ms window with a frame-rate of 10 ms; the frequency analysis range was set to be 50 Hz $\sim$ 8000 Hz. Note that the delta and double delta of the MFCCs were not used here since they were not found to be helpful in discerning between speakers in our experiments. We also found cepstrum mean subtraction was not helpful for both clean and mismatched data; therefore it was not used in our baseline system.

The back-end of the baseline system is the standard GMMs trained using the maximum likelihood estimation (MLE) [44]. Let M_i represent the GMM model for the i-th speaker, and i be the index for speakers. During testing, the testing utterances u match against all hypothesized speaker models (M_i), and the speaker identification decision (J) is made by:

$$J = \arg\max_i \sum_k \log p(u_k|M_i), \tag{11}$$

where u_k is the k-th frame of utterance u and $p(\cdot|M_i)$ is the probability density function. Thirty-two Gaussian mixture components were used in the speaker GMM models. To obtain a fair comparison of the different front-end features, only the front-end feature extraction was varied and the configuration of the back-end of the system remained the same in all the experiments throughout this paper.

C. Analytic Study Using a Development Set

To better understand and optimize the various components of the CFCC feature extraction, we delved into each module in the CFCC feature extraction and experimented with its alternative variations using a separate development set as described in Section III-A. The goal was to determine the effects of each component on the overall performance and ultimately optimize the feature extraction. Specifically, we investigated the effects of the filter width (β), various windowing schemes, with/without equal loudness, and two different nonlinearity schemes. The analytic study was performed on noisy speech with white noise at 0 dB and 6 dB SNR levels.

1) Effect of Filter Bandwidth (β): The first step of the cochlear feature calculation is to pass the speech audio through a band-pass filter bank as described in (5), in which β is varied to adjust the filter bandwidth. We experimented with different β values and empirically optimized its value according to the speaker identification accuracy performance.

Fig. 6 shows the speaker identification accuracy of the CFCC features with different filter bandwidth (β). We found that when $\beta = 0.035$, the CFCC has the best performance for speaker identification.

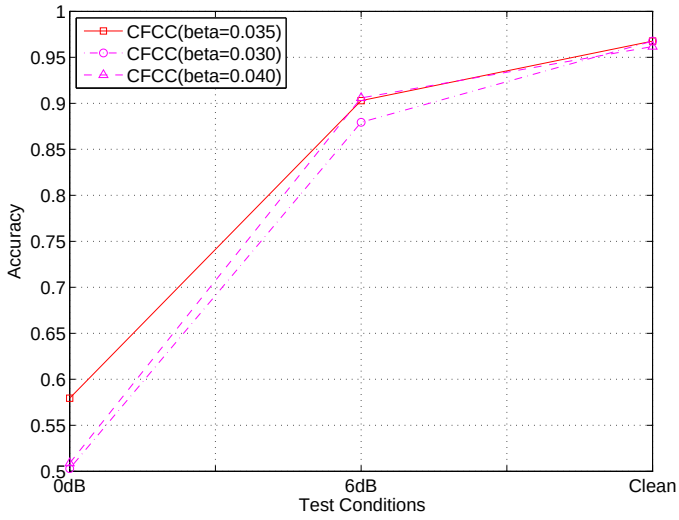

Fig. 6. Speaker identification accuracies of the auditory-based cochlear features (CFCC) with different filter bandwidth adjusted by parameter (β).

2) Effect of Equal-Loudness: The loudness of a sound is a function of both the intensity and the frequency spectrum of a sound stimulus. For pure tone or narrow-band noise, the equal-loudness contour measures the sound intensity across frequency bands needed in order to invoke a sensation of equal-loudness magnitude [45]. The equal-loudness level contours are intended to reflect the frequency characteristics of the human auditory system. To simulate the human loudness perception in our proposed auditory-based feature extraction, we weighted each channel of the filter-bank output by an equal-loudness function, which gives different weights to different frequency bands [46].

Fig. 7 shows a comparison of CFCC systems with or without using the equal-loudness function. It can be seen that the system with equal-loudness weighting consistently performs better than the one without equal-loudness weighting on all testing conditions.

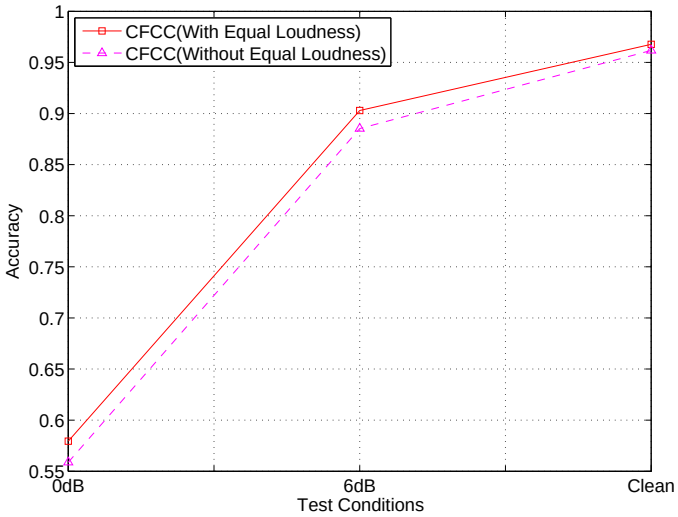

Fig. 7. Speaker identification accuracy results of the auditory-based cochlear features (CFCC) with/without equal loudness.

3) Effect of Various Windowing Schemes: As shown in Fig. 1, after speech is decomposed into travelling waves, a hair cell function with a certain window size is applied to the traveling waves at each frequency band. We experimented with three different types of windowing schemes. The first one is the fixed-length window typically used in many feature extraction approaches. The second one takes into account the multi-resolution characteristics of the Cochlear transform and uses a fixed-epoch window at different frequency bands. The second scheme is more flexible; however, serious data leaking problems can occur at high-frequency bands. For example (assuming the sampling rate is 16 kHz and the target rate of the feature extraction is 10 ms, or 160 samples), when we use the window at the size of 3.5 epochs of the central frequency at each frequency band, the window size at the frequency band with a central frequency of 4 kHz would be 14 samples. That is much smaller than the target rate of 10 ms or 160 samples. To mitigate the data leaking problem, the third approach combines the first two windowing schemes. In low-frequency bands, a fixed-epoch window is used; as the central frequency increases and data-leaking problems start to occur, the fixed-length window is applied.

Fig. 8 shows a comparison of CFCC systems with three different kinds of windowing schemes. It can be seen that a combination of the fixed-length and fixed-epoch window gives the best performance. The fixed-epoch window does not perform as well, which might be due to the aforementioned

data leaking problem.

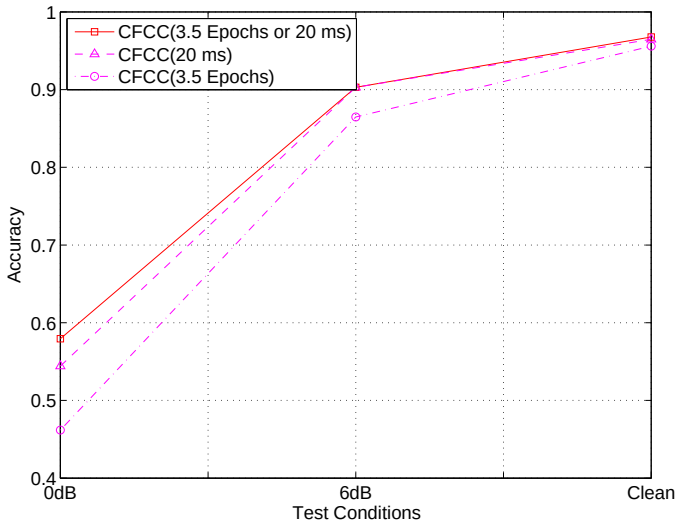

Fig. 8. Speaker identification accuracy results of the auditory-based cochlear features (CFCC) with a fixed-length window (20 ms), fixed-epoch window (3.5 epochs), or a combination of the fixed-length and fixed-epoch window (3.5 epochs or 20 ms).

4) Effect of nonlinearity: As shown in Fig. 1, after the windowing/averaging procedure, a nonlinearity is applied to the simulate the nonlinearity in the human auditory system. We experimented with both the logarithm and cubic-root nonlinearities and empirically found that the cubic-root nonlinearity outperforms the logarithm under all noisy testing conditions as shown in Fig. 9.

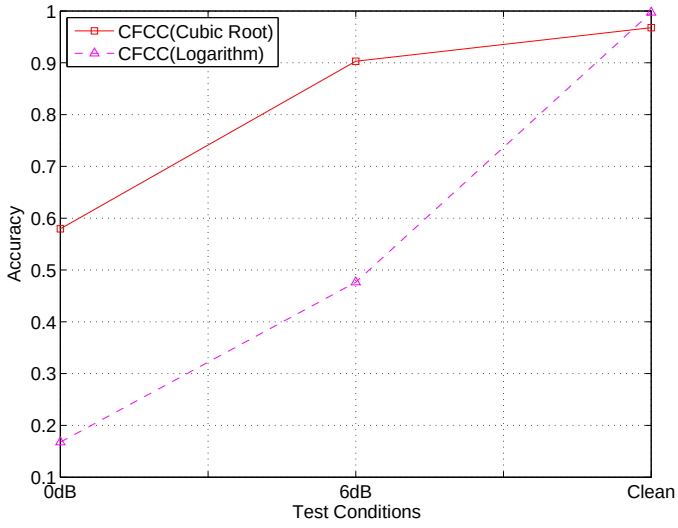

Fig. 9. Speaker identification accuracy results of the auditory-based cochlear features (CFCC) with logarithm and cubic nonlinearity.

It is interesting to observe that the cubic-root nonlinearity performs better than the logarithm, which might have to do with the warping and suppression effect of the cubic function on the noise components. In comparison, the logarithm has a high variance for changes at low energy.

5) Summary of Experimental Study with the Development Dataset: Based on the previous analytic study, the CFCC feature extraction can be summarized as follows: First, the speech audio file is passed through the band-pass filter bank. The filter width parameter β was set to 0.035. The Bark scale is used for the filter bank distribution and equal-loudness weighting is applied at different frequency bands. Second, the travelling waves generated from the cochlear filters are windowed and averaged by the hair cell function. The window length is 3.5 epochs of the band central frequency or 20 ms, whichever is the shortest. Third, a cubic root is applied. Finally, since most back-end systems adopt diagonal covariance based GMM or HMM models, the discrete cosine transform (DCT) is used to decorrelate the features. The 0th component, related to the energy, is removed from the DCT output.

Table I shows a comparison of the speaker identification accuracy of the optimized CFCC features with the MGFCCs and MFCCs tested on the development set.

TABLE I
COMPARISON OF MFCC, MGFCC, AND PROPOSED CFCC FEATURES TESTED ON THE DEVELOPMENT SET.

Testing SNR	-6 dB	0 dB	6 dB
MFCC	6.8%	15.9%	42.1%
MGFCC	9.1%	45.0%	88.8%
CFCC (Proposed)	12.6%	57.9%	90.3%

D. Final Experiments Using a Testing Dataset

Using the optimized CFCC feature extraction based on the development set, we conducted speaker identification experiments on the testing set with the results depicted in Fig. 10. As we can see from Fig. 10, in clean testing conditions, the CFCC features generated comparable results to MFCC features and achieved over 96% accuracy. As white noise is added to the clean testing data at increasing intensity, the performance of the CFCCs is significantly better than both the MGFCCs and MFCCs. For example, when the SNR of the testing condition drops to 6dB, the accuracy of the MFCC system drops to 41.2%. In comparison, the parallel system using the proposed CFCC features still achieves 88.3% accuracy, more than twice as accurate as the MFCC features. Similarly, the MGFCC features have an accuracy of 85.1%, which is better than the MFCC features, but not as good as the proposed CFCC features. The CFCC performance in the testing data set is similar to its performance in the development set. Overall, we see that the proposed CFCC features significantly outperform both the widely used MFCC features and another related auditory-based MGFCC feature set in this speaker identification task.

To further test the noise robustness of our proposed feature, we conducted more experiments on noisy speech data with two kinds of real-world noise (car noise and babble noise) as described in Section III-A using the same experimental setup. Fig. 11 and Fig. 12 present the experimental results for the car noise and the babble noise at -6 dB, 0 dB and 6 dB levels, respectively. The proposed auditory features consistently outperform the baseline MFCC system and the MGFCC system under both real-world car noise and babble noise testing conditions.

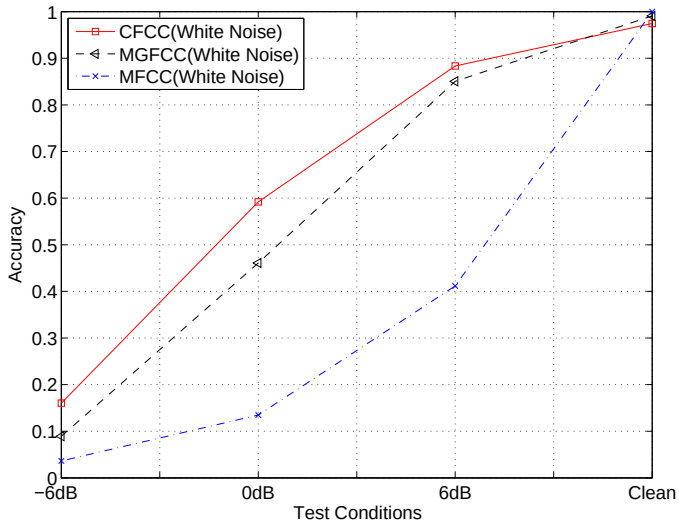

Fig. 10. Comparison of MFCC, MGFCC, and the proposed CFCC features tested on noisy speech with white noise.

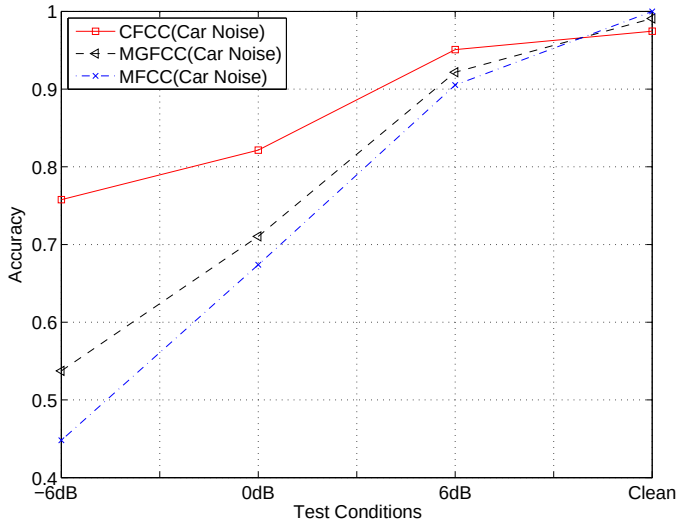

Fig. 11. Comparison of MFCC, MGFCC, and the proposed CFCC features tested on noisy speech with car noise.

We conducted further experiments with PLP and RASTA-PLP features using the same experimental setup as described before [9][47]. The comparative results on white noise, car noise, and babble noise are depicted in Fig. 13, Fig. 14, and Fig. 15, respectively. The CFCC features outperform the PLP features in all three testing conditions. The PLP features minimize the differences between speakers while preserving important speech information via the spectra warping technique [8], which, as a consequence, is typically not used as speech features for speaker recognition. It is interesting to observe that the CFCCs perform significantly better than RASTA-PLP on white noise testing conditions at all different levels; however, for car and babble noise the performance of the CFCCs and RASTA-PLPs is fairly close. It is typically used in combination with PLP, which is referred to as RASTA-PLP [47]. Our experiments show that RASTA filtering largely improves the performance of PLP features in speaker identification under mismatched training and testing conditions. It is particularly helpful when tested under car noise and babble noise, but it is not as effective for white noise. In comparison, the CFCCs consistently generate superior performance in all three conditions.

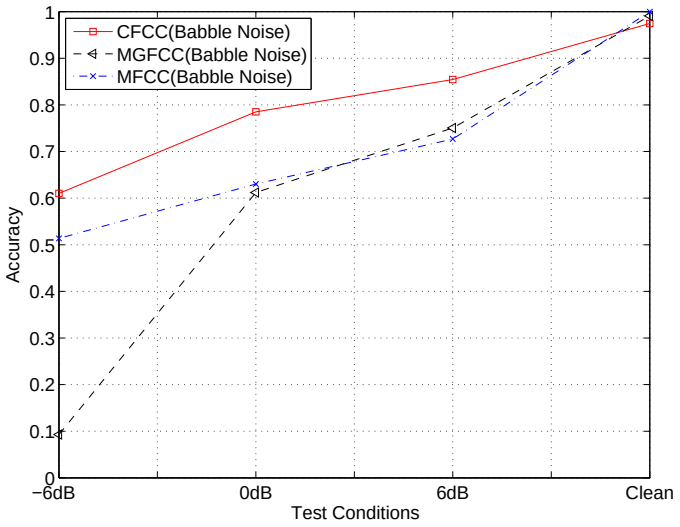

Fig. 12. Comparison of MFCC, MGFCC, and the proposed CFCC features tested on noisy speech with babble noise.

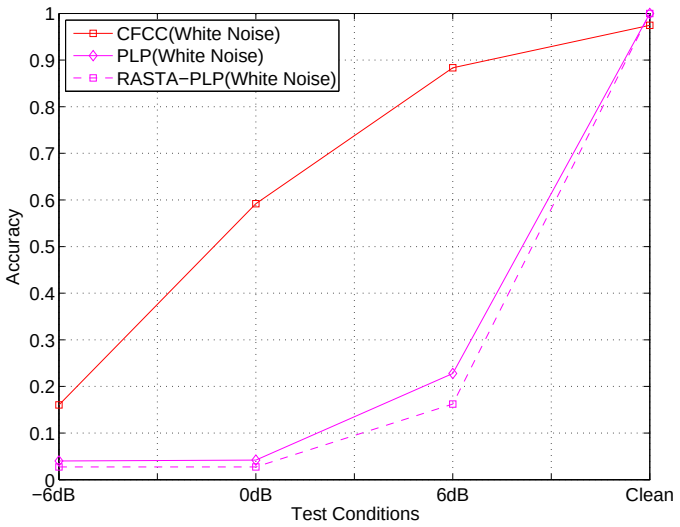

Fig. 13. Comparison of PLP, RASTA-PLP, and the proposed CFCC features tested on noisy speech with white noise.

IV. Conclusions

A new auditory-based feature extraction algorithm for robust speaker identification in mismatched conditions was presented in this paper. Our research was motivated by studies of the signal processing functions in the human peripheral auditory system. The CFCC features are based on a recently presented flexible time-frequency transform (AT) in combination with several components to emulate the human peripheral hearing system. The analytic study for feature optimization was conducted on a separate development set. The optimized

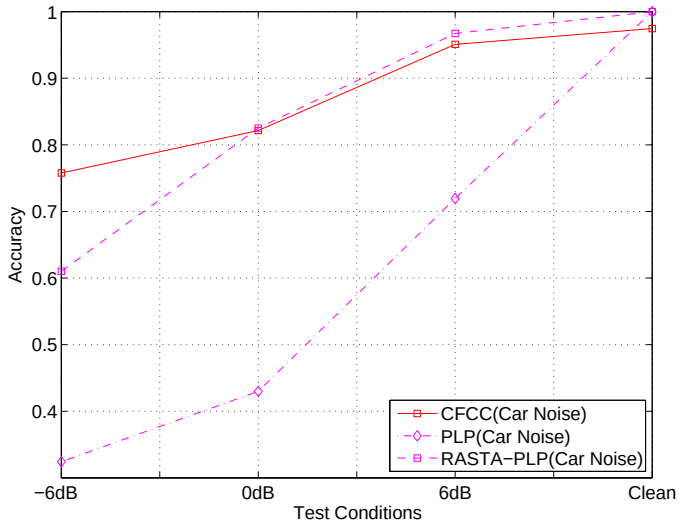

Fig. 14. Comparison of PLP, RASTA-PLP, and the proposed CFCC features tested on noisy speech with car noise.

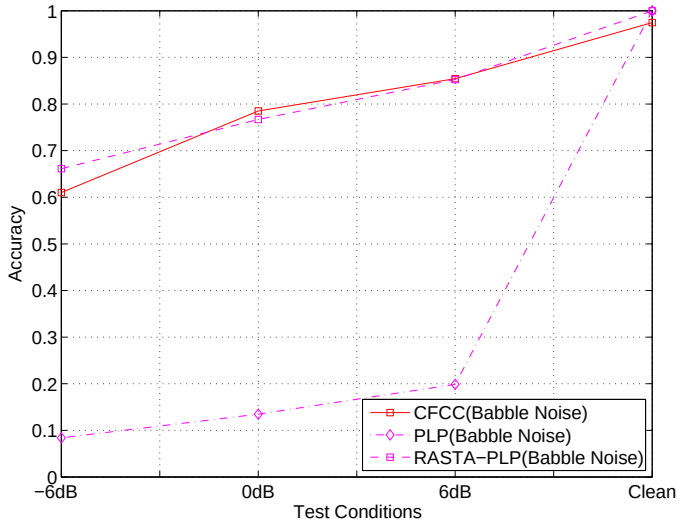

Fig. 15. Comparison of PLP, RASTA-PLP, and the proposed CFCC features tested on noisy speech with babble noise.

CFCC features were then tested under a variety of mismatched testing conditions, which included white noise, car noise, and babble noise. Our experiments show that under mismatched conditions, the new CFCCs perform consistently better than both the MFCC and MGFCC features. Further comparison with PLP and RASTA-PLP features shows that although RASTA-PLP can generate comparable results when tested on car noise or babble noise, it does not perform as well when tested on flatly distributed white noise. In comparison, CFCCs generate superior results under all three noise conditions.

The auditory transform is a new transform for robust feature extraction. In the future, we plan to extend our study of auditory-based features to other speech application tasks, including automatic speech recognition and accent recognition.

ACKNOWLEDGMENT

The authors would like to thank Yan Yin, Craig B. Adams, and Ivan Selesnick for their help and useful discussions. Also, the authors would like to thank the associate editor and anonymous reviewers for their help in improving the quality of this paper.

REFERENCES

[1] Q. Li, "Solution for pervasive speaker recognition," SBIR Phase I Proposal, Submitted to NSF IT.F4, Li Creative Technologies, Inc., NJ, June 2003.
[2] Q. Li, "An auditory-based transform for audio signal processing," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Paltz, NY), Oct. 2009.
[3] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, 1974.
[4] J. Makhoul, "Linear prediction: a tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, April 1975.
[5] Q. Li, F. K. Soong, and O. Siohan, "A high-performance auditory feature for robust speech recognition," in *Proceedings of 6th Int'l Conf. on Spoken Language Processing*, (Beijing), pp. III 51–54, Oct. 2000.
[6] Q. Li, F. K. Soong, and S. Olivier, "An auditory system-based feature for robust speech recognition," in *Proc. 7th European Conf. on Speech Communication and Technology*, (Denmark), pp. 619–622, Sept. 2001.
[7] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, speech, and signal processing*, vol. ASSP-28, pp. 357–366, August 1980.
[8] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.
[9] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech and Audio Proc.*, vol. 2, pp. 578–589, Oct. 1994.
[10] M. grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Trans. on Audio, Speech, and language processing*, vol. 16, pp. 1097–1111, August 2008.
[11] G. von Békésy, *Experiments in hearing*. New York: McGRAW-HILL, 1960.
[12] N. Y.-S. Kiang, *Discharge patterns of signale fibers in the cats auditory nerve*. MA: MIT, 1965.
[13] J. P. Wilson and J. Johnstone, "Capacitive probe measures of basilar membrane vibrations in," *Hearing Theory*, 1972.
[14] E. F. Evans, "Frequency selectivity at high signal levels of single units in cochlear nerve and cochlear nucleus," in *Psychophysics and Physiology of Hearing*, pp. 195–192, 1977. Edited by E. F. Evans, and J. P. Wilson. London UK: Academic Press.
[15] J. P. Wilson and J. Johnstone, "Basilar membrane and middleear vibration in guinea pig measured by capacitive probe," *J. Acoust. Soc. Am.*, vol. 57, no. 3, pp. 705–723, 1975.
[16] A. R. Møller, "Frequency selectivity of single auditory-nerve fibers in response to broadband noise stimuli," *J. Acoust. Soc. Am.*, vol. 62, pp. 135–142, July 1977.
[17] P. M. Sellick, R. Patuzzi, and B. M. Johnstone, "Measurement of basilar membrane motion in the guinea pig using the Mossbauer technique," *J. Acoust. Soc. Am.*, vol. 72, pp. 131–141, July 1982.
[18] R. D. Patterson, "Auditory filter shapes derived with noise stimuli," *J. Acoust. Soc. Am.*, vol. 59, pp. 640–654, 1976.
[19] D. L. Barbour and X. Wang, "Contrast tuning in auditory cortex," *Science*, vol. 299, pp. 1073–1075, Feb. 2003.
[20] S. M. Khanna and D. G. B. Leonard, "Basilar membrane tuning in the cat cochlea," *Science*, vol. 215, pp. 305–306, Jan 182.
[21] B. Moore, R. W. Peters, and B. R. Glasberg, "Auditory filter shapes at low center frequencies," *J. Acoust. Soc. Am*, vol. 88, pp. 132–148, July 1990.
[22] B. Zhou, "Auditory filter shapes at high frequencies," *J. Acoust. Soc. Am*, vol. 98, pp. 1935–1942, October 1995.
[23] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 382–395, Sept. 1995.
[24] J. M. Kates, "A time-domain digital cochlea model," *IEEE Trans. on Signal Processing*, vol. 39, pp. 2573–2592, December 1991.

[25] J. M. Kates, "Accurate tuning curves in cochlea model," *IEEE Trans. on Speech and Audio Processing*, vol. 1, pp. 453–462, Oct. 1993.
[26] B. C. J. Moore and B. R. Glasberg, "Suggested formula for calculating auditory-filter bandwidth and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, 1983.
[27] J. L. Flanagan, *Speech analysis synthesis and perception*. New York: Springer-Verlag, 1972.
[28] G. Zweig, R. Lipes, and J. R. Pierce, "The cochlear compromise," *J. Acoust. Soc. Am.*, vol. 59, pp. 975–982, April 1976.
[29] J. Allen, "Cochlear modeling," *IEEE ASSP Magazine*, pp. 3–29, Jan. 1985.
[30] R. F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Trans. on Acoustics, Speech, and Signal processing*, vol. 36, pp. 1119–1134, July 1988.
[31] W. Liu, A. G. Andreou, and J. M. H. Goldstein, "Voiced-speech representation by an analog silicon model of the auditory periphery," *IEEE Trans. on Neural Networks*, vol. 3, pp. 477–487, May 1992.
[32] J. L. Goldstein, "Modeling rapid waveform compression on the basilar membrane as a multiple-bandpass-nonlinear filtering," *Hearing Res.*, vol. 49, pp. 39–60, 1990.
[33] J. Lin, W.-H. Ki, T. Edwards, and S. Shamma, "Analog VLSI implementations of auditory wavelet transforms using switched-capacitor circuits," *IEEE Trans. on Circuits and systems I: Fundamental Theory and Applications*, vol. 41, pp. 572–583, Sept. 1994.
[34] L. Sellami and R. W. Newcomb, "A digital scattering model of the cochlea," *IEEE Trans. on Circuits and systems I: Fundamental Theory and Applications*, vol. 44, pp. 174–180, Feb. 1997.
[35] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.*, vol. 68, no. 5, pp. 1523–1525, 1980.
[36] B. C. Moore, *An introduction to the psychology of hearing*. NY: Academic Press, 1997.
[37] P. I. M. Johannesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus," *The proceeding of the symposium on hearing Theory*, vol. IPO, pp. 58–69, June 1972.
[38] S. S. Stevens, "On the psychophysical law," *Psychol. Rev.*, vol. 64, pp. 153–181, 1957.
[39] S. S. Stevens, "Perceived level of noise by Mark VII and decibels (E)," *J. Acoustic. Soc. Am.*, vol. 51, pp. 575–601, 1972.
[40] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proceedings of IEEE ICASSP*, pp. 1589–1592, 2008.
[41] D. Wang and G. J. Brown, *Fundamentals of computational auditory scene analysis in Computational Auditory Scene Analysis Edited by D. Wang and G. J. Brown*. NJ: IEEE Press, 2006.
[42] W. M. Fisher, G. R. Doddington, K. M. Goudie-Marshall, C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "Ntimit," Speech Database LDC93S2, LDC, 1993.
[43] http://www.dcs.shef.ac.uk/ martin/SpeechSeparationChallenge/.
[44] D. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.
[45] Y. Suzuki and H. Takeshima, "Equal-loudness-level Contours for Pure Tones," *Journal of Acoustic Society of America*, pp. 918–933, 2004.
[46] I. 266:1987, "Specification for normal equal-loudness level contours for pure tones under free-field listening conditions," Standard BS 3383:1988, ISO 226:1987, BS and ISO, July 1988.
[47] http://www.icsi.berkeley.edu/ dpwe/projects/sprach/sprachcore.html.

Qi (Peter) Li (S'87 - M'88 - SM'01) received a Ph.D. degree in electrical engineering from the University of Rhode Island, Kingston, in 1995.

From 1995 to 2002, he worked at Bell Laboratories, AT&T and then Lucent Technologies, Murray Hill, NJ, as a Member of Technical Staff in the Multimedia Communications Research Lab, where his research focused on speech and speaker recognition, biometric authentication, front-end signal processing, and speech modeling. His research results were implemented in Lucent products and made a contribution to the Bell Labs ASR system, which achieved the top performance in a public robust speech recognition evaluation. Also, he won the best performance in a speaker verification evaluation conducted by one of the largest banks in the U.S. In 2002, he established Li Creative Technologies (LcT), Inc., Florham Park, NJ. LcT is a high-tech company in R&D for speech and image signal processing, multimedia applications, biometrics, and communication products. Dr. Li and his team have been awarded more than thirty research contracts by U.S. government agencies and private companies. He is currently conducting research in hearing, acoustic-signal processing, microphone arrays, speech and speaker recognition, and noise reduction and cancellation for various applications and commercial products. Dr. Li currently holds many issued patents and has filed more than one dozen patents. He has published more than 80 papers in peer-reviewed journals and conferences. He is also the author of the book *Speaker Authentication*, which will be published by Springer in 2011.

Dr. Li has been active as a reviewer for several journals, IEEE publications, and conferences. He is an elected member of the Speech and Language Technical Committee of IEEE Signal Processing Society. He was a Local Chair for the IEEE Workshop on Automatic Identification and a committee member for several IEEE conferences and workshops. He received a best paper award, an achievement award, and several Bell Labs patent awards. He has been listed in Who's Who in America (Millennium and 2001 Editions) and Who's Who in Executives and Professionals (2004 Edition). In 2004, he received the Success Award issued by an agency of the New Jersey Government. He and his team received the Best Consumer Technology/Electronics Company Award issued by the New Jersey Technology Council in 2006 and an Innovations Design and Engineering Award issued by the International CES in 2011.

Yan Huang (M'09) Yan Huang received the M.S.E. degree in Electrical Engineering from Johns Hopkins University, Baltimore, MD, in 2001 and the M.S. degree in Computer Science from University of California, Berkeley, in 2007.

Yan worked at the Center of Language and Speech Processing ('99-'01), Panasonic Speech Technologies Laboratory ('01-'02), and the International Computer Science Institute (ICSI) ('02-'08) as Research Assistant, Research Engineer, and Research Associate, respectively. From 2008 to 2010, she was a Research Scientist with Li Creative Technologies, Inc., located in Florham Park, NJ. She has been working on various components of large vocabulary speech recognition systems, speaker diarization, speaker recognition, and speech synthesis. Her major research interest is machine learning and its applications in speech and natural language processing. Currently, Yan is a Speech Scientist at Microsoft Corporation, located in Redmond, WA, where her major focus has been on unsupervised and lightly supervised acoustic model training, user feedback modeling in voice search. This work was conducted while she was with Li Creative Technologies, Inc.