ROBUST SPEAKER IDENTIFICATION USING AN AUDITORY-BASED FEATURE

Qi (Peter) Li and Yan Huang

Li Creative Technologies (LcT), Inc.
25 B Hanover Road, Suite 140, Florham Park, NJ 07932, USA
{qili, yanhuang}@ieee.org; www.licreativetech.com

ABSTRACT

An auditory-based feature extraction algorithm is presented. The feature is based on a recently published time-frequency transform plus a set of modules to simulate the signal processing functions in the cochlea. The feature is applied to a speaker identification task to address the acoustic mismatch problem between training and testing. Usually, the performances of acoustic models trained in clean speech drop significantly when tested on noisy speech. The proposed feature has shown strong robustness in the mismatched situation. As shown in our experiments, in a speaker identification task, both MFCC and the proposed feature have near perfect performances in a clean testing condition, but when the SNR of input signal drops to 6 dB, the average accuracy of the MFCC feature is only 41.2%, while the proposed feature still achieves an average accuracy of 88.3%.

Index Terms— Speech feature extraction, auditory-based feature, robust speaker recognition, speaker identification, cochlea.

1. INTRODUCTION

Feature extraction is the first crucial component in automatic speech processing. Generally speaking, a successful front-end feature should carry enough discriminative information for classification or recognition, fit well with the back-end modeling, and be robust to the changes of acoustic environments. After decades of research and development, the maintenance of satisfactory system performances under various operating modes remains a major problem, especially when acoustic environments between the training and testing are mismatched. Through a careful study, we have determined that the imitation of the human hearing system is a promising research direction towards improving feature robustness. To this end, we propose an auditory-based feature extraction algorithm based on our recently published auditory-based time-frequency transform [1, 2], which was inspired by the traveling waves in the cochlea.

At a high level, most speech feature extraction falls into the following two categories: modeling the human voice production or modeling the peripheral auditory hearing. For the first approach, one of the most popular features is a group of cepstral coefficients derived from linear prediction, known as linear prediction cepstral coefficient (LPCC) [3]. For the second approach, there are two groups of features, based on either the Fourier transform or the auditory filter bank (or auditory transform). The MFCC (Mel frequency cepstral coefficients) [4] and RASTA-PLP [5] are the two representative speech features that were developed in the first group. The proposed feature belongs to the second group.

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

The Fourier transform (FT) has the 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, when applied in speech processing, the time-frequency decomposition mechanism of the FT is different from the mechanism in the hearing system. First, it uses fixed-length windows, which generate the pitch harmonics in the entire speech bands. Second, its individual frequency bands are in linear distribution, which is different from the nonlinear distribution in human cochlea. Finally, in our recent study [1, 2], we presented that the FFT spectrogram has more noise distortion and more computation noise than an auditory-based transform which we recently developed; thus, it is natural to develop a new feature based on our new auditory-based, time-frequency transform [2], to address the above concerns in the FFT.

In auditory research, the traveling wave of the basilar membrane (BM) in the cochlea and its impulse response have been observed and reported in the literature, e.g. [6, 7, 8]. Moreover, the BM tuning and auditory filters have also been studied in the literature, e.g. [9, 10, 11, 12]. Many electronic and mathematic models have been defined to simulate the traveling wave, the auditory filters, and the frequency responses of the BM, e.g. [13, 14, 15, 16, 17]. Also, there are models to model the entire auditory system, e.g. [18] and references wherein.

The Gammatone filter [19] has been used as a cochlear model to decompose speech signals into the output of a number of frequency bands, but there is no proof to its inverse transform. To provide an invertible auditory-based transform, Li redefined the Gammatone-based filter bank, thus proved the inverse transform [2]. We named it as the *auditory transform* (AT) which includes a pair 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 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. The transform pair has been proven in theory and validated in experiments [2]. Although the invert transform is not necessary in feature extraction, the AT ensures no information loss in the forward transform; therefore, provides a new platform for feature extraction research. In the Gammatone filter bank, the filter bandwidth is locked to the band central frequency, while in the AT, the filter bandwidth can be adjusted easily based on applications.

Compared to the FFT, the new transform has flexible time-frequency resolution and its frequency distribution can take on any linear or nonlinear scales. It is easy to implement a distribution to be similar to that of the Bark, Mel, or ERB scale, which is similar to the frequency distribution of the BM. Most importantly, the proposed

transform has significant advantages in noise robustness and are free from the pitch harmonic distortion as plotted in [2]. Therefore, we use AT as a new platform for feature extraction.

2. PROPOSED AUDITORY-BASE FEATURE

An illustrative block diagram of the proposed feature is shown in Fig. 1. It consists of the following modules to conceptually replicate the hearing system at a high level: a cochlear filter bank, hair-cell function with variable length windows, loudness nonlinearity, and discrete cosine transform (DCT). We name it *cochlear feature cepstral coefficients* (CFCC).

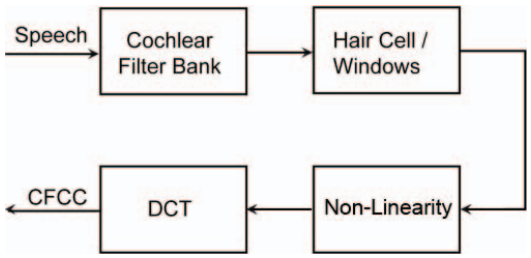

Fig. 1. Diagram of the proposed feature extraction algorithm.

2.1. The Cochlear Filter Bank

The cochlear filter bank is the forward transform of the auditory transform (AT) [2]. Let $f(t)$ be any square integrable function. A transform of $f(t)$ with respect to a function representing the basilar membrane (BM) impulse response $\psi(t)$ is defined as:

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

where a and b are real, both $f(t)$ and $\psi(t)$ belong to $\mathbf{L}^2(R)$, and $T(a,b)$ represents the traveling waves in the BM. The above equation can also be written as:

$$T(a,b) = \int_{-\infty}^{\infty} f(t)\psi_{a,b}(t)\, dt, \tag{2}$$

where

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

Factor a is a scale or dilation variable. By changing a, we can shift the central frequency of an impulse response function. 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 is defined as:

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

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. 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 [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 construct $\psi_{a,b}(t)$ with the lowest frequency along the time axis, the value of a is in $0 < a \leq 1$. If we stretch ψ, the value of a is in $a > 1$. The frequency distribution of the cochlear filter can be in the form of linear or nonlinear scales such as ERB (equivalent rectangular bandwidth) [15], Bark [20], mel scale [4], log, etc. Note that the values of the a need to be pre-calculated for all required central frequency of the cochlear filter.

2.2. Other Operations

The inner hair cells act to transducer mechanical movements into neural activities. When the BM moves up and down, a shearing motion is created between the BM and the tectorial membrane [21]. It causes the displacement of the hairs at the tops of the 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. Our computation of the hair cell function is as follows:

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

where $T(a,b)$ is the filter bank output. Here, we assume that all other detailed functions in the outer ear, middle ear, and the control of auditory 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 in a duration associated with the current band central frequency. At the concept level, we use the following equation:

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

where $d = \lceil 3.5\tau_i, 20\text{ms} \rceil$ is the window length, τ_i is the period of the ith band, and $L = 10$ ms is the window shift duration. We empirically set the computations and the parameters and they may need to be adjusted for different datasets. If we plot S as a 2-D image, it is a kind of spectrogram but much more robust than FFT spectrogram because it has less distortions caused by background noise, less computational noise, and free from pitch harmonics as shown in [2]. Instead of using a fixed length window, we are using a variable length window for different frequency bands. The higher the frequency, the shorter the window. This avoids the high frequency information being smoothed out by long window duration.

Furthermore, we apply the scales of loudness function suggested by Stevens [22, 23] to the hair cell output as:

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

This operation implements cubic root nonlinearity from the physical energy to the perceived loudness. In the last step, the discrete cosine transform (DCT) is applied to decorrelate the feature dimensions and generates the cochlear filter cepstral coefficients (CFCC) as our new auditory-based speech feature.

3. EXPERIMENTS

3.1. Database and Experimental Setup

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) [25]. The training data was recorded under clean conditions. The testing sets were obtained by mixing clean testing utterances with white noises 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 as summarized in Table 1. Note that the training set consists of only clean speech; both the development set and the testing set consist of clean speech and noisy speech at four different SNR levels. Note that we mainly focused on 0 dB and 6 dB conditions in our feature analysis since under -6 dB the performance of all features are closed to random chances.

Table 1. Summary of the Training, Development, and Testing Sets

Data Set	# of Spks.	# of Utters / Spk.	Dur. (sec) / Spk.
Training	34	20	36.8s
Develop.	34	10	18.3s
Testing	34	$10 \sim 20$	29.6s

3.2. The Baseline System

Our baseline system uses the standard MFCC front-end feature and Gaussian Mixture Models (GMMs). Twenty-one-dimension MFCC features (c0 $\sim$ c20) were extracted from the speech audio based on 25 ms window with a frame-rate of 10 ms;the frequency analysis range was set to be 50 Hz $\sim$ 8000 Hz. The 0th component of the MFCC feature corresponding to the energy was discarded and the final front-end feature of the baseline system was 20-dimension MFCCs (c1 $\sim$ c20). 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; therefore it was not used in our baseline system.

The back-end of the baseline system is the standard Gaussian Mixture Models (GMMs) trained using the Maximum Likelihood Estimation (MLE). 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 mixtures were used in the speaker GMM models. To obtain 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.

For comparison, we also implemented the Gammatone filter cepstral coefficient (GFCC) feature following the descriptions in [24], where a *downsampling* procedure was applied to the output of the Gammatone filter bank followed by a cubic root function applied to the absolute values of the downsampling output. An exact implementation following the description in [24] did not give us reasonable experimental results; therefore, we replaced the downsampling procedure in [24] by computing an average of the absolute values on the Gammatone filter bank output using a 20 ms window shifted every 10 ms, followed by a cubic root function and DCT. This procedure is different from the original GFCC, so we name it as *modified GFCC* feature (MGFCC). The MGFCC can be considered as an additional result to support the concept of auditory-based filter bank as an alternative to FFT.

3.3. Compare MFCC, MGFCC, and Proposed CFCC Features

To better understand and optimize the various components of 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 3.1. The goal was to find out the effects of each component to the overall performance and ultimately optimize the feature extraction.

Based on the our analytic study, the details on 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 nonlinearity is applied. Finally, since most back-end systems adopt diagonal Gaussian, discrete cosine transform (DCT) is used to decorrelate the features. The 0th component, corresponding to the energy, is removed from the DCT output.

Table 2 summarizes the speaker identification accuracy of the optimized CFCC feature in comparison with MGFCC and MFCC tested on the development set.

Table 2. 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%

Using the optimized CFCC feature extraction based on the development set, we conduct speaker identification experiments on the testing set with the results depicted in Fig. 2. As we can see from Fig. 2, in clean testing condition, the CFCC feature generates comparable near-perfect results to MFCC. As white noises with increasing intensities are added to the clean testing data, the performances of the CFCC are significantly better than the MGFCC and MFCCs performances. 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 feature still achieves 88.3% accuracy, which is more than two times better

than the MFCC feature. The MGFCC feature has an accuracy of 85.1%, which is better than the MFCC feature, but not as good as the proposed CFCC feature. The CFCC performance in the testing data set is similar to its performance in the development set. Overall, we see that the proposed CFCC feature outperforms both the widely used MFCC feature and another related auditory-based MGFCC feature in this speech identification task.

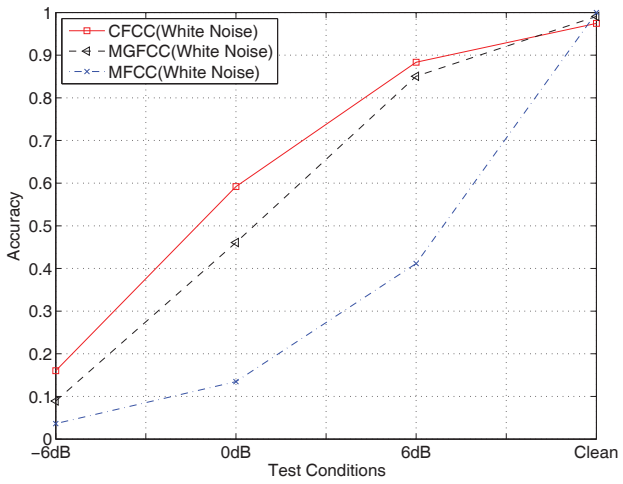

Fig. 2. Comparison of MFCC, MGFCC, and the proposed CFCC features tested on noisy speech with white noise.

4. CONCLUSIONS

A new auditory-based feature for robust speaker identification was presented in this paper. The research was motivated by the studies of the signal-processing functions in the human peripheral auditory system. The feature was developed based on a recently presented invertible time-frequency transform plus several components motivated by the human hearing system. Our experiments suggest that under mismatched acoustic conditions, the new feature consistently performs better than both the MFCC and MGFCC features.

The auditory-based transform provides a new platform for robust feature research. In the future, we plan to extend our study of the CFCC feature to other speech application tasks, including automatic speech recognition, accent recognition, and other applications.

5. 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. Atal and M. Schroeder, "Predictive coding of speech signals," in *Proceedings of the 6th Int. Congress on Acoustics*, (Tokyo), 1968.

[4] 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.

[5] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech and Audio Proc.*, vol. 2, pp. 578–589, Oct. 1994.

[6] G. von Békésy, *Experiments in hearing*. New York: McGRAW-HILL, 1960.

[7] N. Y.-S. Kiang, *Discharge patterns of signale fibers in the cats auditory nerve*. MA: MIT, 1965.

[8] J. P. Wilson and J. Johnstone, "Capacitive probe measures of basilar membrane vibrations in," *Hearing Theory*, 1972.

[9] R. D. Patterson, "Auditory filter shapes derived with noise stimuli," *J. Acoust. Soc. Am.*, vol. 59, pp. 640–654, 1976.

[10] D. L. Barbour and X. Wang, "Contrast tuning in auditory cortex," *Science*, vol. 299, pp. 1073–1075, Feb. 2003.

[11] 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.

[12] B. Zhou, "Auditory filter shapes at high frequencies," *J. Acoust. Soc. Am*, vol. 98, pp. 1935–1942, October 1995.

[13] J. M. Kates, "A time-domain digital cochlea model," *IEEE Trans. on Signal Processing*, vol. 39, pp. 2573–2592, December 1991.

[14] J. M. Kates, "Accurate tuning curves in cochlea model," *IEEE Trans. on Speech and Audio Processing*, vol. 1, pp. 453–462, Oct. 1993.

[15] 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.

[16] J. L. Flanagan, *Speech analysis synthesis and perception*. New York: Springer-Verlag, 1972.

[17] G. Zweig, R. Lipes, and J. R. Pierce, "The cochlear compromise," *J. Acoust. Soc. Am.*, vol. 59, pp. 975–982, April 1976.

[18] M. Zilany and I. Bruce, "Representation of the vowel /ɛ/ in normal and impaired auditory nerve fibers: Model predictions of responses in cats," *J. Acoust. Soc. Am*, vol. 122, pp. 402–417, July 2007.

[19] 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.

[20] 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.

[21] B. C. Moore, *An introduction to the psychology of hearing*. NY: Academic Press, 1997.

[22] S. S. Stevens, "On the psychophysical law," *Psychol. Rev.*, vol. 64, pp. 153–181, 1957.

[23] S. S. Stevens, "Perceived level of noise by Mark VII and decibels (E)," *J. Acoustic. Soc. Am.*, vol. 51, pp. 575–601, 1972.

[24] 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.

[25] http://www.dcs.shef.ac.uk/∼martin/SpeechSeparationChallenge/.