FEATURE EXTRACTION BASED ON HEARING SYSTEM SIGNAL PROCESSING FOR ROBUST LARGE VOCABULARY SPEECH RECOGNITION

Qi (Peter) Li and Xie Sun

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

ABSTRACT

A new auditory-based feature extraction algorithm for robust speech recognition is developed from modeling the signal processing functions in the hearing system. Usually, the performance of acoustic models trained in clean speech drops significantly when tested on noisy speech; thus recognition systems cannot work robustly in the field even when they have good performance in labs. To address the problem, we have developed features based on a set of modules to simulate the signal processing functions in the cochlea, such as auditory transform, hair cells, and equal-loudness functions. The features are then applied to the Wall Street Journal task. To simulate the performance in the field, the training data is near clean speech while the testing data are with added white and babble noise. As shown in our experiments, without added noise, the proposed features have a similar performance as MFCC, RASTA-PLP, and PLP features. When we added noise and tested at different SNR levels, the performance of the proposed auditory features is significantly better than others. For example, at 10 dB SNR level which is often encountered in real applications, the performance of the proposed auditory features is 65.53% while the best from others is 36.33% from the RASTA-PLP. The proposed features provide an absolute gain on recognition accuracy of 29.20%. Overall, our experiments show that the proposed auditory features have strong robustness in the mismatched and noisy situations in speech recognition.

Index Terms— Speech feature extraction, auditory-based feature, robust speech recognition, cochlea, auditory transform.

1. INTRODUCTION

Current automatic speech recognition (ASR) technology can provide good performance in clean or very low noise environments, such as in a quiet office, or when acoustic conditions in training can match the conditions in testing. ASR performance may drop significantly or a system may stop working when encountering background noise or when the mismatch appears. There are many ways to address the problem and a lot of work has been done using different approaches. Since the human hearing system is robust, our approach was as follows: first, model the hearing system, and second, develop feature extraction algorithms based on our hearing models.

Our research started with a study of the time frequency transform in the cochlea and then extended to other functions in the hearing system. The proposed auditory features are an outcome of this research.

The fast Fourier transform (FFT) is the time frequency transform used in popular MFCC [1], PLP [2] and RASTA-PLP [3] features. The FFT 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, 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 pitch harmonics in the entire speech bands. Second, its individual frequency bands have linear distribution, which is different from the nonlinear distribution in human cochlea. Finally, in our recent study we demonstrated that the FFT spectrogram has more noise distortion and more computation noise than our auditory-based transform; thus, it is natural to develop new features based on our new auditory-based, time-frequency transform [4], 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. [5]. Moreover, the BM tuning and auditory filters have also been studied in the literature, e.g. [6]. 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. [7]. Also, there are models to model the entire auditory system, e.g. [8] and references therein. The Gammatone filter [9] 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 proving the inverse transform [4]. The new *auditory transform* (AT) includes a pair of a forward transform and an inverse transform.

Compared to the FFT, the AT 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 AT has significant advantages in noise robustness and is free from the pitch harmonic distortion as plotted in [4]. Compared to 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 by changing a parameter. As we have observed in our experiment, adjusting the bandwidth can improve ASR performance. Given the above analysis, we use AT as the first module when modeling the cochlear system and add others to simulate the major signal processing functions in the hearing system.

In this paper, we present new auditory features based on our recent hearing research in a more complete signal processing platform to model the hearing system. These lead to the features with better performance than the previously published one [10]; therefore, we named the new one as *new auditory features* as presented below.

2. HEARING MODELS AND FEATURE EXTRACTION

An illustrative block diagram of the proposed hearing model is shown in Fig. 1. It consists of the following modules: the auditory transform (AT), energy normalization, Meddis hair cell model, equal-loudness function, windowing, loudness nonlinearity, and discrete cosine transform (DCT). Except for DCT, all the modules are for modeling the signal processing functions in the hearing system.

To facilitate the following discussions, we name the new features from the model as *auditory feature cepstral coefficients* (AFCC). Compared to the auditory features in [10], we added four modules so this frontend is also a model for the hearing system and not just a feature extraction algorithm.

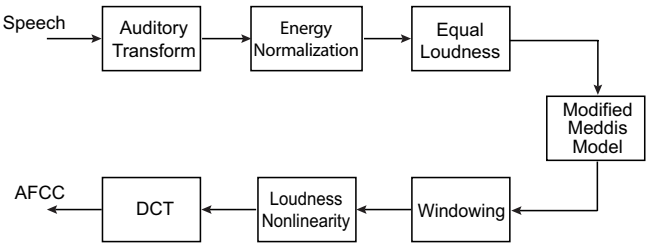

Fig. 1. Diagram of the proposed auditory-based feature extraction algorithm for ASR.

The auditory filter bank in the AT simulates the frequency response of the BM in the cochlea [4]. 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.

The auditory filter in the AT is defined as:

$$\begin{aligned}\psi_{a,b}(t) &= \frac{1}{\sqrt{|a|}} \left(\frac{b-t}{a}\right)^{\alpha} \exp\left[-2\pi f_L \beta \left(\frac{b-t}{a}\right)\right] \\ &\quad \cos\left[2\pi f_L \left(\frac{b-t}{a}\right) + \theta\right] u(-t),\end{aligned} \tag{4}$$

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. $\theta = 0$.

The value of a can be determined by the current filter central frequency, f_c, and the lowest central frequency, f_L, in the auditory filter bank:

$$a = f_L / f_c. \tag{5}$$

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) [13], Bark [14], mel scale [1], log, etc. Note that the values of the a need to be pre-calculated for all required central frequency of the cochlear filter.

As shown in [4], the spectrograms generated from the AT are free from harmonics, have much less computation noise, and are robust to background noise compared to the spectrograms generated from the FFT. In numerical computation, the AT output can be represented as $T(i,n)$, where i represents the number of the frequency band and t represents discrete time. Since we are not using energy in the following computation, the gain of the auditory filters in the AT may need to be renormalized.

Following the AT, an equal-loudness function [15], $g(i)$, is applied to each band of the AT output:

$$E(i,n) = g(i)T(i,n) \;\; \forall i,n \tag{6}$$

where $g(.)$ is actually a weighting function on the different frequency bands.

In the hearing system, the inner hair cells act to transduce mechanical movements into neural activities. When the BM moves up and down, a shearing motion is created between the BM and the tectorial membrane [16]. This 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. We applied the Meddis hair cell model [17] to our computation which includes a feedback loop. For our applications, we applied the following constrains to ensure that the model output is not negative.

$$M(i,n) = \begin{cases} m[E(i,n)] & \text{if } E(i,n) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where the Meddis model is represented as $m(.)$.

In the next step, the hair cell output for each band is converted into a representation of nerve spike count density. To simplify, we use a shifting window to represent the function. The window function with length ℓ can be represented as:

$$S(i,j) = \frac{1}{\ell} \sum_{n=n_j}^{n_j+\ell-1} M(i,n) \tag{8}$$

The window length can be in 20 - 30 ms and shit 10 ms at each step.

Furthermore, we apply the scales of loudness function suggested by Stevens [18, 19] to the hair cell output as:

$$Y(i,j) = S(i,j)^{1/3}. \tag{9}$$

This operation implements cubic root nonlinearity to model the perceived loudness. In the last step, DCT is applied to decorrelate the feature dimensions and generates the auditory filter cepstral coefficients (AFCC) as our new auditory-based speech features. We note that in software implementation, the order of the above computation can be changed for fast and efficient computation.

3. EXPERIMENTS

To investigate the performance of the proposed auditory features, we used large vocabulary recognition corpus Wall Street Journal (WSJ0)

as the original speech data. To simulate a distant talking scenario using a handheld device or hands-free application, the original data was played from an artificial mouth and recorded using a microphone array from a distance of 0.5 meter in a standard office room. The microphone array named CrispMic^{TM} is a small, linear array with four microphone components [20, 21]. The re-recorded WSJ corpus was then partitioned to training, development, and testing data sets without any overlap. The utterances defined in WSJ-SI84 [22] were used as our training set in which 7,138 utterances from 3,586 males and 3,552 females are included. The test set was built based on the standard Nov 92 test set [22] with 330 test utterances. A development set from the Nov 92 speaker-dependent test set was used to tune parameters, in which 310 utterances are included. The dictionary includes 5,000 word vocabularies. Cross-word, tri-phone acoustic models were trained, and a bi-gram language model was used in all the experiments. To evaluate the performance in noisy and mismatched conditions, we added white and babble noise to the pre-recorded data with a SNR of 5dB, 10dB, 15dB, and 20dB, respectively. We named the original WSJ corpus as the original dataset, the microphone array recorded datasets as the distant datasets, and the white and babble noise added datasets as the noisy datasets in the following discussions. The original dataset was not used in our experiments.

The sampling frequency of the distant dataset is 16 kHz. For the MFCCs, RASTA-PLP, and PLP thirteen dimensional cepstral coefficients and their first and second order time derivatives were used in the acoustic model training and testing. For the proposed AFCCs, we investigated the 9 to 13 dimensions of the cepstral coefficients in the development set using the distant dataset. The best performance was obtained at 10 dimensional AFCC which is energy plus 9 dimensional cepstral coefficients while other dimensions also provide similar performances. Thus, we use 30 dimensional AFCC in total in our experiments including base, first and second order time derivatives, which are 9 dimensions fewer than the MFCC features. We note that for a fair comparison, we used the popular HTK toolkit to generate the MFCC and PLP features with cepstral mean normalization (CMN). For RASTA-PLP, we use the exactly the code downloaded from [23] with the same experimental setup as described in [3].

CMN was used for all 4 features. We did not use the Wiener filter or spectral subtraction based noise reduction algorithms in any one of the features for fair comparison. The same speech signals were inputted to each one of the feature extraction software packages directly.

Regarding acoustic models, tri-phoneme models were used and each mode has three hidden Markov model (HMM) states. The model structures are the same for both MFCC and AFCC features although AFCC feature vector has fewer dimensions. All the acoustic models were trained only using the distant dataset and tested in both distant and noisy datasets.

Regarding language models, we did not tune the language model parameters, such as word insertion penalty and grammar scale factor. The language models were the same for all experiments.

Based on the discussion in the last section, the details on the AFCC feature extraction can be summarized as follows: First, the speech waveform is passed through the auditory filter bank which is the forward transform of the AT. The filter width parameter β is set to 0.15. The Bark scale is used for the filter bank distribution. Energy normalization can be applied to ensure the energy representation of each channel of the filter output matches and equal-loudness function can be applied at this stage. Following that, the modified Meddis model is applied to further process the waveform. The output is a rectified and processed waveform. A moving window is then applied to each channel. The window length is 25 ms and shifts every 10 ms. The window output is the average value of the waveform in the window. The window output then goes through the loudness nonlinearity. Finally, since most back-end systems adopt diagonal Gaussian, the DCT is used to decorrelate the features. For AFCCs, we use the energy term c0 plus nine coefficients c1 to c9 as the cepstral coefficients.

Table 1. Comparison on Test Set with Added White Noise in Word Accuracy (%)

Testing SNR	5 dB	10 dB	15 dB	20dB	Clean
AFCC (Proposed)	36.02	65.53	78.95	85.34	89.20
MFCC	10.61	31.76	64.99	82.22	90.88
RASTA-PLP	12.89	36.33	66.02	80.40	87.46
PLP	12.91	35.10	66.11	82.70	91.01

Table 2. Comparison on Test Set with Added Babble Noise in Word Accuracy (%)

Testing SNR	5 dB	10 dB	15 dB	20dB	Clean
AFCC (Proposed)	35.92	73.06	84.2	87.99	89.20
MFCC	28.1	63.68	81.73	88.59	90.88
RASTA-PLP	21.17	52.90	76.46	84.91	87.46
PLP	26.90	61.85	81.90	88.29	91.01

In our evaluation, in the first step, the AFCC feature parameters were adjusted in the development dataset without added noise. In the second step, all features were generated using the noisy datasets. As noises with increasing intensities were added to the distant testing dataset, the performance of the AFCC improved significantly than other feature performances. Tables 1 and 2 summarize the comparison on different features on the speaker independent ASR task at different SNR levels. The overall performance is shown in Figs. 2 and 3. Overall, we see that the proposed AFCC features outperform other features in noisy speech recognition.

4. CONCLUSIONS

New auditory-based features for robust large-vocabulary speech recognition were proposed in this paper. The new features are constructed by modeling the signal-processing functions in the human hearing system. Our experiments suggest that under noisy and mismatched acoustic conditions, the new features consistently perform better than the MFCC, RASTA-PLP, and PLP features. Our models of the hearing system should make a significant contribution to speech recognition robustness.

5. ACKNOWLEDGMENT

This material is based upon work supported by DARPA under Contract No. W31P4Q-10-C-0016. The authors would like to thank the program managers, Dr. Joseph Olive and Dr. Bonnie Dorr, for their support and Yan Yin and Qiru Zhou for their help in the experiments. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Approved for Public Release: Distribution Unlimited.

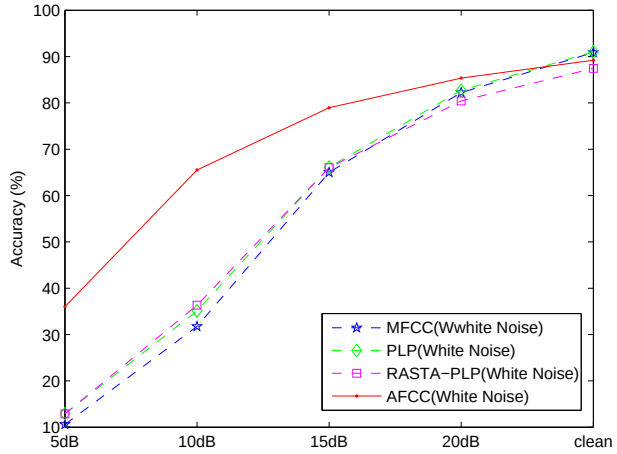

Fig. 2. Word accuracies of features tested on speech with white noise.

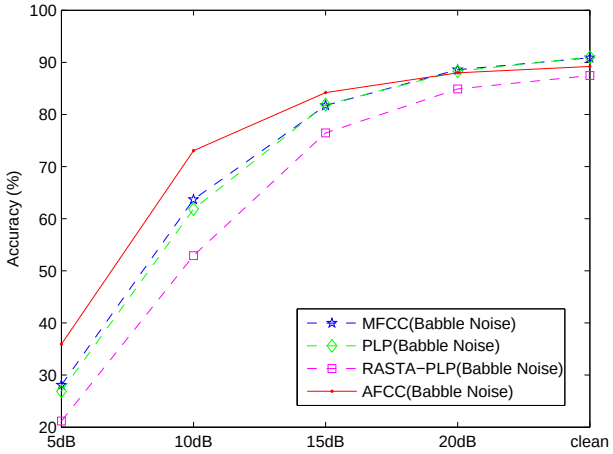

Fig. 3. Word accuracy of features tested on speech with babble noise.

6. REFERENCES

[1] 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.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

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

[4] 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.

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

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

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

[8] 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.

[9] 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.

[10] Q. Li and Y. Huang, "Robust speaker identification using an auditory-based feature," in *ICASSP 2010*, 2010.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] ISO, "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.

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

[17] R. Meddis, "Simulation of mechanical to neural transduction in the auditory recepter," *JASA*, vol. 79, pp. 702–711, March 1986.

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

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

[20] Q. Li, M. Zhu, and W. Li, "A portable usb-based microphone array device for robust speech recognition," in *ICASSP 2009*, 2009.

[21] http://www.crispmic.com.

[22] P. C. Woodland, J. J. Odll, V. Valtech, and S. J. Young, "Large vocabulary continuous speech recognition using htk," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. II, (Adelaid, Australia), pp. 125–128, 1994.

[23] http://www.icsi.berkeley.edu/ dpwe/projects/sprach/sprachcore.html.