AN AUDITORY SYSTEM-BASED FEATURE FOR ROBUST SPEECH RECOGNITION

Qi Li, Frank K. Soong, and Olivier Siohan

Multimedia Communications Research Laboratory
Bell Labs, Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974, USA
{qli,fks,siohan}@research.bell-labs.com

ABSTRACT

An auditory feature extraction algorithm for robust speech recognition in adverse acoustic environments is presented. The feature computation is comprised of an outer-middle-ear transfer function, FFT, frequency conversion from linear to the Bark scale, auditory filtering, nonlinearity, and discrete cosine transform. The feature is evaluated in two tasks: connected-digit recognition and large vocabulary continuous speech recognition. The tested data were under various noise conditions, including handset and hands-free speech data in landline and wireless communications with additive car and babble noise. Compared with the LPCC, MFCC, MEL-LPCC, and PLP features, the proposed feature has an average 20% to 30% string error rate reduction on the connected-digit task, and 8% to 14% word error rate reduction on the Wall Street Journal task in various additive noise conditions.

1. INTRODUCTION

Feature extraction is the first crucial block in any automatic speech recognition (ASR) system. Currently, there are two major approaches to feature extraction: modeling human voice production and perception system. For the first approach, one of the most popular features is the LPCC feature. For the second approach, the most popular feature is the MFCC feature [1]. Both features work well in clean but not so in adverse environments.

After comparing the above approaches, we decide to pursue the auditory system-based approach in our attempt to develop a noise robust front-end since human perception seems to be rather resilient to various types of noise. This approach may have the potential to address more directly the robust problem in ASR. Auditory system study involves many research areas, such as acoustics, physiology, psychology, signal processing etc. We first investigate the results from the hearing research to characterize the functions of auditory system; then model the functions by signal processing operations; finally, implement the functions in equivalent and efficient algorithms. We have reported some preliminary results in [2]. In this paper, we further improve the extraction algorithm, conduct experiments in adverse environments, and reduce the computational complexity.

2. PROPOSED AUDITORY FEATURE

The human auditory system consists of the following modules: outer ear, middle ear, cochlea, hair cells, and nerve system. It converts the sound pressure to auditory nerve firing rates in various frequency bands for auditory cognition in the brain. The overall transfer function (TF) from concha to oval window is a cascaded response of the outer- and the middle-ear transfer functions. As we showed in [2], from 0 to 8 KHz, the TF exhibits a bandpass characteristics while the 0 to 4 KHz a high-pass nature. The outer-middle-ear transfer function outperforms the preemphasis filter for the new feature. Therefore, we use the transfer function to replace the preemphasis filter in the front-end.

For telephone speech, the signal is first sampled at a sampling rate of 8 KHz, and then blocked into 30 ms, or 240 samples. The speech samples are then weighted by a Hamming window for FFT. The window is shifted every 10 ms or 80 samples. After the TF, $T(\omega)$, the power spectrum $P(\omega)$ becomes: $P_T(\omega) = P(\omega) \cdot T(\omega)$. Although the frequency scale of FFT spectrum is linear, the frequency along the basilar membrane is on the Bark [3], ERB-rate, or mel [1] scale. The above three scales are similar and they are from the measurement of human auditory system. We chose the Bark scale because it has showed a slight advantage over the other scales in our recognition experiments.

To convert the frequency from linear to Bark, we first have the same number of points equally spaced in the Bark scale, and then project those points onto the linear axis by:

$$f = \begin{cases} \frac{1000}{0.76}\tan(z/13) & \text{if } 0 \text{ Bark} \le z \le 10.41 \text{ Bark} ; \\ 1000 \cdot 10^{\frac{z-8.7}{14.2}} & \text{if } 10.41 \text{ Bark} < z \le 17.25 \text{ Bark} . \end{cases} \quad (1)$$

The projected value is obtained by linearly interpolating neighboring points on both sides.

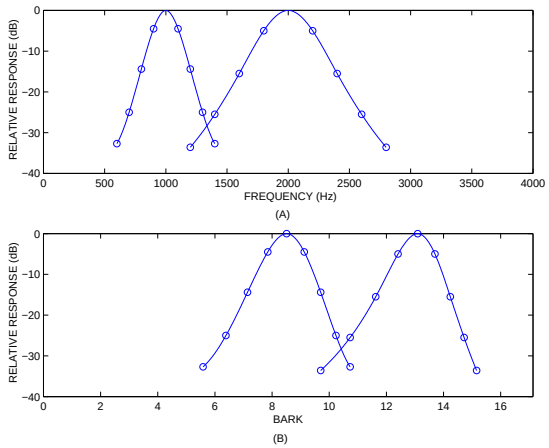

Figure 1: (A) The shapes of cochlear filters, determined using the Patterson's method, centered at 1 KHz and 2 KHz. (B) The shapes of the filters plotted at the Bark scale.

The selective frequency response of the basilar membrane acts as a bank of bandpass filters equally spaced in the Bark scale. The shape of the auditory filter has been studied, and one well-accepted result was from the Patterson's study [4]. It assumed that the shape of the cochlear filter[1] is roughly symmetric. Two cochlear filters, determined by the Patterson's method, centered at 1 KHz and 2 KHz, are plotted in Fig. 1 (A) in Hz and (B) in the Bark scale (adapted from Fig. 3 in [4]). From Fig. 1, we observe the following: first, the shapes of the filters are neither triangular used in MFCC [1] nor the shape used in PLP [5]; second, the widths of the auditory filters are approximately the same along the Bark scale. The cochlear filter performs two functions: first, it works as a bandpass filter to select signal components near the center frequency; second, the cochlear filter smoothes out component noise within its bandwidth.

To simulate the cochlear filters, we first define a digital, prototype auditory filter with a shape similar to that of the Patterson's filter, and then repeat the same prototype at equal distance along the Bark scale. This is equivalent to performing a moving-average operation along the Bark scale. For the nth digital filter centered at the zth point in the Bark scale, the filter output is:

$$P_z(n) = \sum_{k=-K}^{K} P_B(z+k)H(k), \qquad (2)$$

where the size of the filter is $2K+1$, and $H(\cdot)$ is the function of a digital filter. In practice, we make the shape of the filter wide enough to smooth out noise and pitch harmonics as much as possible while still maintaining a good frequency resolution. One of the designed filter for 8 KHz sampling rate and 128-point FFT spectrum is shown in Fig. 2 (left). To reduce the computational complexity, we use 32 filters spaced apart every 4 points, starting from the 4th point of the 128-point FFT power spectrum.

[1]To distinguish the real auditory filter from the digital auditory filter, we refer the real one as the cochlear filter.

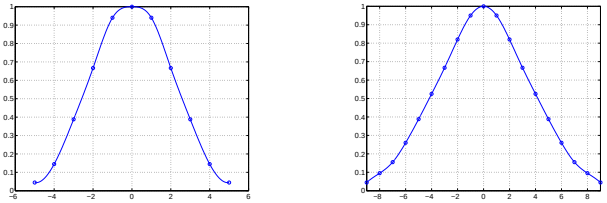

Figure 2: Shapes of the auditory filters for 8 KHz (left) and 16 KHz (right) sampling rates, respectively.

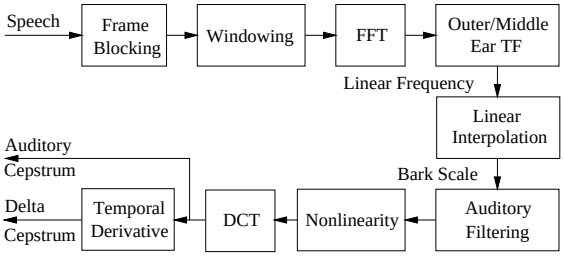

Figure 3: Schematic diagram of the proposed feature.

Auditory research has shown that the basilar membrane response is nonlinear. We choose logarithm as the nonlinearity since it performs better than cubic-root in our recognition experiment. The nonlinearity is followed by a discrete cosine transform (DCT) to convert the logarithmic spectrum to 12 cepstral coefficients. With only finite terms, DCT smoothes out the pitch harmonics in the spectrum. Short-term energy is selected as the energy term in the feature vector. It is computed by accumulating the power of the blocked speech samples before the Hamming window. The procedure for extracting new feature is summarized in Fig. 3.

For various recognition tasks, one set of fixed feature parameters may not always yield the best performance. Fortunately, the proposed feature provides the flexibility for adjusting the filter parameters. Since the proposed auditory filter, $H(x)$, is symmetric, its coefficients can be represented as:

$$H(x) = \{\{F(k)\}_{-W}^{0}, \{F(k)\}_{1}^{W}\}, \qquad (3)$$

where the length of the filter is $2W+1$ and $F(k) = F(-k)$. Since we keep $F(0) = 1$, only $\{F(k)\}_{-W}^{1}$ need to be adjusted with the following constraint:

$$0 < F(k) < F(k+1) < 1, \quad -W \leq k < 0. \qquad (4)$$

In practice, only the filter weights near the center of the filter, such as $F(1)$ and $F(2)$, need to be adjusted during the HMM training to yield the best performance.

Table 1: Comparisons on String Error Rates (SER) and SER Reductions on the task of focusing on hands-free data (digit error rates are in parentheses)

Feature	String Error Rates (%)		SER Reduc. (%) Proposed vs. other	
Databases	Handset	Lapel	Handset	Lapel
Proposed	5.9 (0.8)	11.0 (1.5)	0.0	0.0
LPCC	9.4 (1.0)	13.9 (2.0)	37.2	20.9
MFCC	7.0 (0.9)	14.9 (1.9)	15.7	26.2
Mel-LPC	8.6 (1.0)	12.8 (1.8)	31.4	14.1
PLP	6.6 (0.9)	13.9 (2.0)	10.6	20.9

3. EXPERIMENTAL RESULTS

All experiments used a 39-dimensional feature vector, including short-term energy, 12 cepstral coefficients, plus their first and second order time derivatives. The long-term cepstral mean for each utterance was calculated and removed. There is no overlap between any training and testing utterances for any one of the tasks.

3.1. Connected-Digit Recognition

In the first experiment, we used two CDMA wireless databases named “Handset” and “Lapel”, respectively, collected in a moving car. The Handset database was recorded with handsets while Lapel was recorded with microphones attached to speakers’ lapels. Handset has 769 and 256 utterances for training and test while Lapel has 2,026 and 517 utterances for training and test, respectively. This experiment is intended to show the robustness of the proposed feature in a noisy, moving car environment. The models are context-dependent (CD) digit HMMs (10 states per digit and 7 mixture components on average per state). State tying in the models was determined by a decision-tree based training algorithm [6]. There are totally 1,400 CD models with 800 tied states. The training algorithm is based on maximum likelihood estimation (MLE). Four other feature sets, LPCC, MFCC, MEL-LPCC, and PLP[2], were compared under the exactly same training and testing conditions. Each test utterance consists of a 10-digit string. The testing results are listed in Table 1 in string error rate (SER) and SER reductions. The proposed feature outperforms other features with 11% to 37% SER reductions. In terms of computational complexity, the proposed feature used 1% of real time compared to 0.7% - 0.8% on others. Compared to the decoding complexity, this difference is negligible.

In the second experiment, we evaluate the proposed new feature with different HMM model structures. We trained a set of HBT digit models [7] using 21 databases with a total of 44,123 utterances. The HBT model is context-dependent across the “head" (first 3 states) and “tail" (last 3 states) while the “body" (4 states) is context independent. We first trained initial models by MLE, then applied the generalized probabilistic descent (GPD) algorithm [8] to sharpen the models discriminatively in two iterations using all 21 databases. The GPD-trained models were tested on 4 databases with 10-digit strings. Same training and testing were applied to both LPCC and the proposed features. As shown in Table 2, compared to the LPCC feature, the proposed one has an average 18.1% string error rate reduction on the 4 testing databases, including 2 landline and 2 CDMA databases with 518, 1685, 256, and 517 utterances, respectively.

[2]The Bark-scale PLP feature was implemented based on the HTK package. When the order of autoregressive model is 12, the PLP feature has the best performance as in Table 1 after a careful search from 6 to 14.

Table 2: String Error Rates (SER) and SER Reduction on HBT Models Using Discriminant Training (%)

Data	Tele	Sdn10	Handset	Lapel	Ave.
LPCC	3.9	7.5	6.2	16.2	8.5
Proposed	3.3	8.3	4.7	11.4	6.9
Reduction	15.4	-10.7	24.2	29.6	18.1

Finally, to evaluate the proposed feature at low signal-to-noise ratio (SNR), we tested the proposed feature using the data with additive noise. The models are the same as the above GPD-trained, HBT models. The original testing database is the CDMA “Handset" database. Two kinds of noises, recorded in a moving car and on the trading floor of New York Stock Exchange (NYSE), were digitally added to the original Handset database to generate new, noisy testing data. In total, there are 4 testing datasets for each kind of the noise with different segmental SNRs: 10 dB, 15 dB, and 20 dB, plus the original dataset. The segmental SNR of the original set is about 25 dB.

The results on additive car noise are shown at Fig. 4. The SERs for the auditory feature and LPCC features are

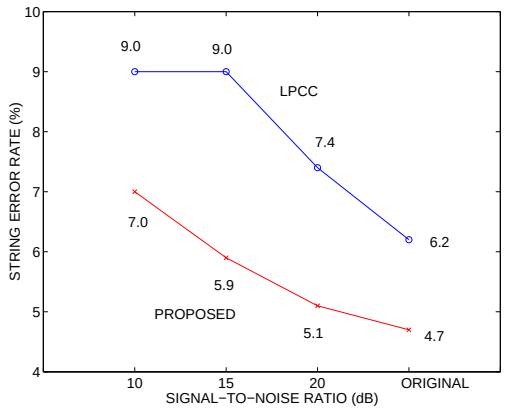

Figure 4: Comparisons on additive car noise.

plotted in two curves, marked with crosses and circles, respectively. The corresponding SERs are labeled. For the datasets, with 10 dB, 15 dB, and 20 dB SNRs, and the original one, the auditory feature provides SER reductions of 22.2%, 34.4%, 31.1%, and 24.2%, respectively, compared to the LPCC features. For the datasets with 10 dB, 15 dB, and 20 dB SNRs on additive NYSE noise, the auditory feature provides SER reductions of 33.1%, 36.0%, and 29.5%, respectively, compared to the LPCC features. To ensure a fair comparison, the search engine was controlled so that the search space was pretty much the same for both features in terms of total number of active hypothesis arcs in decoding.

3.2. Wall Street Journal (WSJ) Recognition Task

The sampling rate for the WSJ database is 16 KHz. The speech signal is blocked into 20 ms window, shifted every 10 ms. After Hamming windowing and zero padding, a 512-point FFT is performed to generate 256-point spectra. The shape of the auditory filter for the 256-point spectrum is plotted at Fig. 2 (right). A total of 32 filters are used. The first filter is centered at the 6th point at the low frequency side. Other filters are spaced apart from each other by 8 points.

For this task, the SI-84 training corpus was used. It contains 7,200 utterances. The language models used in the experiments were the standard trigram language models provided in the WSJ corpus. The model structure is context-dependent, state-tying triphone models. They are trained using MLE based, decision-tree, state-tying algorithm [6]. The original testing data are WSJ (Nov, 1992)-5K closed vocabulary dataset with 330 utterances. To test the robustness, car noise was added to the original data at 10 dB, 15 dB, and 20 dB segmental SNRs.

The experimental results are shown in Fig. 5. The auditory feature yields the same WER as the MFCC feature for clean speech, but it yields a better performance in noise. From 10, 15, to 20 dB SNR, the WER reductions are 7.6%, 14.3%, and 7.6% compared to the MFCC feature. Both features used similar amount of decoding time.

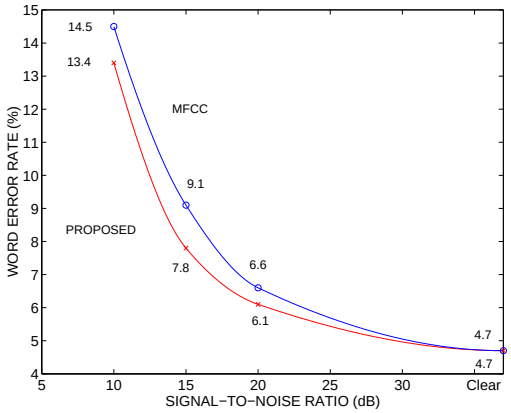

Figure 5: Comparisons on the Wall Street Journal Task

4. CONCLUSIONS

A new auditory system-based feature was proposed in this paper for improving the robustness of recognition performances. Efficient signal processing functions were implemented to mimic human auditory system. Based on the analysis of outer and middle ear, a transfer function was constructed to replace the commonly used preemphasis filter. A new set of digital auditory filters, which bear resemblance to the auditory filter in cochlea, was proposed to replace those used in the MFCC and PLP features. The ASR performance is partially improved by the new outer-middle-ear transfer function, but the major improvement is from the new set of auditory filters. The shape of the auditory filter can be adjusted during training for optimal performance. Since there is only one prototype filter and it is symmetric, the adjustment is minimal. The recognition experiments showed significant improvements on both the connect digit and WSJ tasks in various noise environments, model structures, and training algorithms.

5. ACKNOWLEDGMENT

The authors would like to thank O. Ghitza, R. Chengalvarayan, R. Sukkar, F. E. Korkmazskiy, and C.-H. Lee for useful discussions.

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] 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*, pp. III 51–54, Oct. 2000.

[3] 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.

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

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

[6] W. Reichl and W. Chou, "Decision tree state tying based on segmental clustering for acoustic modeling," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, (Seattle, WA), pp. 801–804, May 1998.

[7] W. Chou, C.-H. Lee, and B.-H. Juang, "Minimum error rate training of inter-word context dependent acoustic model units in speech recognition," in *Proceedings of Int. Conf. on Spoken Language Processing*, pp. 432–439, 1994.

[8] W. Chou, C.-H. Lee, and B.-H. Juang, "Segmental GPD training of an hidden Markov model based speech recognizer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 473–476, April 1992.