ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A HIGH-PERFORMANCE AUDITORY FEATURE FOR ROBUST SPEECH RECOGNITION

Qi Li, Frank K. Soong, and Olivier Siohan

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 proposed. Based on the analysis of human auditory system, the feature extraction algorithm consists of several modules: FFT, outer-middle-ear transfer function, frequency conversion from linear to Bark scales, auditory filtering, nonlinearity, and discrete cosine transform. Three recognition experiments have been conducted on connected digit recognition in wireless and land-line communications using handsets and hands-free microphones. Compared to LPCC and MFCC features, the proposed feature has shown 11% to 23% error-rate reductions on average in handset and hands-free acoustic environments in the experiments.

1. INTRODUCTION

Feature extraction is the first crucial block in any automatic speech recognition (ASR) system. Currently, there are two major approaches to the feature extraction: modeling either human voice production or perception systems. The most popular LPCC and MFCC features are from modeling each of the systems respectively. To achieve better and more robust ASR performance, especially in adverse acoustic environments, a new feature extraction algorithm is desirable. After an analysis of the above two approaches, we decided to pursue an auditory system approach for a new feature.

The human auditory system consists of the following modules: outer ear, middle ear, cochlea, hair cells, and nerve system. It converts the sound represented by air pressure to nerve firing rates in various frequency bands for auditory cognition in the brain. Instead of modeling all of the mechanical, hydro-dynamic, electrical, or chemical activities in each of the modules in detail, we model the functions of each of modules from a view of information and signal processing.

2. PROPOSED AUDITORY FEATURE

A schematic diagram of the proposed auditory feature is shown in Fig. 1. The speech signal is first sampled at 8 KHz sampling rate, then blocked into 240 samples, 30 ms block. Hamming window is then applied, and the window is shifted every 80 samples. The data at each time frame are then zero-padded to produce a 1024-point FFT, which generates a spectrum of 512 values. The signal processing is then performed in the frequency domain from this point on. The magnitude of the spectrum is processed through a transfer function (TF) that models the gain of pressure in both outer and middle ears approximately. The TF is shown in Fig. 2 as the solid line, which is essentially the sum of the outer-ear TF (dash-dot line) and middle-ear TF (dashed line) with little modification on low frequency bands to compensate telephone channels. The TFs were derived from the plots of psychological experiments in [1] and [2], respectively.

The spectrum is then converted to Bark scale [3] to emulate the frequency scale in the cochlea. The relation between the Bark and linear scale is shown in Fig. 3. The 512 data points are equally spaced in the Bark scale between 2.0 to 16.4 Barks, which corresponding to a linear frequency range from 200 to 3500 Hz. The 200 Hz cut-off frequency is suggested by the outer-middle-ear TF in Fig. 2 while the 3500 Hz cut-off frequency is chosen for telephone band applications. Each point in Bark is projected onto a point in Hz as shown in Fig. 3. The value of the projected point is then obtained by linear interpolation using the values of its neighboring points in the linear domain.

In the next step, an auditory filter is applied to smooth

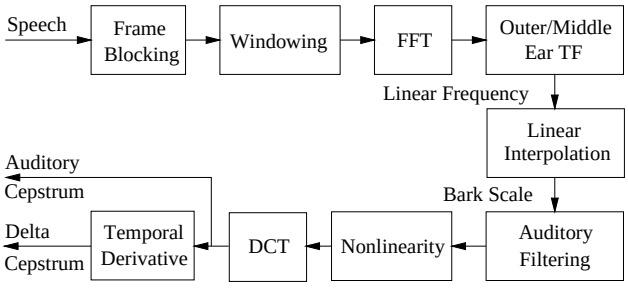

Figure 1: Schematic diagram of the proposed feature.

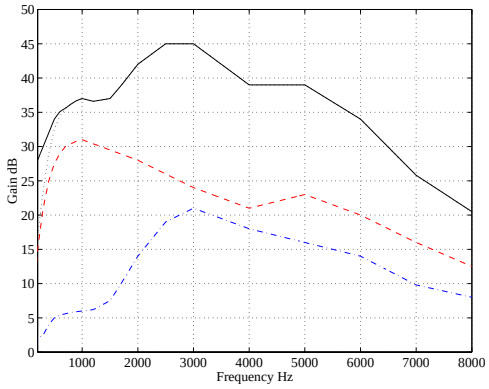

Figure 2: Outer-middle-ear transformation function.

out the speech spectrum as in the cochlea. The shape of the auditory filter is plotted in Fig. 4 in linear (top) and logarithmic (bottom) scales. It is modified from a psychophysical measurement of the frequency response of cochlea using Peterson's method [4]. The auditory filter operates as moving-average filtering has output at every point in the spectrum from 2.0 to 16.4 Barks. In the frequency domain, we consider the spectral envelopes of speech formants as signals while viewing the envelops of pitch harmonics as noise. Since the filter is in a similar shape to the formant envelops, it has high response to speech formants. On the other hand, since the size of the filter is much wider than harmonic periods and noise, the filter has low response to harmonics and noise; therefore, the function of the auditory filter is to improve the signal-to-noise ratio in the frequency domain.

In the last step, the smoothed spectrum is processed through a nonlinear function of logarithm to simulate the nonlinearity in discharge rates of auditory nerves, followed by a discrete cosine transform (DCT) to convert the logarithmic spectrum to 12 DCT coefficients. DCT is similar to the cepstral transform. It actually performs a second step to further smooth out the pitch harmonics in the spectrum. Short-term (ST) energy by accumulating the power of the blocked speech samples before Hamming window is selected as the energy term in the feature.

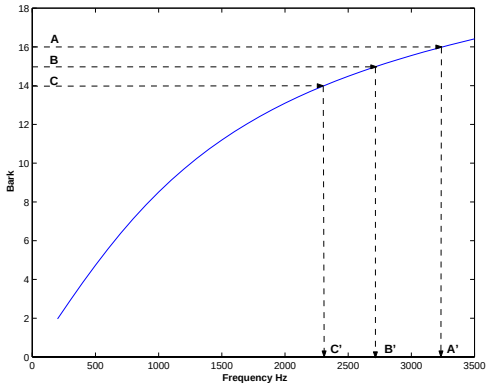

Figure 3: Frequency conversion from linear to Bark scale.

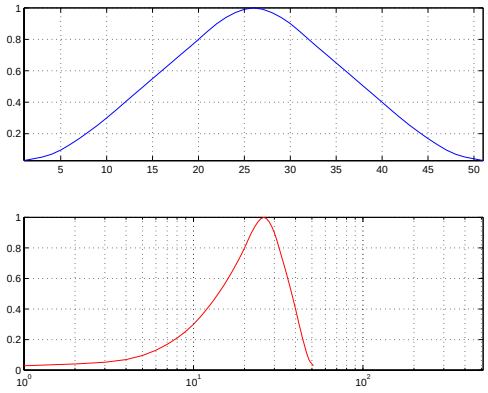

Figure 4: Shape of the auditory filter in linear (top) and logarithmic (bottom) scales.

We use the following example to show the signal processing procedures in the proposed feature. A male's voice: "Redial last number" was recorded simultaneously by a close-talking microphone, Fig. 5 (top), and a hands-free microphone, Fig. 5 (bottom). The corresponding spectrograms after FFT, outer-middle-ear transform function, and linear-to-Bark conversion are shown in Fig. 6. The spectrograms smoothed by the auditory filtering are plotted in Fig. 7. The moving-average filter is operated from low to high Barks for every frame. To show the further smoothing effect of the DCT or cepstral transform, we reconstructed the spectrograms from cepstral coefficients through IFFT in Fig. 8, where we observe that DCT smoothed out the pitch harmonics and blurred the background noise.

Compared to the LPCC feature, the proposed feature is to model the auditory perception system instead of the voice production system. Compared to the MFCC feature [5], the new feature includes an extra outer-and-middle-ear TF and uses an auditory filter determined from psychoacoustic experiments [4] instead of the triangular filters used in MFCC [5]. In addition, the Mel frequency scale are replaced by Bark scale and spectral data are spaced equally in the Bark scale through linear interpolation before auditory filtering. Compared to the PLP feature [6], the proposed feature uses magnitude instead of power spectrum. The shape of the auditory filter in the proposed system is closer to the shape of real auditory system. Also, the resolution of filtering output is the same as the spectrum while the resolution in the PLP implementation is much lower as is in MFCC. In the proposed feature, we use the outer-middle-ear TF from auditory system study directly while PLP uses equal-loudness preemphasis. Compared to RASTA feature [7], the proposed approach uses different filters and different nonlinearities at

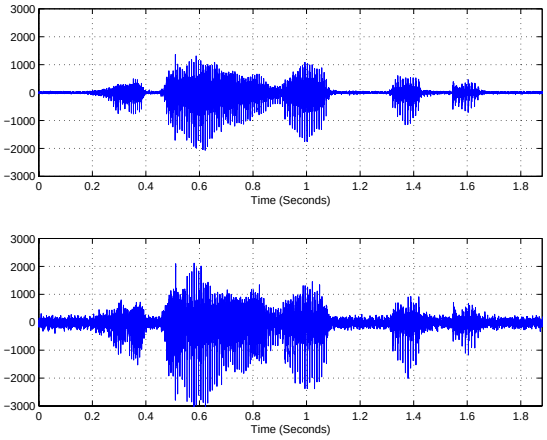

Figure 5: Male's voice: "Redial last number".

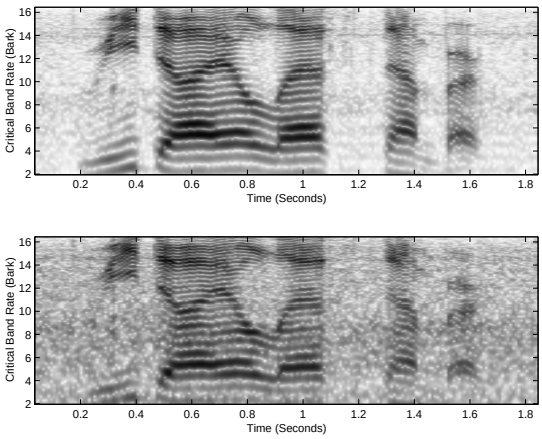

Figure 6: Spectrograms in Bark scale.

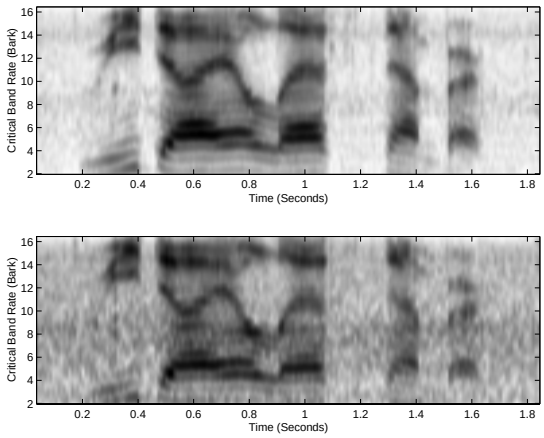

Figure 7: Spectrograms after auditory filtering.

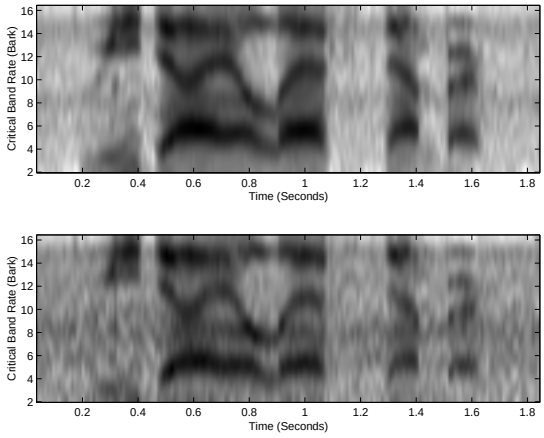

Figure 8: Spectrograms reconstructed from cepstrum.

different processing stages. In summary, all the auditory-based feature approaches mimic human auditory system one way or another while different configurations may yield different performances. We try to construct a system as close to the human auditory system as possible by applying the psychoacoustic experimental results directly while the computation speed, the system complexity, and the ASR back-end were also considered in developing the proposed feature.

3. EXPERIMENTAL RESULTS

The proposed feature has been tested on a connected digit recognition task and compared with LPCC and MFCC features. All the experiments used 39 dimensional features including energy, 12 cepstral or DCT coefficients, plus their first and second order time derivatives. Ten-digit utterances are used in all the tests and the performances are reported in string error rates. There is no overlap between any training and test utterances.

3.1. Focus on Hand-Free Data

In this experiment, we used two CDMA wireless databases named "Handset" and "Lapel", respectively. The utterances in Handset were recorded with handsets while the utterances in Lapel were recorded with the microphones located on 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. Since Lapel has many more training utterances than Handset, this experiment is intended to show the robustness of the proposed feature in a hands-free environment. The models are context-dependent digit HMMs (10 states per digit and average 7 mixtures per states). The state tying for every states in the models is determined by decision trees [8]. There are totally 1,400 CD models with 800 tied states. The training algorithm is based on maximum likelihood estimation (MLE). In Table 1, the numbers are string error rates while the digit error rates are given in parentheses. On average, the string error rates for LPCC, MFCC, and the proposed features are 11.7%,

Table 1: Comparisons on String Error Rates (%) (digit recognition error rates are in parentheses)

Data	Handset	Lapel	Average
LPCC	9.4 (1.0)	13.9 (2.0)	11.7 (1.5)
MFCC	7.0 (0.9)	14.9 (1.9)	11.0 (1.4)
Proposed Feature	6.6 (0.9)	11.4 (1.5)	9.0 (1.2)

11.0%, and 9.0%, respectively. The proposed feature has shown a 23% error rate reduction on average compared to the standard LPCC cepstral feature.

3.2. Focus on Close-Talking Microphone Data

We further extended the experiment using 20 databases for training and 4 databases for test. Since 19 training databases were recorded with handsets in wireless and land-line environments and only Lapel was recorded by hands-free microphones, this experiment focuses on handset environment. The model structure and training algorithm are the same as above. In total, there are 42,243 training utterances. The testing databases: Tele, Sdn10, Handset, and Lapel, have 518, 1685, 256, and 517 utterances, respectively. The experimental results in term of utterance error rates were listed in Table 2. One average, the proposed feature shows an 18% error reduction compared to the LPCC feature.

Table 2: Comparisons on String Error Rates (%)

Data	Tele	Sdn10	Handset	Lapel	Ave.
LPCC	7.0	14.7	7.0	21.3	12.5
Proposed	5.0	11.7	4.7	19.3	10.2

3.3. Head-Body-Tail (HBT) Digit Model

To evaluate the proposed feature with different HMM structure, we trained a set of HBT models [9] using the above 20 databases plus a database “Visor" (recorded with microphones mounted on the visor of a moving car), a total of 44,123 utterances. The digit HBT model is context-dependent across the “head" (first 3 states) and “tail" (last 3 states) while the “body" (4 states) is context independent. After MLE training, the model was tested on 4 databases as above. The utterance error rates are listed in Table 3. The HBT models work better than the CD models. Again, the new feature reduced the error rates on all the tested databases. On average, the proposed feature has an 11% error rate reduction compared to the LPCC feature.

Table 3: String Error Rates of HBT Models (%)

Data	Tele	Sdn10	Handset	Lapel	Ave.
LPCC	5.8	13.1	5.9	18.0	10.7
Proposed	4.8	11.5	5.5	16.4	9.5

4. CONCLUSIONS

An auditory based feature extraction algorithm was proposed in this paper. There are several steps in computing the proposed feature: FFT, outer-middle-ear TF, Bark scale conversion, auditory filtering in the frequency domain, non-linearity, and DCT. The experiments have shown that the proposed feature reduced the string recognition error rates in both close-talking and hands-free environments. On average, the error rate reductions are from 11% to 23%. The error rates can be further reduced if we apply a discriminant training algorithm with the proposed feature. We are in the process of extending the proposed feature to wide band and large vocabulary continuous speech recognition.

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] E. A. G. Shaw, *The external ear*, in Handbook of Sensory Physiology. New York: Springer-Verlay, 1974. W. D. Keidel and W. D. Neff eds.

[2] V. Nedzelnitsky, “Sound presures in the casal turn of the cat cochlea,” *J. Acoustics Soc. Am.*, vol. 68, 1980.

[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, 1980.

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

[5] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentnces,” *IEEE Trans. on Acoustics, sppech, and signal proc.*, vol. ASSP-28, pp. 357–366, 1980.

[6] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *J. Acoust. Soc. Am.*, vol. 87, no. 4, 1990.

[7] H. Hermansky and N. Morgan, “Rasta processing of speech,” *IEEE Trans. Speech and Audio Proc.*, vol. 2, Oct. 1994.

[8] W. Reichl and W. Chou, “Decision tree state tying based on segmental clustering for acoustic modeling,” in *Proc. IEEE ICASSP*, pp. 801–804, May 1998.

[9] W. Chou, C.-H. Lee and B.-H. Juan, “Minimum error rate training of inter-word context dependent acoustic model units in speech recognition,” in *Proc. ICSLP*, pp. 439-432, 1994.