AN ACOUSTIC FRONT-END USING WARPED FREQUENCY AND TEMPORAL RESOLUTIONS

B.T. Lilly and K.K. Paliwal

Signal Processing Laboratory
School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia
B.Lilly, K.Paliwal@me.gu.edu.au

ABSTRACT

Typically, the power spectrum of a speech frame used in speech recognition is estimated for a fixed length window using the fast Fourier transform. Each frequency component represented in this power spectrum is an estimate over that speech frame. The power spectrum calculated in this way has a constant time and frequency resolution. An example of this type of front-end is the LPC-derived cepstral front-end commonly used is recognition systems today [1]. The acoustic front-end presented in this paper employs both a warped frequency and temporal resolutions. We show that a front-end that utilises both warping functions, outperforms a front-end that employs only a warped frequency scale. We also show that this new front-end is unsuitable for noisy conditions.

1. INTRODUCTION

Typically, the power spectrum of a speech frame used in speech recognition is estimated for a fixed length window using the fast Fourier transform. Each frequency component represented in this power spectrum is an estimate over that speech frame. The power spectrum calculated in this way has a constant time and frequency resolution. An example of this type of front-end is the LPC-derived cepstral front-end commonly used is recognition systems today [1].

A front-end that employs a warped frequency resolution is the mel-space cepstral front-end [2] which warps the frequency scale similar to what is found in the human ear. It is well-known that when the human ear performs frequency analysis of a speech signal, it has different temporal (time) resolutions at different frequencies [3, 4]. The wavelet-transform based acoustic front-ends [7, 8] utilize this property of human ear to some extent. These front-ends divide the time-frequency plane in a discrete manner such that the power spectrum for lower frequencies is estimated over a longer time window than for the higher frequencies. However, they are not flexible enough to utilize this property to its full extent.

In this paper, we try to simulate the model of the human auditory system as described in [4, 5]. This model includes a warped temporal window that is inversely proportional to the bandwidth of the corresponding filter in the filter bank. Section 2 describes this model in more detail and describes the method in which we have implemented it. Section 3 shows the experimental setup used in the recognition experiments in this paper. Section 4 explains the results obtained and Section 5 presents our conclusions.

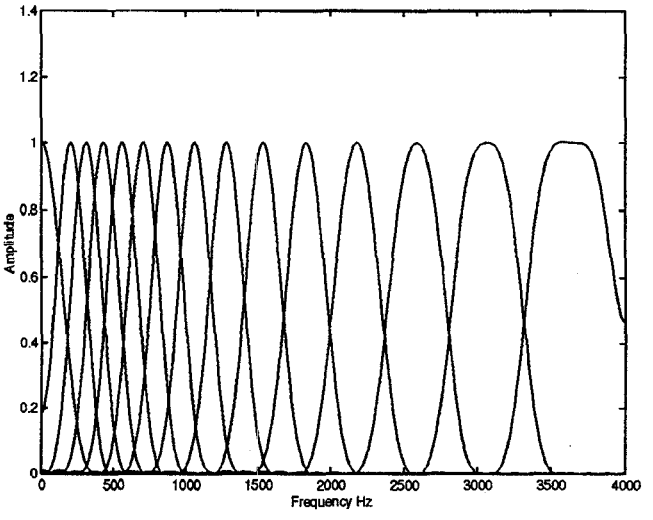

Figure 1. Filter bank using 65 tap FIR filters.

2. ACOUSTIC MODEL IMPLEMENTATION

As stated in the introduction, we have designed a front-end that simulates the model of human auditory system described in [4, 5]. This model can be described in terms of four successive stages:

1. A bank of bandpass filters,
2. each followed by a nonlinear device,
3. a sliding temporal integrator (or temporal window) and
4. a decision device.

For the first stage, we use a bank of FIR bandpass filters. These filters have 1 Bark spacing on the frequency scale. The bandwidth of each filter is equaled to the critical bandwidth corresponding to the filter's center frequency [6] (see Table 1). We have experimented with two different filter banks. Figure 1 shows the filter bank where the impulse response length of the FIR filters was kept constant. FIR filters containing 65 taps are used for all filters in the filter bank. We also used a filter bank that used an impulse response length of the FIR filter that was inversely proportional to its bandwidth as shown in Table 1. The impulse

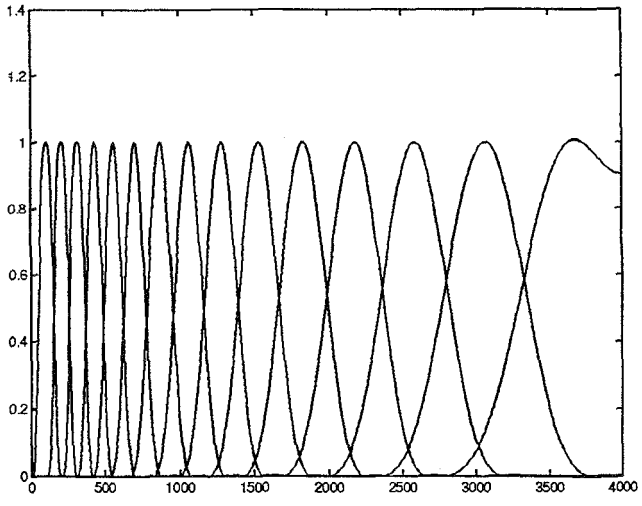

Figure 2. Filter bank that varies the impulse response length with the bandwidth of the FIR filter. The impulse response length corresponding to the bandwidth as given in Table 1.

response length/bandwidth product was kept constant with an initial impulse length of 27 fixing the highest frequency filter.

For the second stage of our auditory model, we have used the squaring function as a nonlinearity. For the third stage, we have used a rectangular window. The duration of this window is taken to be inversely proportional to the bandwidth of the filter. Table 1 shows the duration of the window used as a function of bandwidth. Again, the window length/bandwidth product was held constant with an initial window length of 5ms for the highest frequency filter.

The output of the third stage gives a measure of power of the signal at a frequency equaled to the filter's center frequency. Let us denote output of the ith filter by p_i and its frequency being f_i. If there are N filters in the filterbank, then their outputs can be used to compute the p cepstral coefficients as follows:

$$c_k = \sum_{n=1}^{N} \log p_n \cos(2\pi f_n k / F_s).$$

for $k = 1, 2, ..., p$. In this equation, F_s is the sampling frequency. These p cepstral coefficients derived from this front-end are used as recognition features.

The fourth stage then consists of multi-mixture HMMs to perform the decision. We use this model to design several front-ends for use in speech recognition experiments to observe any improvements in recognition performance.

Table 2 shows a summary of the front-ends used in experiments presented in this paper. The LPCC front-end warps neither the frequency nor the temporal resolutions. It is used as a benchmark to evaluate any improvements in performance. All the other front-ends utilise a warped frequency scale. The BARK-FIR front-end uses a simple filter bank of FIR filters to calculate the power in each band similar to the mel-space cepstral front-end in [2]. The T-BARK-FIR front-end uses a warped temporal scale and a constant

Table 1. Bandwidths used for each filter bank with the corresponding window length and impulse response length.

Filter	Bandwidth (Hz)	Window Length (ms)	Impulse Response Length
1	100.7296	30.1	163
2	102.9805	29.4	159
3	106.940	28.4	153
4	112.929	26.8	145
5	121.406	25.0	135
6	132.989	22.8	123
7	148.471	20.4	111
8	168.857	18.0	97
9	195.402	15.5	83
10	229.678	13.2	71
11	273.650	11.1	59
12	329.783	9.2	49
13	401.172	7.6	41
14	491.711	6.2	33
15	606.298	5.0	27

impulse response length for all the filters in the filter bank. The last two front-ends (BARK-VFIR and T-BARK-VFIR) are similar to the previous two front-ends except they use a filter bank where the impulse response length is inversely proportional to the bandwidth of the corresponding filter. These lengths are shown in Table 1.

3. RECOGNITION AND DATABASE SETUP

Experiments in this paper are performed on the e-set alphabets (B,C,D,E,G,P,T,V,Z) of the ISOLET database. There are 1620 utterances from 90 speakers used for training and 540 utterances from 30 speakers used for testing. The database consists of the same number of males as females and different speakers are used for training and testing.

The ISOLET database is sampled at 16kHz with a 16 bit resolution. We have decimated it to 8kHz using a low pass filter with a cutoff frequency of 3.5kHz. Speech frames are computed every 5ms. The LPCC, BARK-FIR and BARK-VFIR front-ends use a constant temporal resolution with frequency with a width of 20ms. The T-BARK-FIR and T-BARK-TFIR front-ends vary the temporal resolution according to Table 1. Fifteen filter bank energies are used to calculate 12 cepstral coefficients. Delta coefficients are also used in additional experiments to evaluate the recognition performance of this auditory model.

We use simple left-to-right HMM models for the recognition experiments. The isolated word models are modeled with 5 states each with 5 Gaussian mixtures with diagonal covariances.

4. RESULTS

Table 3 shows the recognition results. From this table, it is clear that a front-end that uses a warped frequency scale (BARK-FIR) has little improvement over a front-end that uses a uniform frequency scale (LPCC). However, a significant performance improvement is observed for the front-end that employs both warped frequency and temporal resolutions (T-BARK-FIR).

Table 2. Features of the various front-ends used in experiments presented in this paper.

Front-end	Warped Frequency Resolution	Warped Temporal Resolution	Varying Impulse Response
LPCC	×	×	×
BARK-FIR	√	×	×
T-BARK-FIR	√	√	×
BARK-VFIR	√	×	√
T-BARK-VFIR	√	√	√

Table 3. Recognition performance for front-ends using a 12th order cepstral vector.

Front-end	Recognition Accuracy (%)
LPCC	54.44
BARK-FIR	55.19
T-BARK-FIR	61.85
BARK-VFIR	53.52
T-BARK-VFIR	60.00

Table 4. Recognition performance for front-ends using a 24th order cepstral vector.

Front-end	Recognition Accuracy (%)
LPCC	67.41
BARK-FIR	73.89
T-BARK-FIR	74.81
BARK-VFIR	72.22
T-BARK-VFIR	72.59

This can be attributed to the power spectrum of high frequency components having a higher resolution. This provides more detail to the frequency components of speech that are normally averaged out in a uniform temporal analysis.

Table 3 also shows that the front-ends that vary the impulse response length with bandwidth, show a corresponding decrease in performance. However, the front-end that does utilise a warped temporal resolution (T-BARK-VFIR) does perform better than the front-end that doesn't (BARK-VFIR). We can therefore conclude that a warped temporal resolution provides an increase in performance no matter which filter bank is chosen.

To verify these results, we have also performed experiments by appending 12 delta coefficients to the original feature vectors. As can be seen from Table 4, the T-BARK-VFIR front-end shows only a slight improvement in recognition performance over any of the other front-ends. The addition of delta coefficients has shadowed the effect the warped temporal resolution had on the recognition performance.

Table 4 also confirms that by varying the impulse response of the filters in the filter bank with respect to their bandwidth, does not provide any improvements in recognition performance. Another interesting result is the T-BARK-VFIR front-end showed virtually no increase in performance using the filter bank with varying impulse response lengths.

5. CONCLUSIONS

This paper has shown that a front-end that incorporates a warped temporal resolution significantly increases the recognition performance over a front-end that doesn't. The addition of delta coefficients however, did not show the same increase in performance using a warped temporal resolution as would have been expected.

REFERENCES

[1] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech and Signal Processing*, Vol. 28, No. 4, pp. 357-366, Aug. 1980.

[3] D.A. Eddins and D.M. Green, *Hearing*, Academic Press, USA, 1995.

[4] B.C.J. Moore, *An Introduction to the Psychology of Hearing*, 3rd Edition, Academic Press, London, 1989.

[5] B.C.J. Moore, B.R. Glasberg, C.J. Plack and A.K. Biswas, "The shape of the ear's temporal window", *J. Acoust. Soc. Am.*, Vol. 88, pp. 1102-1116, 1988.

[6] B.C.J. Moore and B.R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", *J. Acoust. Soc. Am.*, Vol. 74, pp. 750-753, 1983.

[7] H. Wassner and G. Chollet, "New cepstral representation using wavelet analysis and spectral transformation for robust speech recognition", *Proc. Int. Conf. Spoken Language Processing*, Vol. 1, pp. 260-263, October 1996.

[8] C.J. Long and S. Datta, "Wavelet Based Feature Extraction for Phoneme Recognition", *Proc. Int. Conf. Spoken Language Processing*, Vol. 1, pp. 264-267, October 1996.