Fifth International Symposium on Signal Processing and its Applications,
ISSPA '99, Brisbane, Australia, 22-25 August, 1999
Organised by the Signal Processing Research Centre, QUT, Brisbane, Australia

USE OF SPECTRAL SUBBAND MOMENTS IN MFCC COMPUTATION

Eigil Gjelsvik and Kuldip K. Paliwal

School of Microelectronic Engineering
Griffith University
Brisbane,QLD 4111, Australia
eigil@excalibar.me.gu.edu.au, K.Paliwal@me.gu.edu.au

ABSTRACT

Mel frequency cepstral coefficients (MFCCs) are currently the most popular form of parameterization of the speech signal in speech recognition systems. In this paper, we look at a way to improve the extraction of these features using information about the spectral characteristics of the signal to modify filter-bank shapes. This information is captured in the form of spectral moments of the subbands. We show that this improves speech recognition performance, but the improvement is not very significant.

1. INTRODUCTION

Selection of proper acoustic features representing the speech signal is one of the most important tasks in the design of a speech recognition system. One wishes to extract parameters that represents the maximum information necessary for speech recognition and that are, at the same time, independent of irrelevant information such as speaker characteristics, manner of speaking, background noise, channel distortion etc. Cepstral coefficients, along with its time derivatives, are the most common features used in today's speech recognition systems. MFCC features are currently the most popular of these features.

It has been shown [1] that information about the subband spectral centroids in the speech signal can be used to improve speech recognition. The centroids provide a crude estimation of the formants of the signal, which in turn have physical interpretation as vocal tract resonances. Formant frequencies were actually used as recognition features in the sixties but they have been lately abandoned due to difficulties of accurate estimation. However, spectral centroids still provide useful information on where to find local energy maxima in the signal. Because the speech power spectrum in the formant frequency area of the signal will be less affected by additive white noise, spectral centroids are also robust to noise.

The spectral centroid of a given subband is computed as the first moment of the power spectrum within the subband. In this paper, we want to use the information captured by the higher-order moments of the power spectrum as well. We utilize this information to improve extraction of MFCC coefficients from the FFT spectrum. More precisely, we wish to apply Gaussian windows, derived from the first and second order spectral moments, as filters, instead of the triangular filters in the MFCC filter-banks. This is explained in more detail in the next section.

2. MFCC FEATURES AND SPECTRAL CENTROIDS

Computation of traditional MFCC coefficients consists of 4 steps [2]:

1. Computation of FFT from input speech signal.
2. Smoothing of FFT spectrum by integrating the spectral coefficients within *triangular* frequency bins (Fig. 1) arranged uniformly on the nonlinear *mel-frequency* scale.
3. Discrete Cosine Transform of the logarithm of the filter-bank output.
4. (optional) Append first and second order time differentials to incorporate dynamic information about the signal.

In this paper, we will try to improve recognition performance by modifying step 2. The triangular filters used for MFCC computation are totally independent of the nature of the speech signal. For instance, if there is a high-energy area in the left part of a filterbank, this might be partially suppressed as a result of the filtering. If we instead manage to find a window that is adapted to the shape of the power spectrum within each subband frequency bin, this could help us to better capture the energies in that bin, thus giving a better modeling of the signal. Our choice of filter window is a Gaussian filter centered at the spectral centroid (first order moment) in each bin and with variance equal to the second order moment.

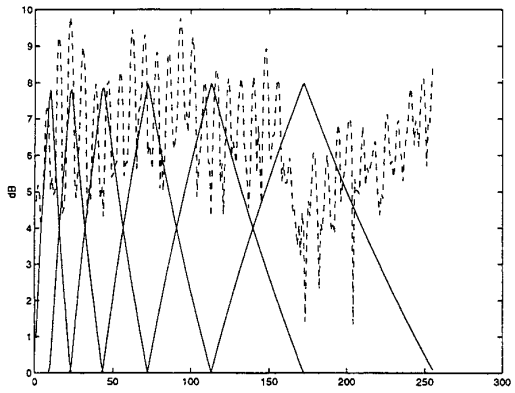

Figure 1: The mel-frequency triangular filters

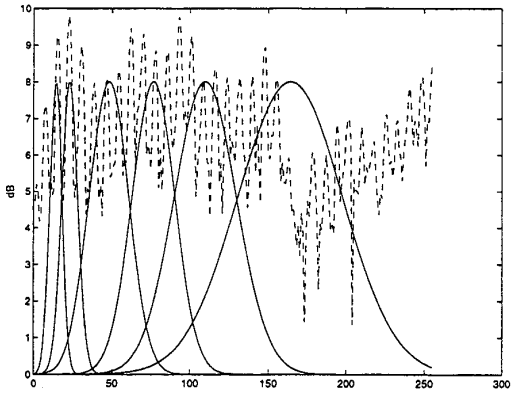

Figure 2: FFT spectrum of a signal and its adapted Gaussian filter-banks when $h(\sigma_m) = 1$.

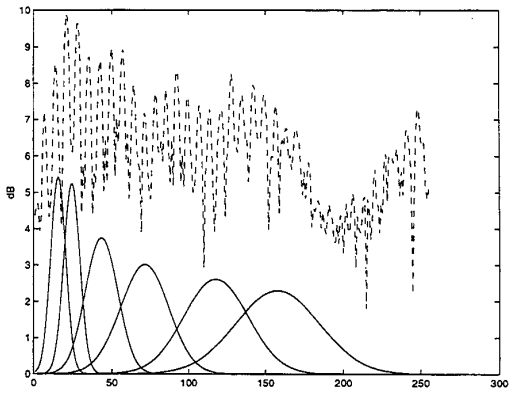

Figure 3: FFT spectrum of a signal and its adapted Gaussian filter-banks when $h(\sigma_m) = \frac{1}{\sqrt{2\pi}\sigma_m}$.

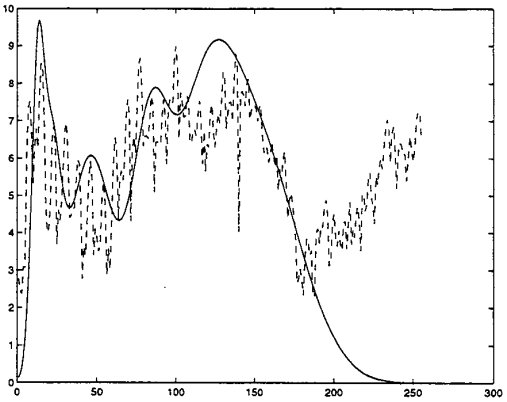

Figure 4: FFT spectrum of a signal and the envelope of all its adapted Gaussian windows for $h(\sigma_m) = 1$.

Let us assume that the frequency band $[0, F_s/2]$ is divided into M subbands, uniformly distributed on the mel-frequency scale. Let the lower and higher edges of the mth subband be l_m and h_m respectively. In general the subbands overlap, that is, $h_m > l_{m+1}$. We define the mth subband spectral centroid C_m as follows:

$$C_m = \frac{\int_{l_m}^{h_m} f w_m(f) P^\gamma(f) df}{\int_{l_m}^{h_m} w_m(f) P^\gamma(f) df} \tag{1}$$

where $P(f)$ is the power spectrum of the speech signal. γ is a constant determining the dynamic range, and can be set to a value that optimizes the recognition performance. $w_m(f)$ is a window that weights the significance of the spectral coefficients in each bin [1]. The second order moment, σ_m^2 is defined as follows:

$$\sigma_m^2 = \frac{\int_{l_m}^{h_m} (C_m - f)^2 w_m(f) P^\gamma(f) df}{\int_{l_m}^{h_m} w_m(f) P^\gamma(f) df} \tag{2}$$

We then define the mth Gaussian window as

$$g_m(f) = h(\sigma_m) e^{-\frac{(f-C_m)^2}{2\sigma_m^2}} \tag{3}$$

where $h(\sigma_m)$ is a weighing function that can be used to set the weighing of each subband m as a function of σ_m. We have investigated two realizations of the function $h(\sigma_m)$:

1. $h(\sigma_m) = 1$
2. $h(\sigma_m) = \frac{1}{\sqrt{2\pi}\sigma_m}$

The shapes of the Gaussian filters for these two cases are shown in Figs. 2 and 3, respectively. The smoothed spectrum will now be obtained by integrating the product of the spectral coefficients and the Gaussian window within each frequency bin, e.g; we integrate between l_m and h_m.

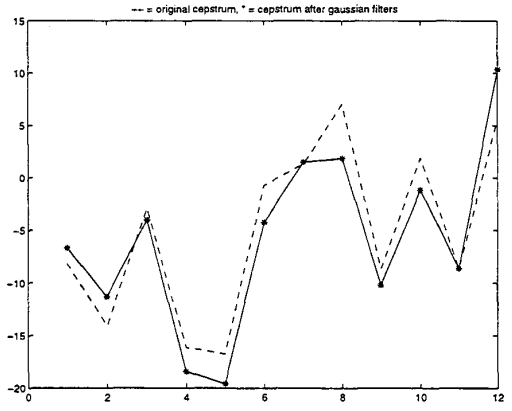

Figure 5: Cespstral coefficients using 12 parameters after Gaussian filtering compared to the coefficients after triangular filtering for the phone 'eh'

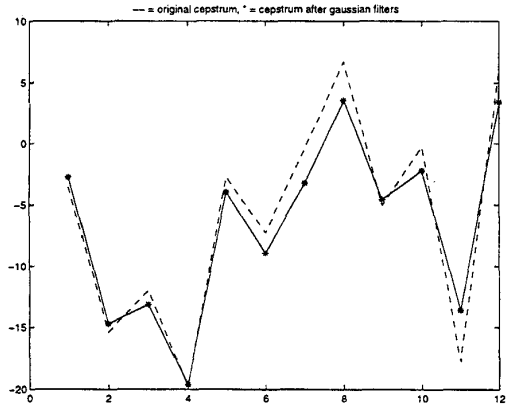

Figure 6: Cespstral coefficients using 12 parameters after Gaussian filtering compared to the coefficients after triangular filtering for the phone 'ah'

In order to have a smooth transition from one subband to the other, it may be desirable to increase the overlap between the subbands. One way to do this is to combine the effect of all the Gaussian windows into one filter-shape that covers the whole spectrum, thus producing an envelope of all the Gaussian filters. This envelope will then be used as a weighing function for the spectral coefficients inside each bin, replacing the triangular filters. The envelope function $e(f)$ will look like this:

$$e(f) = \sum_m h(\sigma_m) e^{-\frac{(f-c_m)^2}{2\sigma_m^2}} \quad (4)$$

The power in each bin will now be calculated by integrating the product of the FFT coefficients and the envelope inside each bin. An example of such an envelope is given in Fig. 4.

A third realization is worked out by combining the Gaussian envelope with the traditional triangular filters. By multiplying the original FFT spectrum with the envelope spectrum before integrating within the triangular bins, we achieve a sort of 'prefiltering before the filters' that 'fits' the signal to the bins before we integrate up the coefficients.

3. RESULTS FROM SIMULATIONS

We evaluate the performance of the modified MFCCs using the HTK toolkit for monophone HMMs on the TIMIT database. The database consists of sentences spoken by speakers from 8 different regions of the USA. The sentences are both phonetically balanced and designed to show differences between dialects (e.g.; we use all the *sa-* and *sx* - sentences). The training data and the test data are different. Speech is digitized at a sampling rate of 16 kHZ. The signal is analyzed every 10 ms with a frame width of 25 ms (with Hamming window and preemphasis). The experiments are carried out using 24 subband filters, deriving 12 cepstral coefficients plus the total energy. Then their first order differentials are appended, giving a total vector-size of 26 parameters for each frame.

At first, the experiments are run on clean speech. Then a second experiment is run with white Gaussian noise added to the test data. The SNR ratio is 20 dB. The constant γ is set to 0.5 [1].

Simulations were carried out using both triangular and rectangular shapes for the windows $w_m(f)$ from equations (1) and (2). Triangular shapes were found to be better, and are used throughout this paper. As for the weighing function $h(\sigma_m)$ from equation (3), the best results were obtained by setting $h(\sigma_m) = 1$.

We also experimented with the effect of the band edges l_m and h_m while calculating the moments. When the subbands overlapped, e.g; the lower band edge was set at the center of the previous subband, we got better results for all realizations than when the band-edges were set consecutively next to each other.

We can see from Table 1 that recognition performance has changed very little compared to traditional MFCC computation. However there is a slight improvement for all the new realizations when in presence of noise.

Filter Realization	**Recognition accuracy**	
	Clean	*Noisy*
Triangular	59.46%	33.43%
Gaussians	59.32%	33.96%
Envelope	59.27%	33.84%
Envelope & triangular	59.49%	33.67%

Table 1: Speech recognition results using 24 subbands

To order to understand these results, we show in Figs.5 and & 6 the cepstral coefficients for the vowels 'a' and 'e' after applying Gaussian filtering in each bin. We can see that cepstral coefficients are slightly changed, but their overall trend is more or less the same.

4. CONCLUSION

We have investigated how we can modify the shape of the subband filters used in MFCC parametrisation using information about the first and second order moments in each subband of the speech signal. Our results so far do not give evidence that this leads to a significant improvement of speech recognition performance. However, the idea of using spectral knowledge about the signal to shape the subband filters remains interesting.

5. ACKNOWLEDGEMENT

The research reported in this paper was partially funded by the ARC Grant No. A49906288. The second author is thankful to Dr. Alain Biem for useful discussion.

6. REFERENCES

[1] K. K. Paliwal. "Spectral subband centroids features for speech recognition." *Proc. ICASSP 98, pp 617-620.*

[2] S. Young. "A review of large-vocabulary continuous-speech recognition." *IEEE Signal Processing Magazine, pp 45-57, Sept. 1996.*