Spectral Subband Centroids as Features for Speech Recognition

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

Abstract - Cepstral coefficients derived either through linear prediction (LP) analysis or from filter bank are perhaps the most commonly used features in currently available speech recognition systems. In this paper, we propose spectral subband centroids as new features and use them as supplement to cepstral features for speech recognition. We show that these features have properties similar to formant frequencies and are quite robust to noise. Recognition results are reported in the paper justifying the usefulness of these features as supplementary features.

1 Introduction

Selection of proper acoustic features is perhaps the most important task in the design of a speech recognition system. It directly affects the performance of a speech recognizer. These features should be selected in such a manner that they should contain maximum information necessary for speech recognition and, at the same time, discard irrelevant information such as speaker characteristics, manner of speaking, background noise, channel distortion, etc. Feature selection is a difficult task and a great deal of research has been done to identify these features (see [1] and references given therein for different front-ends). Once these features are selected, they are extracted from the speech signal on the frame-by-frame basis.

Cepstral coefficients derived either through linear prediction (LP) analysis or from filter bank are perhaps the most commonly used features in currently available speech recognition systems [2]. These features provide a reasonable recognition performance, and have served the speech recognition community quite well for the last two decades. However, we believe that time has come for investigating the use of new features, if we want to improve the speech recognition performance significantly. Here, we are not suggesting to replace the cepstral features (which have been a great success in the past) by the new features. Instead we want the new features to be used alongwith the cepstral features. If this increases the dimensionality of the feature space, linear discriminant analysis may be used for dimensionality reduction [3, 4].

0-7803-3698-4/97/$10.00 © 1997 IEEE.

One of the major problems with the cepstral features is that they are very sensitive to additive noise distortion. Addition of white noise to the speech signal affects the speech power spectrum at all the frequencies, but the effect is less noticable in the higher amplitude (formant) portions of the spectrum (i.e., signal-to-noise ratio is more in the formant regions than in the non-formant regions). Since cepstrum features use formant as well as non-formant regions of the power spectrum in their computation, they become very sensitive to additive white noise. This problem can be overcome by using formant frequencies as features, as the formant locations are not disturbed by the additive noise distortion. In addition to this robustness to noise, formants have many other advantages. For example, they provide most parsimonious representation of the spectral envelope and have physical interpretation as vocal tract resonances. Because of these advantages, the formant frequencies were used as recognition features in the sixties. But they have been lately abandoned mainly due to the problems associated with their estimation from the speech signal. These problems arise due to merging of peaks in the spectrum and appearance of spurious peaks in the spectrum. These problems cause gross errors in formant extraction. If we can overcome these problems or devise features which have properties similar to formant frequencies, we can improve the speech recognition performance.

In this paper, we want to investigate some formant-like features for speech recognition. Obviously, these features should provide information for speech recognition different from the cepstral features, if they are to be used as a supplement to the cepstral features. In this paper, we propose to use spectral subband centroids (SSCs) as supplementary features for speech recognition. These features are having similarities with the formant frequencies and can be extracted easily and reliably (without any estimation errors) from the power spectrum of the speech signal. In this paper, we provide recognition results justifying the use of these features as supplementary features for speech recognition.

2 Spectral Subband Centroids (SSCs)

In order to define spectral subband centroids, we divide the frequency band (i.e.; 0 to $F_s/2$, where F_s is the sampling frequency in Hz) into a fixed number of subbands and compute the centroid for each subband using the power spectrum of the speech signal. Though this definition looks simple, there are a number questions we have to answer here. For example, we have to decide how many subbands to be used, how should the frequency band be divided into subbands (i.e., what should be the center and cutoff frequencies of the subband filters and whether should the subbands be disjoint or overlap each other), what should be the shape of subband filters, whether to use unsmoothed (FFT) power spectrum or the smooth spectral envelope (computed through LP analysis) for computing the centroids, whether to compress the dynamic range of the power spectrum for centroid computation and to what extent, etc. In this paper, we try to provide answers to these questions through experimentation using recognition performance as the criterion.

Let us assume that the frequency band $[0, F_s/2]$ is divided into M subbands. Let the lower and higher edges of mth subband be l_m and h_m, respectively, and its filter shape be $w_m(f)$. 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 and γ is a constant controlling the dynamic range of the power spectrum. By setting $\gamma < 1$, the dynamic rage of the power spectrum can be reduced.

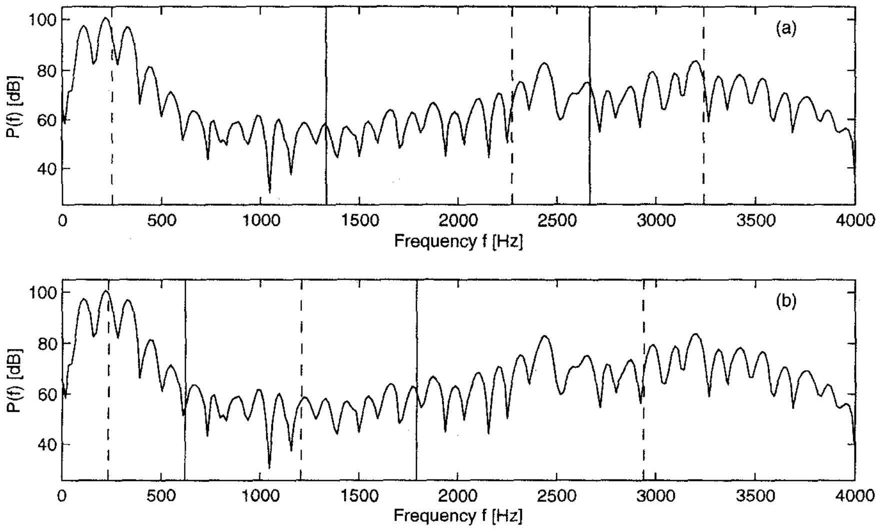

Figure 1: Subband spectral centroids (SSCs) from the FFT power spectrum of the vowel /ee/. The subband boundaries are shown in solid vertical lines and the centroids by dashed vertical lines. (a) Uniform subband division on Hz scale and (b) uniform subband division on mel scale.

3 Analysis Results

For illustration purposes, we compute M SSCs using the unsmoothed (FFT) power spectrum. We divide the frequency band $[0, F_s/2]$ into M equal-length disjoint subbands and employ a rectangular shape for each subband filter. The lower and higher edge frequencies of M subbands are given by

$l_1 = 0$, $h_M = F_s/2$ and $l_{m+1} = h_m = m * F_s/(2 * M)$, for $m = 1, 2, ..., M - 1$. We use a 30-ms long segment of vowel /ee/ sampled at 8000 Hz to perform SSC analysis for $M = 3$ and $\gamma = 0.5$. Results are shown in Fig. 1(a). Note that the centroids are located in this figure at frequencies different from the formant frequencies. If the aim were to make the centroids nearer to the formant frequencies, it could have been done by choosing a larger value for γ. Instead of dividing the frequency band uniformly on the Hz scale, we can divide it uniformly on the mel scale. The resulting subband centroids are shown in Fig. 1(b). When the frequency band is divided uniformly on the Hz scale, we call the resulting SSCs as the Hertz Frequency Centroids (HFCs). Similarly, the SSCs computed by dividing the frequency band uniformly on mel scale are called the Mel Frequency Centroids (MFCs).

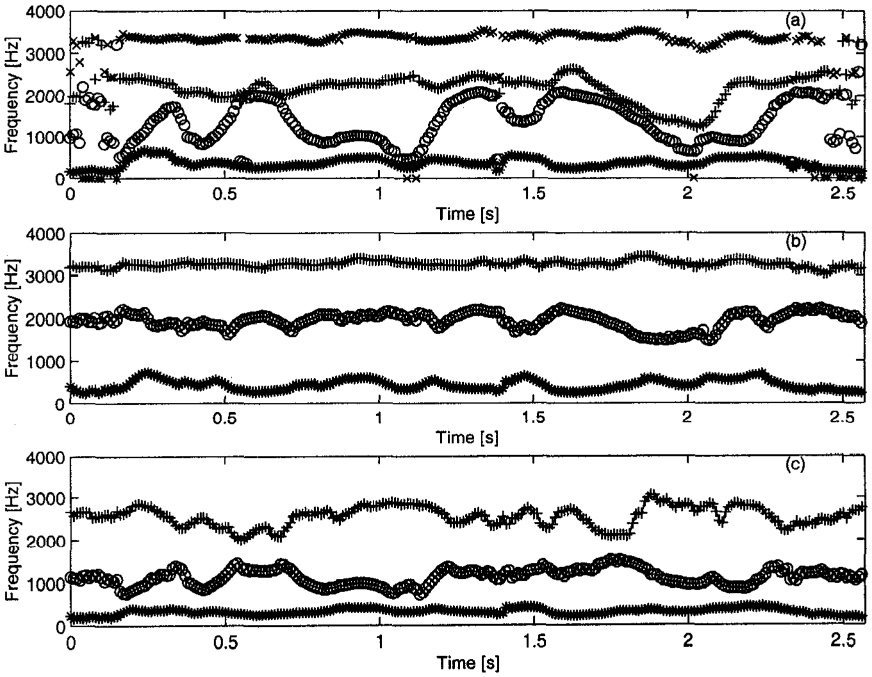

Figure 2: SSCs for the utterance "Why were you away a year Roy?" with $M = 3$ and $\gamma = 0.5$. (a) 4 poles from LP analysis, (b) HFC-FFT, and (b) MFC-FFT.

In Fig. 1, we have used the unsmoothed (FFT) power spectrum for computing the subband spectral centroids. We can also use the smooth (LP) power spectrum for the centroid computation. This will give another two sets of centroids depending on the division of the frequency band on the Hz scale or the mel scale. Thus, we have introduced four types of subband spectral centroids. These are: 1) HFC-FFT (FFT power spectrum and Hz scale),

2) MFC-FFT (FFT power spectrum and mel scale), 3) HFC-LP (LP power spectrum and Hz scale), and 4) MFC-LP (LP power spectrum and mel scale).

In order to show the similarity between formants and SSCs, we use a sentence "Why were you away a year Roy?" spoken by a male speaker and sampled at 8000 Hz frequency. We perform frame-wise analysis of this speech utterance with frame-update of 10 ms and frame-duration of 30 ms. In order to provide an indication of formant frequencies, we compute a 10-th order LP analysis and plot the first 4 poles as a function of time in Fig. 2(a). Only those poles having bandwidth less than 600 Hz are shown in this figure. From this figure, we can clearly observe problems we usually encounter in formant extraction process due to merging of peaks and introduction of spurious peaks. Results of SSC analysis are shown in Fig. 2(b) and (c) for the HFC-FFT and MFC-FFT cases, respectively, with $M = 3$ and $\gamma = 0.5$. Comparison of Fig. 2(b) and (c) with Fig. 2(a) clearly shows similarities between formants and SSCs.

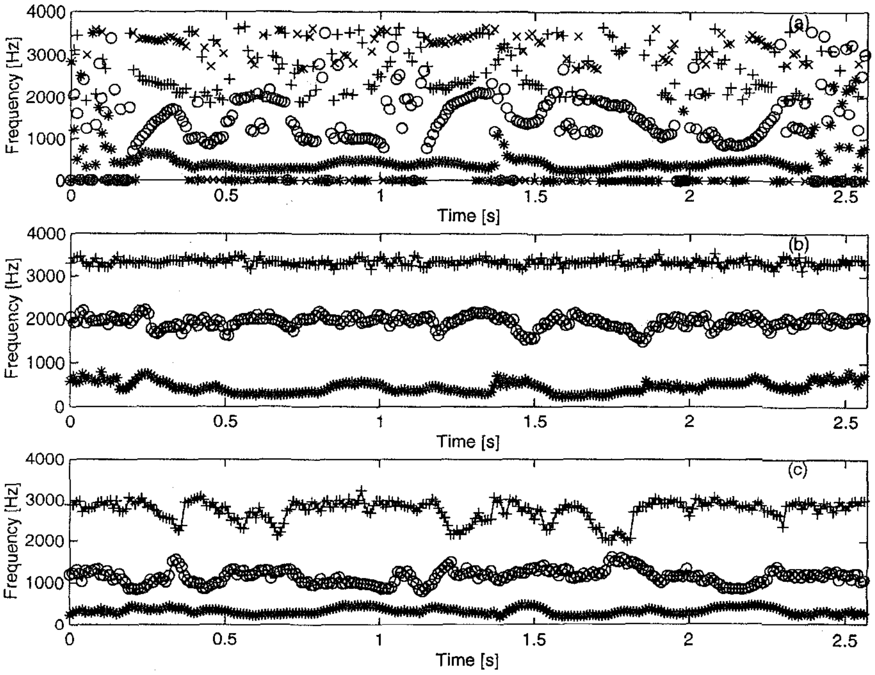

Figure 3: SSCs for the noisy utterance "Why were you away a year Roy?" (SNR = 10 dB) with $M = 3$ and $\gamma = 0.5$. (a) 4 poles from LP analysis, (b) HFC-FFT, and (b) MFC-FFT.

In order to show the effect of additive noise on SSCs, we add simulated white Gaussian noise to the utterance "Why were you away a year Roy?" and make its signal-to-noise ratio (SNR) 10 dB. Analysis results for this noisy utterance are shown in Fig. 3. We can see from Fig. 3(a) that the poles from LP analysis are greatly affected by the noise distortion. However, the SSCs shown in Fig. 3(b) and (c) are not affected that much by the noise distortion (as evident by comparing these plots with Fig. 2(b) and (c)). Also, note that the trajectories in Fig. 3(b) and (c) remain quite smooth and continuous in spite of the presence of noise. This illustrates the robustness of SSC features to noise distortion.

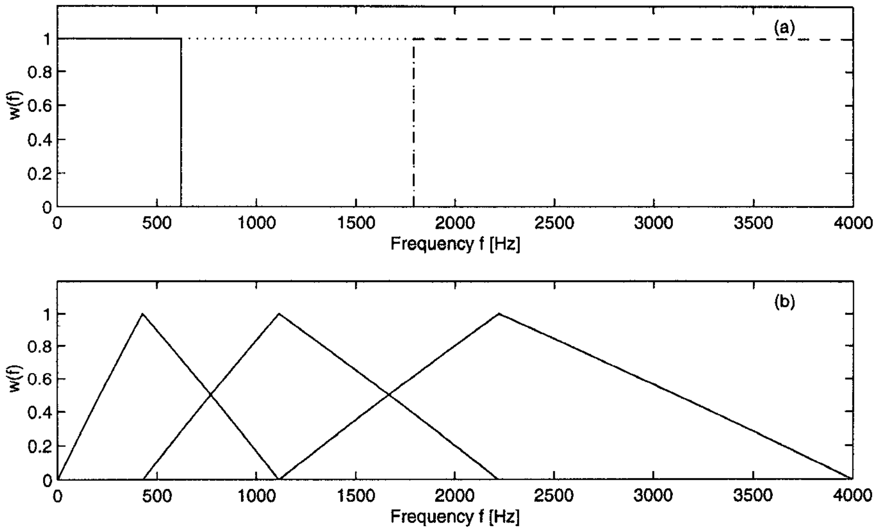

Figure 4: Subband filter shapes for computing MFCs with $M = 3$, (a) when subbands are disjoint and (b) when subbands are overlapped.

So far, we have used disjoint subbands for computing SSCs. In order to make a smooth transition from one subband to the next subband, it may be desirable to introduce some overlap between neighboring subbands. We use here a triangular shape for the overlapped subbands, while a rectangular shape has been used in the preceding paragraphs for the disjoint subbands. The filter shapes are shown in Fig. 4(a) and (b) for the disjoint and overlapped subbands, respectively. Note the amount of overlap used from Fig. 4(b). Fig. 5 shows the SSC trajectories for the utterance "Why were you away a year Roy?". Comparison of Fig. 4 with Fig. 2 shows the improvement in SSC estimation obtained by overlapping the subbands.

4 Recognition Results

We evaluate the performance of the subband spectral centroids as recognition features using a speaker-dependent HMM-based isolated-word speech recognizer as a test-bed. We use a vocabulary of 9 English e-set alphabets.

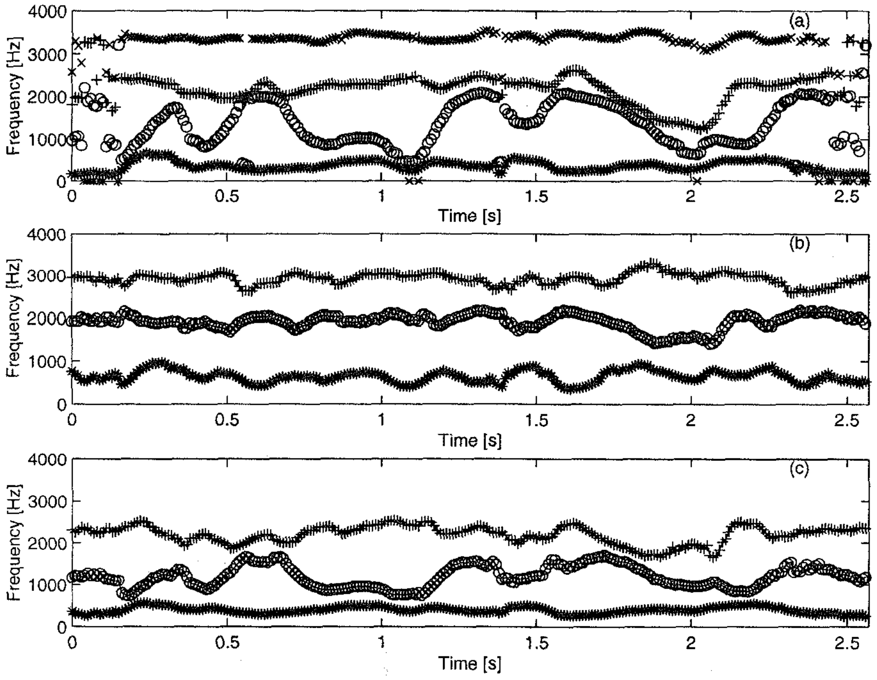

Figure 5: SSCs computed from the overlapped subbands for the utterance "Why were you away a year Roy?" with $M = 3$ and $\gamma = 0.5$. (a) 4 poles from LP analysis, (b) HFC-FFT, and (b) MFC-FFT.

The data base consists of speech from a single male talker. Sixty utterances of each word are used for training and an additional 60 utterances for testing. Speech is digitized at a sampling rate of 8 kHz. The speech signal is analyzed every 10 ms with a frame width of 30 ms (with Hamming window and preemphasis). Endpoints of each utterance are manually determined.

Here we use 3 HFC-LP type of SSCs computed with $\gamma = 0.5$ and 10 LP derived cepstral coefficients. For LP analysis, we use a linear predictor of order 10 (with Hamming window and preemphasis). We list here results for the close condition (training and test data sets are identical) as well as for the open condition (test data set is different from training data set). We can see that just 3 subband centroids can provide a good recognition performance. When they are used as supplementary features to the cepstral features, they improve the recognition performance significantly.

Table 1: Speech recognition results.

Recognition features	Recognition accuracy (in %)	
	close	open
10 cepstrum	90.6	84.3
3 centroid	90.2	84.1
10 cepstrum + 3 centroid	94.5	90.8

5 Conclusions

In this paper, spectral subband centroids (SSCs) are proposed as features for speech recognition. It is shown that these features have properties similar to formant frequencies and are quite robust to noise. When these features are used as a supplement to cepstral features, we have shown that recognition performance improves. This indicates that the SSC features provide additional information (not captured by cepstral features) for speech recognition.

References

[1] J.W. Picone, "Signal Modeling techniques in speech recognition", *Proc. IEEE*, Vol. 81, No. 9, Sept. 1993.

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

[3] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, New York: John Wiley, 1973.

[4] K.K. Paliwal, "Dimensionality reduction of the enhanced feature set for the HMM-based speech recognizer", Digital Signal Processing, Vol. 2, No. 3, pp. 157-173, July 1992.