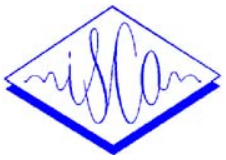

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

A MULTIRESOLUTIONALLY ORIENTED APPROACH FOR DETERMINATION OF CEPSTRAL FEATURES IN SPEECH RECOGNITION*

S. Dobrišek & F. Mihelič & N. Pavešić
e-mail: simond@fe.uni-lj.si
University of Ljubljana, Faculty of Electrical Engineering
Tržaška cesta 25, SI-1000 Ljubljana
SLOVENIA

ABSTRACT

This paper presents an effort to provide a more efficient speech signal representation, which aims to be incorporated into an automatic speech recognition system. Modified cepstral coefficients, derived from a multiresolution auditory spectrum are proposed. The multiresolution spectrum was obtained using sliding single point discrete Fourier transformations. It is shown that the obtained spectrum values are similar to the results of a nonuniform filtering operation. The presented cepstral features are evaluated by introducing them into a simple phone recognition system.

Keywords: Speech Signal Features, Multiresolution Auditory Spectrum, Cepstral Coefficients

1. INTRODUCTION

Speech processing for speech recognition is a perceptual signal analysis. Its goal is to identify a relatively small number of perceptually significant speech signal features. In general, such features are of finite extent in time and there may be several in any given time interval. Conventional feature extraction methods, used within the "state of the art" speech recognition systems are based on the short-term features in conjunction with dynamic features [3, 5]. All these features, merged in a feature vector, are usually of the same extent in time.

It is known that speech signals exhibit many non-stationary phenomena which are reflected in some local properties of a signal. Using only a fixed-window signal analysis, these local properties are poorly described. This is the reason, why multiresolution signal analysis was introduced [10]. Wavelet transforms have become well known as useful multiresolution tools for analysis of signals [4, **?**]. Another successful tool for multiresolution analysis, which has also inspired our research, is the multiresolution Fourier transform [10]. These methods have been successfully used for many signal processing applications. However, in the speech recognition domain both transforms are still being explored to develop a better speech signal representation [1].

We decided to investigate the multiresolution concept and to try to incorporate it into the procedure of deriving the well known cepstral features, which are widely used within successful speech recognition systems [11].

In the following sections, we present an approach for determination of the multiresolution auditory spectrum, the cepstral features derived from this spectrum, and finally, we evaluate the presented features through results of a simple phone recognition task.

*This work was partly funded by the Commission of the European Community under COP-94 contract No 01634 (SQEL)

2. MULTIRESOLUTION AUDITORY SPECTRUM

One of the basic signal representations is its spectrum. An important consideration to be taken here is the equivalence between a spectrum measurement and the output of a filter (for a single spectral point) or a bank of filters (for multiple spectral points) [8]. Consequently, we can notice that the Discrete Fourier Transformation (DFT) represents spectrum measurements for the equally spaced spectral points and according to the above consideration it actually equals the output of a bank of uniform filters.

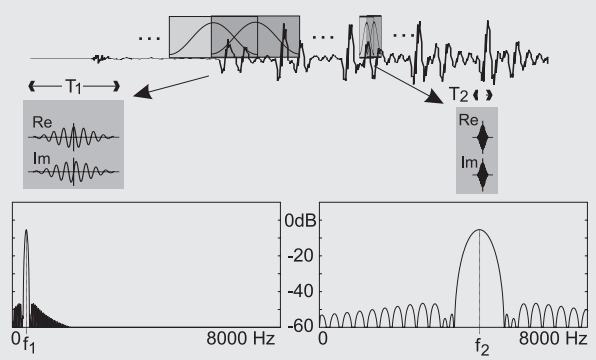

Figure 1: *Single spectral point measurements and magnitude responses of the corresponding filters.*

Spectral measurement is usually evaluated over a certain number of signal samples. If N_k denotes the number of samples and t_s denotes the sampling interval then the spectral measurement $F_i(\omega_k)$ of a sampled signal $f(n) = f(nt_s)$ at position it_s can be defined according to the following equation.

$$F_i(\omega_k) \doteq t_s \sum_{n=i-N_k/2}^{i+N_k/2-1} f(n) e^{-j\omega_k(n-i)t_s} \qquad (1)$$

As mentioned above this spectral measurement corresponds to the output of a particular filter. It can be shown that by varying the number N_k the effective bandwidth of the filter is changed. The filter shape may be also altered by introducing a window, $w(n) = w(nt_s)$, that multiplies

each term in the selected portion of the signal.

$$\widehat{F}_i(\omega_k) \doteq t_s \sum_{n=i-N_k/2}^{i+N_k/2-1} w(n) f(n) e^{-j\omega_k (n-i) t_s} \qquad (2)$$

Figure 1 illustrates single spectral point measurements and the magnitude responses of the corresponding filters. In this example the measurements are done using the Hamming window function.

The speech signal spectrum constantly changes with time. In general, spectral measurements should be repeated for each successive signal sample. This type of measurement is so called sliding or running spectral measurement and is computationally inefficient. However, the efficiency can be improved using the standard method that allows computation of running spectra approximately by hopping rather than by sliding analysis window. The hopped mea-

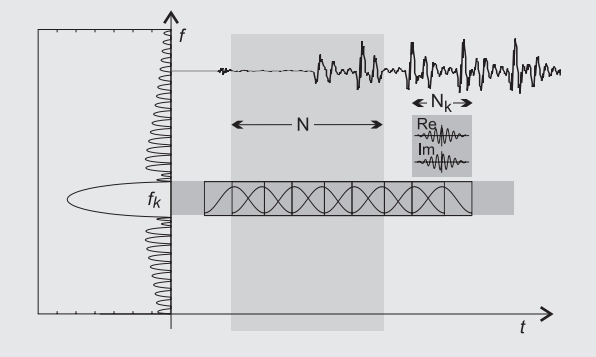

Figure 2: *Calculations of hopped spectral measurements.*

surement is simply a sampling of sliding measurements. To reduce the effect of folding the analysis windows have to overlap and the 2:1 overlap seems to be a reasonable choice [8]. Figure 2 shows how the successive single point spectral measurements are actually computed. It can be seen that a complete calculation for all measurements within a frame of N samples requires $2N$ complex summations of products.

2.1. The Auditory Spectrum

The auditory spectrum is "auditory" in the sense that it has a nonuniform frequency resolution, and that the resolution is defined according to some parameters of the human auditory system [2, 12]. The nonuniform frequency resolution is usually obtained using nonuniform filterbanks. Such filterbanks are defined within the computational models of the human auditory system [9]. It can be also implemented using the Wavelet transforms [7]. However, these two approaches have some disadvantages. The first one requires heavy computational load and the second one becomes complicated when an arbitrary time-frequency resolution is required.

As mentioned before, the conventional DFT represents spectrum measurements for the equally spaced spectral points and it corresponds to the output of a bank of uniform filters. The nonuniform filterbank can be obtained by implementing a large uniform filterbank and then the nonuniformity is created by combining two or more subsequent uniform channels. These combinations can take many forms. The most popular is the triangular weighted sum of subsequent uniform channels. This approach is widely used for determination of speech signal features due to the Fast Fourier Transformation (FFT) algorithm, which is used for the DFT calculation and is computationally very efficient.

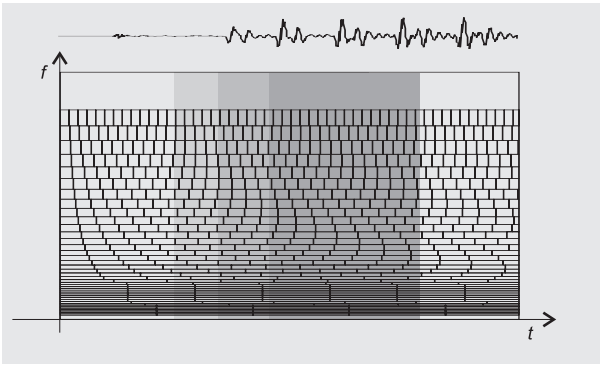

Figure 3: *The nonuniform time-frequency resolution obtained by a set of single point spectral measurements.*

Unfortunately, the above approach has also a disadvantage. The problem is in the fixed time resolution of the FFT. The auditory spectrum should have not only nonuniform frequency resolution but also nonuniform time resolution [12].

We decided to fulfil this requirement by using the already discussed single point spectral measurements. We carefully designed each measurement and the corresponding filter. As described before, each filter can have its own effective bandwidth and each bandwidth represents the duration of a corresponding analysis window. Consequently, a nonuniform frequency resolution and a nonuniform time resolution are achieved simultaneously.

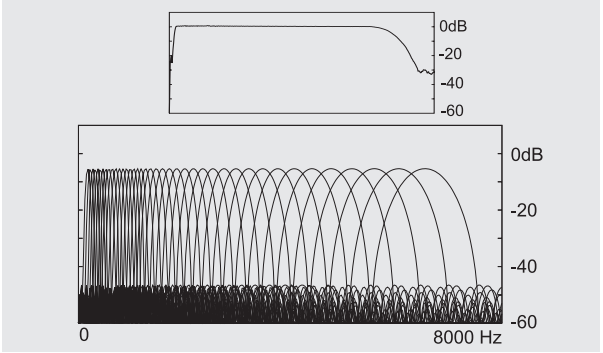

Figure 4: *The magnitude response of the filterbank and its composite magnitude response.*

Figure 3 illustrates an example of a nonuniform time-frequency resolution obtained by a set of 40 single point spectral measurements, which are based on the Hamming

window function. In this particular example, the spectral points are spaced according to the BARK scale [12]. In Figure 4 the magnitude response of the filterbank and its composite magnitude response are depicted.

The actual auditory spectrum we propose represents log-power values of the discussed spectral measurements. Figure 5 shows an example of the multiresolution auditory spectrum of a speech signal.

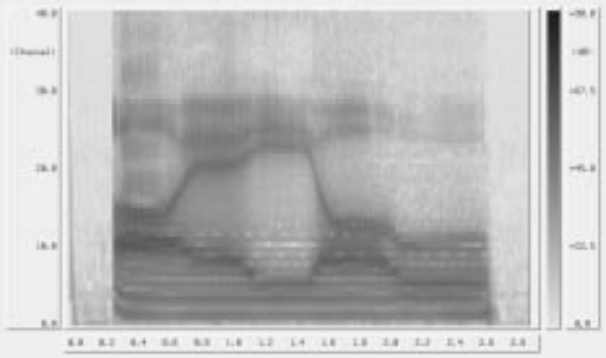

Figure 5: *The multiresolution auditory spectrum of a speech signal.*

3. CEPSTRAL FEATURES

A frame of the auditory spectrum can be used as a feature set for speech recognition. However, these numerous features are expected to be highly correlated. The additional discrete cosine transform (DCT) applied to the log-power spectrum is usually used to reduce the number of features and, furthermore, the derived cepstral ceofficients are expected to be more uncorrelated.

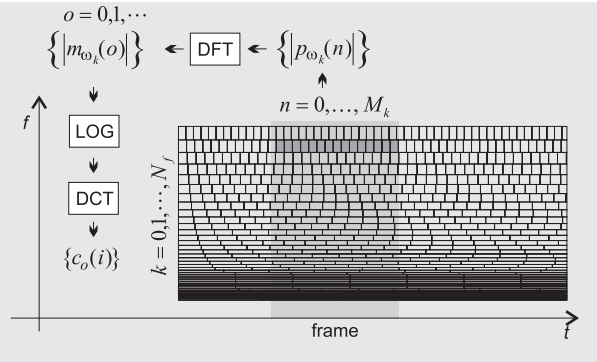

Figure 6: *The cepstral coefficients derived from the multiresolution spectrum.*

The conventional cepstrum coefficients are derived by taking the DCT of the log-power of the filterbank output. If the filterbank output is obtained using the DFT then the coefficients are derived by taking the DCT of the short time log-power spectrum. The question is, how the DCT can be applied to the multiresolution spectrum.

The multiresolution auditory spectrum has higher time resolution for high frequency spectral measurements and lower time resolution for low frequency spectral measurements. This means that there are more subsequent measurements for high frequencies than for low frequencies within a period of time. The conventional spectrum values approximately equal average values of the subsequent measurements within a frame. However, this averaging eliminates the most important feature of the auditory spectrum, which is its ability to describe some local properties of a speech signal.

On the other hand, the average value can be treated as a zeroth order coefficient of the additional DFT applied to the subsequent measurements, and there is no theoretical constraint to introduce the additional higher order coefficients, which are expected to describe local properties of a signal. In general, this DFT coefficients are complex and for speech signals it is reasonable to take only the amplitude information.

Figure 6 illustrates how the cepstral coefficients are calculated from the multiresolution auditory spectrum. In Figure 6 M_k denotes the number of subsequent single point spectral measurements, $|p_{\omega_k}(n)|$, within a speech signal frame, and $|m_{\omega_k}(o)|$ denotes the absolute value of the o-th order coefficient of the DFT applied to these measurements.

Finally, the DCT of the obtained absolute values gives the required cepstral coefficients $c_o(i)$. It can be seen from the equations in Figure 6 that the zeroth order ($o = 0$) cepstral coefficients are approximately equivalent to the conventional cepstral coefficients and that the higher order ($o > 0$) cepstral coefficients represent local properties of the signal frame, especially for high frequency measurements.

4. PHONE RECOGNITION EXAMPLE

For the very first evaluation of the presented features we used a simple phone recogniser, which is based on the Gaussian mixture probability density functions (pdf). The pdf parameters were initialised using the K-means algorithm and estimated using the EM-algorithm. The speech database, used for model parameters estimation, consists of 1512 utterances of 6 speakers (3 males and 3 females). The database corpus consists of 252 unique phonetically balanced words and short phrases. Speech signals are labelled by 33 phone classes. Each vowel class was modelled using 5 component densities and all the other classes were modelled using 3 component densities.

Recognition results of the phone recogniser were compared for different types of features. The conventional mel-frequency cepstral features (MFCC) were derived using the Hamming window, 1024 point FFT and 40 mel-scaled triangular filters. The cepstral feature vectors were generated with the frame shift of 8 ms and the frame duration of 24 ms.

The multiresolution auditory spectrum had 40 mel-scaled spectral points. The spectral measurements were derived from the hopped Hamming windows with a 2:1 overlap. The actual selection of multiresolution cepstral features

(MRCC), derived from the multiresolution spectrum, was defined experimentally. Due to space limitations we can present only the most important results. Table 1 shows the recognition results for feature vectors with 12 features and table 2 the results for feature vectors with 24 features. Table 1 contains recognition results for the MFCC features

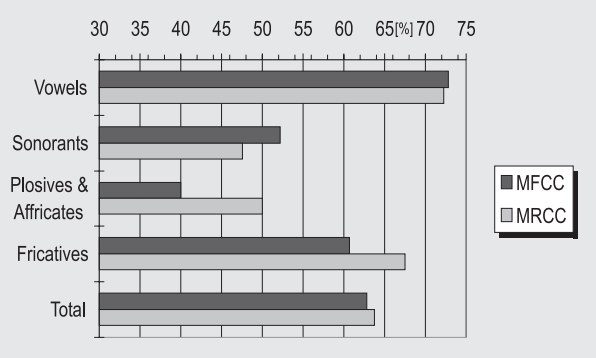

Table 1: *Recognition results for 12 features.*

with the frame duration of 24 ms, and for the MRCC features with the frame duration of 32 ms. The MFCC feature vectors consist of log power and the first 11 cepstral coefficients. The MRCC consist of log power, 8 zeroth order ($o = 0$) multiresolution coefficients ($i = 4, \cdots, 11$) and 3 first order ($o = 1$) multiresolution coefficients ($i = 1, 2, 3$).

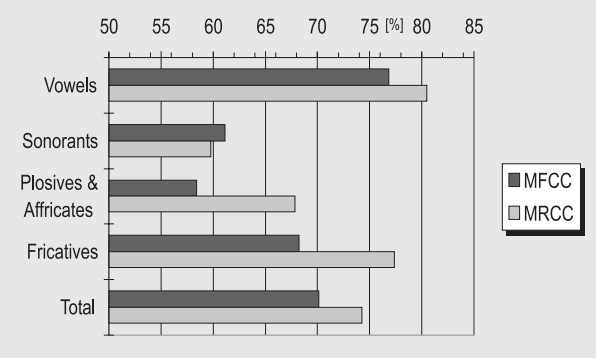

Table 2: *Recognition results for 24 features.*

Table 2 shows the recognition results for the MFCC features with the same frame duration of 24 ms, and for the MRCC features with the frame duration of 64 ms. The MFCC feature vectors consist of log power, first 11 cepstral coefficients, and their first derivatives. The MRCC vectors consist of log power, 11 zeroth order ($o = 0$) multiresolution coefficients ($i = 3, \cdots, 13$), 8 first order ($o = 1$) multiresolution coefficients ($i = 1, \cdots, 8$), and 4 second order ($o = 1$) multiresolution coefficients ($i = 1, \cdots, 4$).

The above results show the main advantage of the presented cepstral features, which lies in their ability to describe local properties of speech signals. When we use the conventional cepstral features then stretching of a signal frame improves the recognition rate for vowels but reduces recognition rate for other more instantaneous phonemes. On the other hand, the feature vectors composed from the presented cepstral coefficient improve the recognition results for all groups of phonemes when a signal frame is stretched.

5. CONCLUSIONS

Cepstral features derived from the multiresolution auditory spectrum have been proposed. Presented features are an extension of the conventional cepstral coefficients and have proved to be very promising for describing local properties of speech signals. Simple phone recognition tests demonstrated that some experimentally defined selections of the presented features outperform the conventional cepstrum.

REFERENCES

[1] E. Ambikairajah, M. Keane, L. Kilmartin, G. Tattersall, "The Application on the Wavelets for Speech Recognition", *Proc. of EUROSPEECH'93*, Vol. 1, pp. 151 - 154, Berlin, Germany, 1993.

[2] J. B. Allen, "How Do Humans Process and Recognize Speech", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 4, pp. 567 - 576, October, 1994.

[3] H. Bourlard, H. Hermansky, N. Morgan, "Towards Increasing Speech Recognition Error Rates", *Speech Communication*, No. 18, pp. 205 - 231, 1996.

[4] A.Cohen, R. D. Ryan, *Wavelets and Multiscale Signal Processing*, Chapman & Hall, Paris, 1995.

[5] R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, V. Zue, *Survey of the State of the Art in Human Language Technology*, Center of Spoken Language Understanding, Oregon Graduate Institute, November, 1995.

[6] I. Daubechies, "The Wavelet Transforms, Time-frequency localisation and Signal Analysis", *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 38, No. 2, pp. 674 - 690, March, 1992.

[7] U. K. Laine, "Speech Analysis Using Complex Orthogonal Auditory Transform (COAT)", *Proc. of the ICSLP'92*, Vol. 1, pp. 69 - 72, Alberta, Canada, 1992.

[8] L. R. Rabiner, *Theory and applications of digital signal processing*, Prentice-Hall, London, England, 1975.

[9] S. Seneff, "A Computational Model for the Peripheral Auditory System: Application to Speech Recognition Research", *Proc. of the ICASSP'86*, Vol. 3, pp. 1983 - 1986, Tokyo, Japan, 1986.

[10] R. Wilson, A. D. Calway, and E. R. S. Pearson, "A generalized Wavelet Transform for Fourier Analysis: The Multiresolution Fourier Transform and Its Application to Image and Audio Signal Analysis", *IEEE Trans. on Information Theory*, Vol. 38, No. 2, pp. 674 - 690, March, 1992.

[11] S. Young and G. Bloothrooft (eds.), *Corpus-Based Methods in Language and Speech Processing*, Kluwer Academic Publishers, The Netherlands, 1997.

[12] E. Zwicker, H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, Berlin - Heidelberg, Germany, 1990.