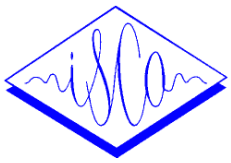

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

USING WAVELET DYADIC GRIDS AND NEURAL NETWORKS FOR SPEECH RECOGNITION

Richard F. Favero and Fikret Gurgen*
Speech Technology Research Group
Department of Electrical Engineering
University of Sydney, Australia

Abstract

This paper describes two multi-rate feature vectors derived from the wavelet transform coefficients for speech recognition. The feature vectors ensure time and frequency alignment across the dyadic grid. The first strategy to compose the feature vector is based on grouping by location in time. This produces frame synchronous data that can be applied to a recogniser without the addition of interpolated points on the dyadic grid. The second strategy is to compose groups of vectors according to frequency region and the sampling rate of each region. Then, the feature vectors are applied to a window based neural network (WNN) to assess speech recognition performance. The WNNs are designed to enhance the resolution of various frequency bands to improve speech recognition performance.

Experiments are performed using the words /b,d,g/. The results show that the performance of the WNN using this wavelet based feature vector is comparable to that of the HMM based system reported in [3,4].

1 Introduction

Acoustic feature extraction is important for speech recognition and several approaches have been adopted which include LPC, and mel-scaled cepstrum coefficients[1]. Despite the acoustic feature extraction being an independent module in a speech recognition system, the overall recognition performance of the system is determined by the quality of the feature extraction and its match to the classifier.

The wavelet transform has become a popular tool for image and speech processing. It has been used in speech processing for analysis [9] and pitch detection [8] with much success. The wavelet transform has recently been applied to speech recognition[3,4].

The wavelet transform of a speech signal generates coefficients on a dyadic grid. This grid has fine time resolution at the high frequencies and fine frequency resolution at low frequencies. This is advantageous for speech recognition since it improves frequency resolution in the formant region and it also models the plosive bursts contained in high frequency features.

The experiments outlined in [3,4] used the wavelet transform for speech recognition with Hidden Markov Models (HMMs) and an improvement in recognition performance was observed compared with MFCC. The coefficients generated by the WT on the dyadic grid were interpolated to produce frame synchronous data. This increased the amount of data without increasing the information content.

Recent studies [5,6] have demonstrated that feed-forward neural networks (NN) with a window based structure, can perform very well for small but difficult pre-segmented phonemic discrimination tasks. Furthermore it has been shown that extension of these networks for the design of speech recognition systems can be easily realized [6].

This paper describes techniques for grouping data on a dyadic grid so that the wavelet based coefficients can be used by the WNN architecture. The results are compared with those obtained with an HMM. Section 2 introduces a brief wavelet theory and section 3 describes the generation of multi-rate feature vectors. Section 4 discusses WNN architecture and section 5 presents the results and the conclusions of the work.

2 Wavelet Theory

Wavelet theory is based on generating a set of wavelets by dilation and translation of a generating wavelet. All of the wavelets are scaled versions of the "mother wavelet". This means that only one filter needs to be designed and the others will follow the scaling rules in both the time and frequency domain.

A set of wavelets is generated from the mother wavelet $\Psi(t)$ [2] by:

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right)$$

The wavelets are contracted ($a < 1$) or dilated ($a > 1$) and are moved over the signal to be analysed by time step b (which is real valued). Contraction and dilation scale the frequency response to allow the set of wavelets to span the desired frequency range. The set of wavelets can be considered as a filter bank for speech analysis.

The continuous wavelet transform (CWT) is defined as:

$$CWT(b,a) = \frac{1}{\sqrt{a}}\int s(t)\,\Psi\left(\frac{t-b}{a}\right)dt$$

*Fikret Gurgen is on leave from Comp. Eng. Dept. of Bogazici University.

The discrete wavelet transform (DWT) is given by:

$$DWT(a^i, a^i n) = \frac{1}{\sqrt{a^i}} \sum_k \Psi\left(\frac{k}{a^i} - n\right) s(k)$$

where i is an integer. The DWT computes data points on a dyadic grid if $a = 2$. (A dyadic grid has half of the number of data points at each successive lower octave[2] (see figure 1)).

By restricting the values of i to be integers and $a = 2$, each of the wavelets will be an octave space apart. Choosing other values for a will change the number of wavelets that are required to cover a given frequency range. Thus if the initial generating wavelet is defined appropriately then sub-octave resolution can be accommodated. The scaling within each octave will need to be preserved so that a given number of voices appears within each octave. This can be varied by changing the value of a or by choosing i to be a real fraction of the number of voices in each octave[2].

3 Multi-rate feature vectors of wavelets

Figure 1 shows the dyadic grid. Each dot represents a coefficient generated by the inner product of one of the wavelets with the incoming speech. From 4000Hz to 2000Hz there are 6 voices and the same from 2000Hz to 1000Hz. From 1000Hz to 80Hz there are 6 voices since we are using a mel-scale for the coefficient generation.

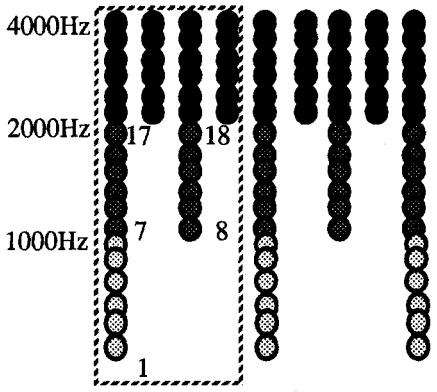

Figure 1:Dyadic Grid Spacing for TLBG.

3.1 Time Location Based Grouping (TLBG)

The dyadic grid is difficult to use because synchronisation of the coefficients presented to the recogniser is crucial. If we consider the dotted square that is drawn on figure 1, we can construct a vector that is composed of all the frequency components of this time segment. This square captures the broad features of the signal in the low frequency coefficients and the fine details in the high frequency coefficients.

The coefficients are grouped as shown in figure 2. The dotted square is moved over the entire dyadic grid to produce a set of frame synchronous vectors for a speech sample. This process retains the time and frequency localisation and does not add interpolated coefficients as used in [3,4]. The new vector will have 42 coefficients.

Figure 2:Dyadic Grid Spacing for FRBG

3.2 Frequency Region Based Grouping (FRBG, Zooming)

This strategy can be considered as a focusing mechanism. It is motivated by mechanisms in vision for recognising objects.

Consider viewing an object at a distance. The broad outline of the object can be determined and used to eliminate unlikely objects. As the distance of the object decreases the detail improves and a correct identification can be performed.

The data in a dyadic grid is of a similar form. The low frequency data describes the broad outline of the speech signal and can be used to eliminate unlikely candidates. Increasing details (using higher frequency detail) can improve the identification.

The data are grouped into the regions of high, mid and low. frequency. These frequency bands are given to the recogniser as independent data. The number of low frequency vectors will be half that of the mid range which will be half of the high range region. To synchronise the data, windows must move over the data at different speeds or the windows must be different sizes.

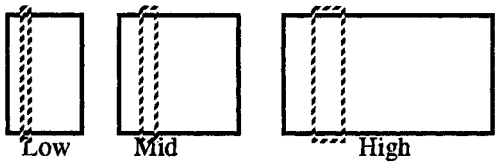

Figure 3:The FRBG grouping with shifting windows

4 Window-based Neural Networks

For efficient classification of speech signals using wavelet parameters, a windowed feed-forward NN (WNN) architecture with a supervised training algorithm is used. Windows consisting of delayed units of frames (Figure 3) are shifted over the multi-resolution dyadic frequency grid of the wavelet data. This allows the short duration features of fricatives and plosive bursts to be learned by the network. As a result, the actual features at each level of resolution

are extracted by using a time-invariant architecture.

A four layer WNN (Figure 4) is employed as an elementary architecture. The number of windows is changed in the input layer according to whether the TLBG or FRBG are used. The TLGB uses one window (with 3 shifts) and covers all the frequencies, FRBG uses three windows for the parameters of three frequency regions. The FRBG window sizes and translation speed synchronously cover the contents of the same data and transfers it to the first hidden layer. In both groupings, the information from input layer is processed by another window (5 frames) which extracts the higher level features using a shift-size of 2 frames. The resulting information is combined at the second hidden layer. At the output layer, a tied connection is used between 3 output units and all of the second hidden layer. All windows are tied connected to the succeeding layer.

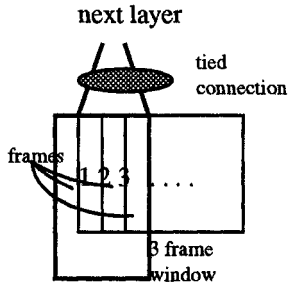

Figure 4: Window structure

The basic unit of the network evaluates the weighted sum of its inputs through a sigmoid function which is a continuous non-linear function. In this case, each sigmoid function in the unit receives the weighted sum of input values from delayed units.

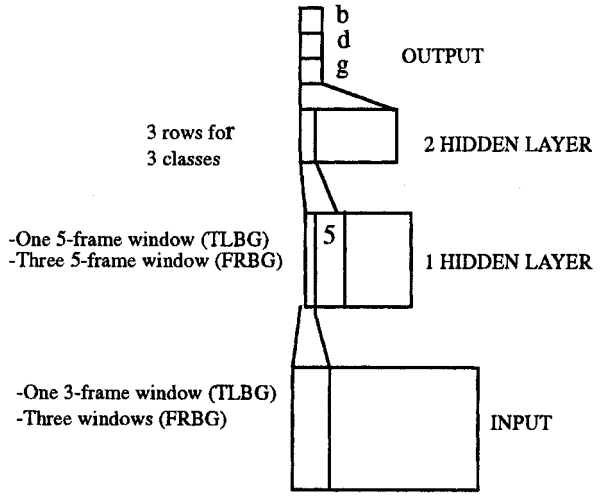

Figure 5: Window-based Neural Network Architecture

A back propagation based recognition algorithm is used to train the networks. The training strategy gradually increases the number of samples used for training . The performance of the algorithm improves by using variable momentum and learning rates during the computation of the new weight values. Convergence occurs after 500 iterations and a maximum performance is obtained around iteration 1000-1500. There is a small difference between the maximum performances and the convergence performance.

5 Experiments

We used the NIST TI-46 word database and restricted ourselves to /b, d, g/ (a subset of the English E set) of the alphabet. The database is down-sampled to 8000Hz. The data is end point detected and then the lengths are normalised to 5000 samples by the addition of sampled noise. This is done to create speech samples of a fixed length for use with the WNN. There are 478 training utterances and 764 testing utterances.

The generating wavelet is based on the Hanning window. This window has a sufficiently narrow main lobe (centred about zero frequency) and the first lobe is more than 40dB below the main lobe [2]. The generating wavelet is designed to be the highest frequency wavelet filter which is modulated to 4000Hz.

From [3,4] the maximum recognition performance is obtained with the generating wavelet with a length of 32 samples and a 16 sample time advance for data sampled at 12500Hz. These same parameters have been chosen for this experiment since the frequency range to be covered is from 80Hz to 4000Hz. The perceptually based mel-scale is used to dilate the wavelets over the desired frequency range.

The experiments were conducted with the WNN and comparisons are made with a HMM based system where appropriate. The HMM used was a CDHMM with 5 emitting states and 5 mixtures per state. It was trained with the same data but with the dyadic grid filled with redundant data points. This has meant that the data presented to the HMM had twice the data but the same information content.

6 Results and Discussion

The results obtained with the time based grouping are shown in Table 1 and those obtained with the frequency

region based grouping in Table 2

Table 1: Results for b,d,g with time based grouping

	WNN	HMM
Training	99.1%	82.4%
Testing	66.9%	68.0%

Table 2: Results for b,d,g experiment with frequency region grouping

	WNN	HMM
Training	99.8%	82.4%
Testing	71.5%	68.0%

The results indicate that the new vector based on the time grouping criteria has retained the information content. The coefficients located on the dyadic grid are sufficient to parameterise the speech signal and to capture the necessary detail to perform the recognition task. The grouping of the high and low frequency detail from a small but longer time frame of the dotted square on Figure 1 has shown that packets of coefficients can be used for speech recognition.

This concatenated vector provides a packet of information over a small global time. In this vector the broad view of the signal is captured by the low frequency signals. They highlight the slow changing parts of the signal. In the same packet the small details are highlighted by the high frequency data. As we move along the vector we are zooming in on specific details of the signal. The global details and the finer details of a segment of speech are being captured at different resolutions. This process of analysis is used widely in image analysis schemes.

The results using the frequency region grouping give an insight into the way that recognition could be performed and is intuitively a human process. The results obtained are the best reported with either the HMM or WNN. The neural network was able to be configured to match the data representation. The nature of the zooming in on the specific speech features is observed in the recognition result.

The neural networks offer the flexibility to deal with data in many and varied formats. This offers several possibilities as to how different feature vectors can be created that can exploit the nature of a particular NN architecture.

The architecture and the notion of multi-rate systems for speech recognition fit well with the understanding of the auditory system. The firing of the auditory nerve is proportional to the frequency content of the signal. For instance low frequency signals will cause synchronous firing of low frequency nerves and high frequency signals will cause high frequency firing. The new feature vectors offer these features in each packet of data and the neural network architecture can take advantage of these features.

While the network that has been used is large and complex, the surfaces of discrimination have been determined at least as well as the HMM with the filled dyadic grid. The convergence time for the neural networks are in the order of that of the HMM. A further reduction in data points and several iterations of the NN architecture and training algorithm will increase the recognition performance and reduce the training time.

This work provides a technique for speech recognition systems that wish to use multi-rate data. Future work will investigate the techniques to apply these feature vectors to HMM systems and to provide WNN architectures that improve the match to the acoustic feature extraction.

7 Acknowledgements

Richard Favero is supported by an Australian Postgraduate Research Award and a Telecom Postgraduate Fellowship.

8 References

[1]Furui, S. "Digital Signal Processing, Synthesis and Recognition", Marcel Decker 1989

[2]Daubechies I, "Ten Lectures on Wavelets", Philadelphia 1992

[3]Favero R.F, King R.W, "Wavelet Parameterization for Speech Recognition" ICSPAT Vol 2 pp. 1444-1449 1993.

[4]Favero R.F, King R.W "Wavelet Parameterization for Speech Recognition: Variations in scale and transaction parameters" to be published at ISSIPNN94

[5]Gurgen F S, Aikawa K, Shikano K, "Phoneme recognition with neural networks using a novel fuzzy training algorithm," IEEE IJCNN'91 Singapore, pp 572-577, 1991.

[6]Gurgen F S, "Phoneme recognition neural networks," ISCIS VII, pp 569-572, 1992.

[7]Waibel A, Hanazawa T, Hinton G, Shikano K, Lang K J, "Phoneme recognition using Time-delay neural networks", IEEE Trans on ASSP Vol 37, No 3, 1989.

[8]Kadambe S,. Boudreaux-Bartels G.F, "Application of the wavelet transform for pitch detection of speech signals" IEEE Trans. Info. Theory Vol.38, pp917-24, 1992.

[9]Kronland-Martinet R, "The wavelet transform for analysis,synthesis,and processing of speech & music sounds" Computer Music Journal Vol.12 No.4 pp.11-20, 1988

[10]Rioul O, Vetterli M, "Wavelets and Signal Processing" IEEE Signal Processing, pp. 14-38, October 1991