COMPARISON OF MOTHER WAVELETS FOR SPEECH RECOGNITION

Richard F. Favero

Speech Technology Research Group
Department of Electrical Engineering
University of Sydney, Australia

ABSTRACT - This paper compares four modulated wavelets for speech recognition. The modulated wavelets are based on well understood window functions. These are the Hanning, Hamming, Blackman and Gaussian windows.The wavelet parameterisation of speech using each of the above mother wavelets is applied to a multi-speaker E-set discrimination task. The results show that the Gaussian and the Blackman windows give a slight improvement in recognition performance.

INTRODUCTION

Wavelets have been shown to be useful front end processors for speech recognition systems for discriminative tasks. These speech recognition systems have been based on Hidden Markov Models (HMMs) (Favero and King, 1994) and neural networks (Favero and Gurgen, 1994; Kadambe and Srinivasan, 1994; Szu et. al. 1992). Discriminative tasks have been chosen for use with wavelet parameterisations to show that improved time and frequency resolution will better parameterise these difficult areas of speech for speech recognisers.

The choice of mother wavelet can determine:

- the number of wavelets to parameterise the signal,
- the sampling rate of the wavelet transform,
- the types of features that are to be extracted.

The modulated Gaussian was used by Morlet (1982) in his work on Geophysics. Modulated wavelets have been used in speech analysis (Basile et. al., 1992) and music analysis (Kronland-Martinet, 1989). Application of a modulated Hanning window to speech recognition was successful in improving speech recognition performance over MFCC.

This paper reports the results of comparisons of four modulated wavelets when applied to a discriminative speech recognition task. The four window functions that the wavelets are based on are outlined and how the wavelets are produced.

WAVELET THEORY

Wavelet theory is based on generating a set of filters by dilation and translation of a generating wavelet (mother wavelet). The mother wavelet is usually a band-pass filter. All of the generated wavelets are scaled versions of the "mother wavelet". Increasing the scale of a wavelet will increase its time duration, reduce the bandwidth and shift the centre frequency to a lower frequency value. Decreasing the scale does the opposite.

A set of wavelets is generated from any defined mother wavelet $\Psi(t)$ by:

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

The wavelets are contracted ($0<a<1$) or dilated ($a>1$) and are moved over the signal to be analysed by time step b. Contractions and dilations scale the frequency response of the generating wavelet to produce a set of wavelets that span the desired frequency range. The generated set of wavelets can be considered as a filter bank for speech analysis.

The continuous wavelet transform (CWT) performs the inner product (correlation) of a signal s(t) with all scales and dilations of a mother wavelet. The CWT will produce a two dimensional output similar to a spectrogram. The CWT is defined as ($a > 0$, b is real):

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

The discrete wavelet transform (DWT) is the CWT sampled at a defined set of points. The DWT of a sampled signal s(k) is given by (i, k are indexing integers):

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

The scaling value is made discrete by i being discrete. The DWT computes data points an octave space apart on a dyadic grid if $a = 2$ since the scale values would be -2, -1, 0, 1, 2, 4, 8.... (A dyadic grid has half of the number of data points at each successive lower octave (Daubechies, 1992; Rioul and Vetterli, 1991). The value of a can be chosen such that more than one wavelet coefficient per octave is generated (voices of an octave). If the initial generating wavelet is defined appropriately then sub-octave resolution can be accommodated. This can be achieved by choosing:

$$a = 2^{\left(\frac{1}{numberOfVoices}\right)}$$

The sampled CWT (SCWT) is a variation of the DWT. This produces frame synchronous data (redundant at lower frequencies) but retains the features that are offered by the wavelet transform. The sampled CWT is given by:

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

The wavelet transform in this paper is calculated using the SCWT. The SCWT is modified to reduce the computational load. Coefficients off the dyadic grid are filled with adjacent coefficients that lie on the dyadic grid.

A piece-wise mel scale is used as the frequency scaling operation. This scale is logarithmic above 1000Hz and linear below 1000Hz. There are 12 wavelets above 1000Hz and 6 below 1000Hz. Thus the wavelet transform generates 18 coefficients per sample, every 2ms.

MODULATED WAVELETS

Morlet (1982) used the modulated gaussian in his work in geophysics. Since then, modulated window functions have become popular wavelets because they have desirable time-bandwidth products. This ensures the best possible time and frequency resolution for a given task.

Modulation of window functions allows the control of the centre frequency and bandwidth of the wavelet. The four window functions examined here have been studied extensively and their frequency responses are well known. They have suitable frequency responses and time duration properties for the present application. Different modulated window functions may affect recognition performance due to the different frequency and time responses for the data in a particular task. The four window functions used are outlined below (Oppenheim and Schafer, 1975).

Hanning

The Hanning window has been used in wavelet analysis (Basile et. al. 1992) and for speech recognition in previous work (Favero and King, 1993,1994; Favero and Gurgen, 1994). The peak side lobe is -31db below the main lobe. The hanning window is given by:

$$w(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

Hamming

The Hamming window is widely used in MFCC calculations. The peak side lobe is -41dB below the main lobe.The Hamming window is given by:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

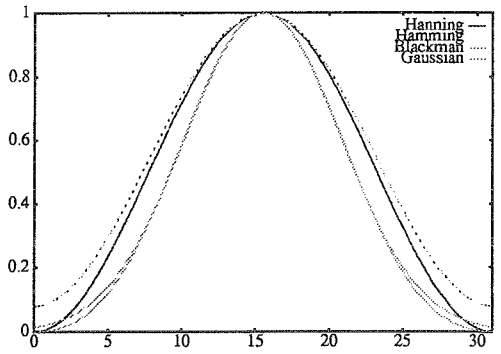

Figure 1:Plot of the window functions in the time domain

Blackman

The Blackman window is given by:

$$w(n) = 0.42 - 0.5\cos\left(\frac{2\pi n}{N-1}\right) + 0.08\cos\left(\frac{4\pi n}{N-1}\right), 0 \le n \le N-1$$

The peak side lobe is -57dB below the main lobe.

Gaussian

The Gaussian window is a widely used windowing function. The Gaussian function has the same shape in both the time and frequency domain. The Gaussian function has the best time bandwidth product of $1/4\pi$. The Gaussian function is given by:

$$w(t) = e^{-\left(\frac{t-\mu}{\delta}\right)^2}$$

Figure 1 shows that the Hanning and Hamming windows have a similar shape as do the Gaussian and the Blackman windows. Experimental results for these two pairs of windows are expected to correlate.

EXPERIMENT

A discrimination task is chosen for the experiment based on the NIST TI-46 word database. This database contains 16 speakers and each speaker repeats each word 26 times. Ten of the words are used for training and the 16 remaining are used for testing. The database is down-sampled to 8kHz. The leading and trailing silence is removed from each utterance prior to performing the wavelet transform.

We have chosen the "E-set" (b, c, d, e, g, p, t, v, z) because the difficulty in discriminating the initial consonant makes this a difficult multi-speaker recognition task.

The experiments described here use continuous density HMMs with 5 states and 5 weighted mixtures (Rabiner, 1989).

The wavelet transform is performed with each of the mother wavelets. The lengths of the mother wavelets is set to 32 samples (4ms). The gaussian standard deviation is set to 7.5 samples which produces a gaussian window near 32 samples (this depends where the cut-off is determined).

RESULTS

The results are shown in the following table.

	Hanning	Hamming	Blackman	Gaussian
Training	78.4	76.6	79.0	80.4
Testing	67.5	66.2	68.5	68.5

Table 1: Recognition performance as a percentage correct

DISCUSSION

The results show that the Blackman and Gaussian windows performed 1% better that the Hanning window and 2.3% better than the Hamming window. The improvement in performance is not significant, nor are the differences between the window functions. The results imply that changes to the mother wavelet can affect recognition performance. These small changes could be amplified if the output of the wavelet transform is passed through a post-processor (such as Principal Components Analysis).

The Blackman and Gaussian windows have a similar shape and the recognition performance corresponds with this similarity. The window shapes had a better time resolution than that of the Hanning and Hamming windows. This seems to indicate that mother wavelets that have a better time resolution should provide slight improvements in recognition performance.

Despite the Hanning and Hamming windows being similar there was 1.3% difference in recognition performance. The Hamming window has a step at the edges of the defined time domain. The frequency impacts of this may have affected the recognition performance.

While the Gaussian and the Blackman modulated wavelets performed equivalently, the Blackman window has several features that make it a valuable wavelet. These include:

- finite time duration,
- time resolution equivalent to the Gaussian.

These factors make the Blackman window particularly suitable for wavelet parameterisation.

ACKNOWLEDGEMENTS

The author is grateful to Robin W. King for many useful discussions and editorial comments. Richard Favero is supported by an Australian Postgraduate Research Award and a Telecom Postgraduate Fellowship.

REFERENCES

Basile, P., Cutugno, F. and Maturi, P. "The Wavelet Transform and Phonetic Analysis", Proc. of the Fourth Australian Intl. Conf. on Speech Science and Technology, Brisbane, pp 2-7.1992

Daubechies,I. "Ten Lectures on Wavelets", Philadelphia, 1992.

Favero, R.F, King R.W, "Wavelet Parameterization for Speech Recognition" Int. Conf. Signal Processing Applications and Technology, Santa Clara, Vol 2 pp. 1444-1449 1993.

Favero, R.F, King, R.W, "Wavelet Parameterisation for Speech Recognition: Variations in the scale and translation parameters" Int Symp. Speech, Image Processing and Neural Networks Hong Kong, Vol 2, pp. 694-697, 1994.

Favero, R.F and Gurgen, F, "Using Wavelet Dyadic Grids and Neural Networks for Speech Recognition" ICSLP94 pp 1539 - 1542 1994

Kadambe, S, Srinivasan, P "Applications of adaptive wavelets for speech", Journal of Optical Engineering, Vol. 33, No. 7, pp. 2204-11, 1994

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.

Morlet, J., Arens, G., and Giard, D, "Wave propagation and sampling theory-Part 1: Complex signal and scattering in multi-layered media" Geophysics, Vol. 47, No. 2 pp. 203-221, 1982.

Oppenheim, A.V and Schafer, R.W "Digital Signal Processing" Prentice-Hall New Jersey, 1975.

Rabiner, L "Tutorial on Hidden Markov Models" Proceedings of the IEEE, Vol. 77, No. 2, pp.257-85, 1989.

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

Szu, H., Telfer, B., Kadambe, S. "Neural Network adaptive wavelets for speech representation and classification" Journal of Optical Engineering, Vol. 31 pp. 1907-16, 1992

A HIERARCHICAL APPROACH TO PHONEME RECOGNITION OF FLUENT SPEECH

David B. Grayden and Michael S. Scordilis

Department of Electrical and Electronic Engineering
The University of Melbourne

ABSTRACT - An overview is presented of a hierarchical phoneme recognition system which performs the task in a number of steps: segmentation, manner of articulation classification and then place of articulation classification. A combination of knowledge-based techniques and neural networks are used within these modules.

INTRODUCTION

Automatic speech recognition (ASR) systems designed to handle multiple speakers, a large vocabulary and a large grammar must be able to recognise speech at the lowest level, typically the phonemic level, in order to achieve the required versatility. The phoneme recognition approach has proved popular for building many of the large vocabulary systems in existence (Waibel & Lee, 1990). Information derived from this basic level is used at higher levels of the recognition process such as word selection and syntax and semantic analyses (Lee, et. al., 1990, Sagayama, et. al., 1992, Deller, et. al., 1993).

Fluent speech is composed of sequences of words with little or no pauses between them. Since this is the natural form of speech for humans, it is preferred to uttering connected isolated words. However, when speaking fluently, the boundaries between words become difficult to locate as words blend into one another. Coarticulation is also more prevalent within words as well as between words. This leads to a sequence of phonemes that is different to that produced in isolated word speech, but human listeners are still able to understand without difficulty.

A hierarchical approach has been used in this system for recognition of phonemes in fluent speech. The tasks performed by the recogniser are divided into modules which relate to different acoustic and phonetic processes present in the utterances thus restricting the classification tasks that must be performed. The modular approach also allows flexibility in using different techniques to perform each task. Bayesian classifiers, neural networks and knowledge-based techniques are used where they can provide the best results. This method produces a versatile system that may be altered and adjusted as better classifiers are developed.

SYSTEM FRAMEWORK

An outline of the phoneme recognition system is shown in Figure 1. The first processing operation on the speech signal was the extraction of features relevant to the recognition process. In this system, the features were: 16 mel-scale filterbank energies, low and high frequency energies and a ratio of low frequency to high frequency energy.

Each sentence of speech was then segmented into phonemes with careful attention paid to accurately locating obstruent and sonorant regions. This was performed using a combination of Bayesian classifiers and Time-Delay Neural Networks (TDNNs) (Grayden & Scordilis, 1994).

The next stage in the system classified the phonemes into manner of articulation classes and, finally, determined the place of articulation of each phoneme. Both of these stages were performed by TDNNs (Grayden & Scordilis, 1993).