COMPARISON OF PERCEPTUAL SCALING OF WAVELETS FOR SPEECH RECOGNITION

Richard F. Favero

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

ABSTRACT - Recent work has applied wavelets to speech recognition and has shown that the use of perceptual scaling of the wavelet set can reduce the number of coefficients generated per feature vector compared to standard log frequency scaling. This paper examines the formulation and performance of four frequency scaling operations applied to a wavelet set for the parameterisation of speech in a speech recognition system. Three perceptually based frequency scales: two mel-frequency scales and a bark scale are compared with standard log scaled wavelets.

Application to a multi-speaker E-set discrimination task shows that the piece-wise mel scale provides a recognition accuracy of 67.5%, outperforming the other perceptually based scales slightly, and the standard wavelet log scale by nearly 7%.

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). The perceptually based mel-frequency scale has become the standard parameterisation for HMM-based speech recognition, and perceptually based frequency scales have also been used with wavelets to limit the number of coefficients that were presented to a classifier (Favero and King, 1993; Favero and Gurgen, 1994).

Perceptual scaling of wavelet sets is non-trivial because of the dependence on the mother wavelet. The mother wavelet determines the number of wavelets required to parameterise a given signal, the number of coefficients that are generated per second, and can determine where the wavelets are located in the frequency domain. The modulated wavelets can be modified during scaling to create a wavelet set that has an arbitrary frequency scale. Control of both the resolution of a wavelet (which determines the bandwidth) and its frequency location are necessary conditions for achieving an arbitrary frequency scaling of a wavelet set that has a sufficiently tight frame (Daubechies, 1992; Vetterli and Herley, 1992).

This paper reports comparisons of four frequency scaling operations of wavelet sets for speech recognition. The work outlines the processes of generating wavelets that are perceptually scaled in the frequency domain and shows how the scale (resolution) of a wavelet can be varied independently of its centre frequency.

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 discrete values of a determine the frequency location and bandwidth of the generated wavelet. A series of values for a can be chosen to represent an arbitrary frequency scale. But while the wavelet set is perceptually spaced in the frequency domain, the bandwidths of the wavelets no longer have cut-off frequencies that adequately cover the frequency range. Independent control of the frequency location and the bandwidth of a wavelet (thus the frequency and time resolution) will allow adequate coverage of the frequency domain of the signal to be analysed when an arbitrary frequency spacing is desired.

This control is achieved using modulated wavelets. The centre frequency location is determined by the modulation and the bandwidth by the envelope. The frequency modulation is modified in accordance with the perceptual frequency scale. The bandwidth of the modulating envelope is chosen (based on the centre frequency of the surrounding wavelets) so that the wavelet set is a sufficiently tight frame (covers the frequency domain adequately). Figure 1 shows how different modulating envelopes affect the frequency resolution of the wavelets. Figure 1(a) shows two wavelets with different frequency modulation but the same 10ms envelope. The two wavelets have a cut-off frequency of 7dB. Figure 1(b) shows two wavelets with the same frequency modulation but the envelope is twice as long. This has reduced the bandwidth by half and the spectrum is not covered adequately between the wavelets.

The wavelets used throughout this paper are based on a modulated Hanning window. The Hanning window is an easy to use wavelet since it has finite time duration. This allows easy computation of the wavelet transform. The Hanning window is 32 samples (4ms).

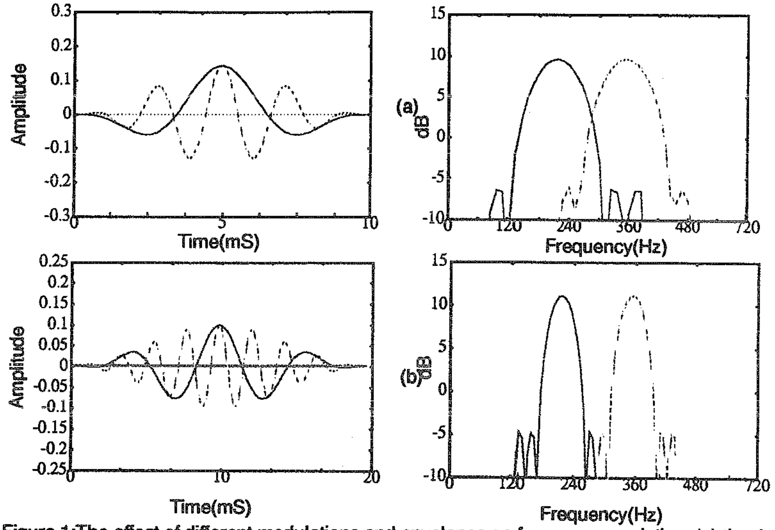

Figure 1:The effect of different modulations and envelopes on frequency resolution. (a) the time and frequency domain representations of an 10ms envelope(b) 20ms modulated envelope

The wavelet transform is performed with the SCWT. The SCWT generates 18 coefficients per sample, every 2ms. The SCWT is modified to reduce computational effort. The modification requires coefficients that lie on the dyadic grid to be copied to adjacent coefficients that lie off the dyadic grid. Since these redundant coefficients were not calculated the computational effort is reduced by 43%.

FREQUENCY BAND SCALING OPERATIONS

The frequency band scaling operations that have been used are outlined below. Each of the frequency scaling operations are evenly spaced to produce 18 wavelets in this experiment.

Log Scaling

The wavelet transform produces a log spaced frequency scale. In this experiment the number of voices per octave has been set to 3 and the number of octaves to cover the frequency range has been set to 6, thus a total of 18 wavelets. Although this number of wavelets is not sufficient to cover the frequency range adequately, it is used for comparison with the perceptual scales. Table 1 shows that the log scale has 12 wavelets below 1000Hz and 6 above. This will allow the log scale to parameterise low frequencies better than high frequencies.

Mel Scaling

This frequency scale is that used by the Hidden Markov Model Toolkit (Young, 1992). The frequency scale is given by:

$$Mel(f) = 2595 \log\left(1 + \frac{f}{700}\right)$$

The 18 wavelets are evenly spaced over this scale. Table 1 shows the number of wavelets that this produces in each of the three frequency bands. The scale is effectively linear below 1000Hz and logarithmic above 1000Hz. Below 1000Hz the modulating envelope is constant for all wavelets.

Piece-wise Mel Scaling

This frequency scaling was used in previous work (Favero and King, 1993,1994). This scale follows that of the log scale above 1000Hz and is linear below 1000Hz. The number of voices per octave above 1000Hz is 6 and there are 2 octaves producing 12 wavelets above 1000Hz. There are 6 evenly spaced wavelets below 1000Hz that have the same modulating envelope.

Bark Scaling

The bark scaling is that used by Hermanski (1990) and is given by:

$$Bark(f) = 6\log\left(\frac{f}{600} + \left(\left(\frac{f}{600}\right)^{2} + 1\right)^{0.5}\right)$$

This scale is very similar to that of the Mel scaling but with a slight shift in the location of the centre frequencies of the wavelets.

	Piece-Mel	Mel	Bark	Log
High(2000-4000Hz)	6	6	5	3
Mid(1000-2000Hz)	6	4	5	3
Low(<1000Hz)	6	8	8	12

Table 1: Number of coefficients located in frequency ranges

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 uses 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 speaker independent recognition task. The /b/, /d/, /p/, and /t/ are a particularly confusable subset of the E set. The plosive burst is of a short duration (if at all). The formants rise on the onset of voicing for the /b/, /d/ and /t/ but F2 falls to a stable level for /p/. The /v/ often has a low intensity and hence can be confused as /c/, /e/ or /z/. The /c/ and /g/ are high intensity during the consonant and hence are easily discriminated.

The experiments described here use continuous density HMMs with 5 states and 5 weighted Gaussian mixtures (Rabiner, 1989; Young, 1992).

DISCUSSION

Table 2 contains the recognition results for testing and training data for each of the frequency scaling operations as a percentage correct score.

	Piece-Mel	Mel	Bark	Log
Training	78.4	77.8	78.6	79.1
Testing	67.5	65.6	65.0	60.9

Table 2: Recognition performance for each frequency scaling operation

There is little difference in the effect of perceptual scaling on recognition performance. The log scale performs poorly for this particular recognition task, although this is not surprising given that there are so few wavelets located in the frequency region where the discriminating information is located. The training result for the log scale highlights that despite a training process performing well, recognition performance with unseen data requires an adequate parameterisation.

The recognition results are consistent with the analysis in Table 1. There is a correspondence between the number of coefficients above 1000 Hz and the recognition performance. The Mel and Bark scales have the similar recognition performance and this corresponds with the number of coefficients above 1000Hz. The piece-wise mel scale has the highest number of wavelets above 1000Hz and the best performance. The log scale has the smallest number of wavelets above 1000Hz and the poorest recognition performance.

The discriminative information is located at the beginning of each utterance where the plosive is occurring and onto the coarticulation with the vowel. Given the plosives contain a significant high frequency content, the results indicate that a frequency scaling operation that has a large number of coefficients in the higher frequency range will provide better recognition performance.

Despite the constant Q analysis being lost through using an arbitrary scaling operation, opportunities exist for determining wavelet scaling operations and wavelets that maximise recognition performance for a given task. Understanding the nature of the recognition task will aid in determining which frequency scaling operation to use.

Future work will investigate scaling operations and wavelets that can be used with recognition tasks that have larger vocabularies. Kadambe (1994) points out that for a particular task with a large vocabulary, different wavelets may be necessary to achieve the best recognition performance. This would extend to the frequency scaling operations. This provides opportunities to use wavelets adaptively for a particular task and to control the number and resoluticn of each of the wavelets.

ACKNOWLEDGEMENTS

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

REFERENCES

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

Hermanski, H. "Perceptual linear predictive analysis of speech" Journal of the Acoustical Society of America, Vol 87, No. 4 pp. 1738-52 1990

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

Rabiner, L., "Tutorial on Hidden Markov Models" Proc. 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

Vetterli, M. Herley, C "Wavelets and Filter Banks: Theory and Design" IEEE Trans. Sig. Proc. Vol. 40, No. 9 pp. 2207-2232, 1992.

Young, S.J "HTK: Hidden Markov Model Toolkit V1.4 Reference Manual" Cambridge University 1992