COMPOUND WAVELETS AND SPEECH RECOGNITION

Richard F. Favero

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

ABSTRACT - This paper reports on a method for improving a speech parameterisation for speech recognition by increasing the bandwidth of a mother wavelet without significantly altering its time resolution. The linear combination of wavelets that have centre frequencies near each other produce a compound wavelet with a larger bandwidth. This paper also shows how more complex wavelets can be constructed for use in other correlation tasks.

This work applies a wavelet parameterisation using compounded wavelets to a discriminative recognition task. The wavelet transform of the speech sample using the resultant wavelets is applied to a HMM classifier. Recognition performance on the E-set discrimination task improves from 67.5% to 70.0% through the use of compounding.

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 to test with wavelet parameterisations to show that improved time and frequency resolution will better parameterise these difficult areas of speech for speech recognisers.

Speech recognition performance has been shown to be sensitive to the choice of mother wavelet and the number of wavelets that cover the frequency domain (Favero, 1994). If there are insufficient wavelets generated in the set, then there will be regions of the spectrum which are not well parameterised. An excessive number of wavelets produces wavelet coefficients with a high correlations that do not assist the parameterisation. In the time domain, the mother wavelet determines how often the wavelet transform must be sampled. Mother wavelets that have a wide time duration require slower sample rates of the wavelet transform than short duration mother wavelets. Thus the time duration-bandwidth product (often referred to the time-bandwidth product) determines the number of wavelet coefficients per second that are required to adequately parameterise a speech sample (Daubechies, 1992).

Underlying speech analysis, as applied to speech recognition, is the assumption that an improved parameterisation will improve recognition performance. This also assumes that a recogniser can exploit the improved feature set. Improving the parameterisation of an already useful feature set would validate the above assumptions.

This paper shows how the method of compounding wavelets can improve the parameterisation of speech for speech recognition. The bandwidth of a mother wavelet used in previous work (Favero and King, 1993, 1994) is extended without significantly affecting its time duration. These wavelets are applied to a discriminative speech recognition task and show an improvement in speech recognition performance. The method is extended to show how more complex and non regular wavelets can be constructed for different classification tasks.

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 used is a modulated Hanning window. The highest frequency wavelet is 4ms wide (32 samples). The SCWT is used to perform the wavelet transform. 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 to locate the wavelets in the frequency domain. There are 12 wavelets above 1000Hz and 6 wavelet below 1000Hz. The wavelet transform generates 18 coefficients per sample every 2ms.

COMPOUND WAVELETS

A pair of wavelets are compounded using the following equation:

$$\Psi a, b, i\,(t) = \alpha \frac{1}{\sqrt{a^{\frac{2i}{n}}}} \Psi\left(\frac{t-b}{a^{\frac{2i}{n}}}\right) + \beta \frac{1}{\sqrt{a^{\frac{2i+1}{n}}}} \Psi\left(\frac{t-b}{a^{\frac{2i+1}{n}}}\right)$$

where *n* is the number of voices per octave, and α and β are weighting coefficients for each wavelet. The equation uses *n* wavelets to produce *n/2* compounded wavelets. This can be extended to compound any number of wavelets.

The frequency response of the compound wavelet is the sum of the two contributing wavelets. The time duration of the wavelet is the longer of the two contributing wavelets. The weighting coefficients can be used to tune the contribution in both the time and frequency domain to achieve a particular response.

Compound wavelets obey the reconstruction admissibility condition (Daubechies, 1992) since the sum of the compound wavelet coefficients will be zero if the sum of the contributing wavelets is zero. While this is a desirable property for wavelet analysis, it is not consequential for speech analysis as applied to speech recognition.

When more than two wavelets are compounded edge effects must be considered. The wavelets that are located on the edge of the compounded band should be weighted so that the frequency response of the compound wavelet is consistent across the compounded bandwidth.

Compound wavelets are computationally efficient. Compounding is performed once during the initialisation of the wavelet transform. The contributing wavelets are computed and then compounding performed. The inner products are then calculated between each of the compound wavelets and the signal of interest.

Increasing the bandwidth of a wavelet.

When both a regular time and frequency domain representation are required, wavelets that are to be compounded should be sufficiently close to avoid ripples in the pass-band. This is the case in speech analysis where the formants are visualised assuming that each block on the time-frequency plot is a regular shape. This and the consistency with the spectrogram makes for easy interpretation of the wavelet transform (scalo-gram) (Grossman et al., 1990).

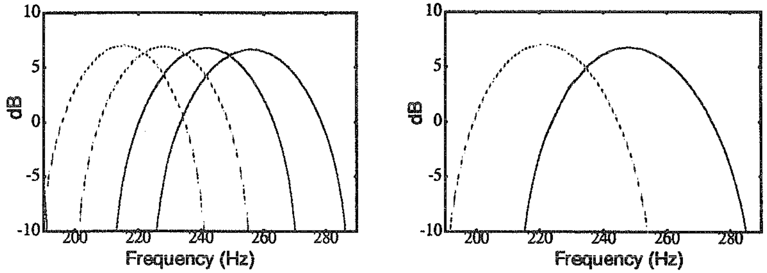

Figure 1:Compounding wavelets to increase bandwidth. (a) Four adjacent wavelets used for compounding. (b) Resulting compounded wavelets

Consider Figure 1(a). An additional wavelet is inserted between adjacent wavelets (the wavelets on the right of a pair. The wavelet pairs are then compounded. Figure 1(b) shows the outcome of the compounding using adjacent pairs. The compounded wavelet has an increased bandwidth and a slightly improved cut-off frequency.

The phase relationships between the two wavelets need to be considered for a particular task. This can be critical for analysis that relies on the phase properties of the signal. Since these compound wavelets are being used for recognition of rather long acoustic events (> 1ms) the phase relationships are not consequential.

Special Signal Correlators

Arbitrarily complex compound wavelets can be created for particular tasks. The weights α and β can be used to provide arbitrary linear combinations of wavelets to achieve a desired analysis. The resultant wavelet continues to satisfy the admissibility condition with the use of these weighting values.

Consider Figure 2. Suppose we wish to use a wavelet to identify third-octave chords within musical signals. Compounding two wavelets that are one third apart can be used as the mother wavelet to be scaled over the entire frequency range. Figure 2 shows how the wavelets are compounded. This new wavelet can now be scaled and translated over the musical range of interest using 12 voices per octave to locate third-octave chords.

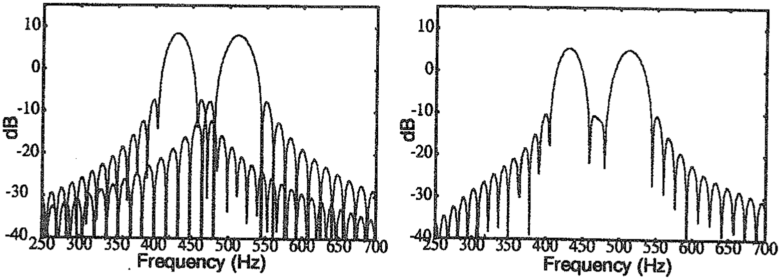

Figure 2:How two wavelets separated by an third-octave can be compounded to create a wavelet for third-octave signal detection

Extending the compound level

The compound level can be extended to increase the bandwidth further and improve the frequency response of the compounded wavelet. Figure 3 shows how successive compounds removes ripples from the pass band of a wavelet that has a narrow bandwidth. With successive compounds the passband becomes smoother.

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. The speech is parameterised with wavelets with compound levels from 2 to 10.

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).

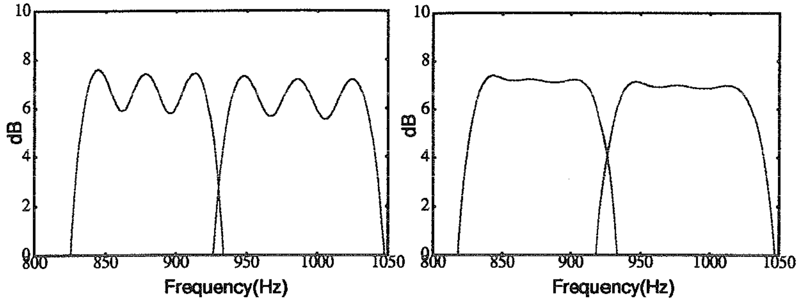

Figure 3:Increasing the compound level reduces the ripple in the passband.(a) 16ms hanning wavelet with compound level 3. (b) 16ms hanning wavelet with compound level 4

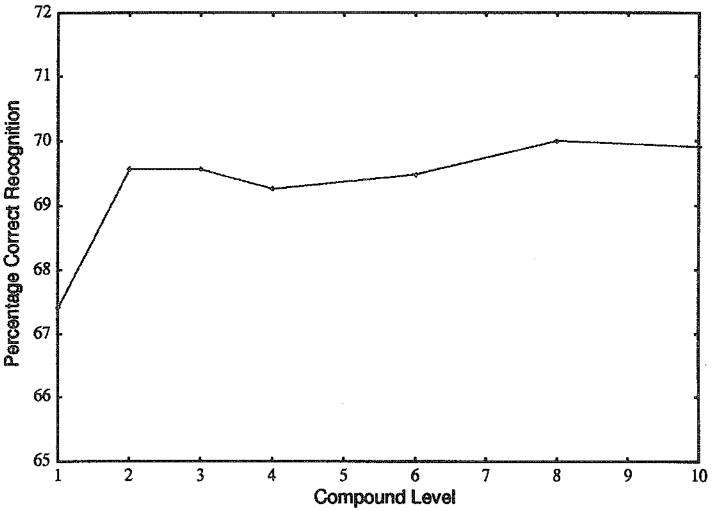

Figure 4:Recognition results for different compound levels

DISCUSSION

The graph in Figure 4 shows that the speech recognition performance increases with increasing compound level. The recognition performance does not improve indefinitely but approaches an upper limit for this particular task and initial wavelet set. The greatest increase in recognition performance occurs in at compound level 2 from which there is little significant gain. There is a small local minimum at compound level 4. The movement is less than 0.5% of the recognition performance and is not statistically significant.

Intuitively, the compounding of the wavelets is, for a fixed number of wavelets, improving the bandwidth of the wavelet so as to improve the coverage of the frequency range. As the frequency range is covered sufficiently, the parameterisation of the speech improves for the given number of wavelets, and hence the recognition performance improves. Given that there is a limit to the amount of information that a fixed number of coefficient can contain, the recognition performance also finds this limit.

CONCLUSION

The recognition results show that by improving the speech parameterisation by improving the frequency domain coverage, speech recognition performance can be improved.

The method of compounding wavelets presented in this paper provides facilities to create wavelets through the linear combination of simple modulated wavelets and how wavelets with arbitrary frequency responses can be constructed.

ACKNOWLEDGEMENTS

The author is grateful to Robin King and Andrew Hunt for many useful 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

Favero, R.F "Comparison of Mother Wavelets for Speech Recognition" SST94 in this proceedings.

Grossman, A Kronland-Martinet, R Morlet, J "Reading and Understanding Continuous Wavelet Transforms" In Wavelets: Time Frequency Methods and Phase space, Springer-Verlag Berlin, pp. 2-20, 1990.

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.

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

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