USING PRINCIPAL COMPONENT ANALYSIS WITH WAVELETS IN SPEECH RECOGNITION

Andrew Hunt and Richard Favero

Speech Technology Research Group
Department of Electrical Engineering
University of Sydney

ABSTRACT - Recent work has shown that wavelets can provide an effective spectral representation for use in speech analysis and speech recognition because of their ability to merge both wide-band and narrow-band spectral representations. There are, however, some difficulties associated with using wavelet parameterisation with HMM-based speech recognition. This paper presents a method which uses PCA to transform the feature space of wavelets so that they can be more effectively used with HMMs. The approach yields good results; up to 25% error-rate reduction is achieved on the difficult E-set discrimination task, along with a seven times reduction in the number of parameters. Further, the training and execution times with the PCA features is greatly reduced.

INTRODUCTION

Recent work (Basile et. al., 1992) has suggested that wavelets may provide a more effective spectral representation for speech analysis than conventional wide-band and narrow-band spectrograms. The primary advantage offered by wavelets is the constant time-bandwidth product across different frequencies which provides the relative advantages of both wide and narrow-band representations. At higher frequencies (> 2 kHz) a narrow time window with wide frequency range is used to give accurate representations of short term events (e.g. plosive transients which occur in only a few milliseconds). At lower frequencies (< 1kHz) a longer time window with more accurate frequency resolution is used which gives more precise representation of slow moving formants and harmonics. The use of this construction gives this wavelet design similar time-frequency capabilities to human audio perception.

Previous work by one of the authors has investigated the use of wavelets for speech recognition as an alternative speech parameterisation technique. Initial work (Favero and King, 1993) indicated that wavelets may offer effective parameterisation for HMM-based recognition and providing slightly better accuracy than Mel-Frequency Cepstral Coefficients (MFCC). There were, however, three primary difficulties with the use of wavelets along with conventional HMM and ANN-based recognisers.

- The wavelet parameters are not time-synchronous; the parameter rate at higher frequencies is greater than at low frequencies.
- The parameter rate is considerably higher than established parameterisation methods; MFCC at around 1200 parameters per second; time-grouped wavelets at 5250 parameters per second; frame-synchronous wavelets at 9000 parameters per second.
- Because of the spectral characteristics of speech, adjoining wavelet parameters are highly correlated which reduces the efficiency of HMM modelling.

Principal Component Analysis (PCA), a multivariate statistical technique (described in more detail below), was applied to the wavelet parameters to address each of the three difficulties described above and with the aim of improving the representation of wavelets for use with HMMs. The approach achieved promising results. A seven-fold reduction in data rate was achieved along with an 25% error-rate reduction on the E-set recognition task compared to MFCC. This paper describes the wavelet parameterisation, the application of PCA, the experimental setup, results and discussion.

WAVELETS

Wavelet theory is based on generating a set of filters by dilation and translation of a generating wavelet. All of the wavelets are scaled versions of the "mother wavelet". This requires that only one filter 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 $\Psi(t)$ by:

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

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

The continuous wavelet transform (CWT) of a signal s(t) is defined as ($a > 0$, b is real):

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

The discrete wavelet transform (DWT) of a sampled signal s(k) is given by (i, k are integers):

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

The DWT computes data points an octave space apart on a dyadic grid if $a = 2$. (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:

$$a = 2^{\frac{1}{numberOfVoices}}$$

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:

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

Feature Vectors

Two forms of feature vector derived from the SCWT are used in this work. The first is the direct use of the 18 point vector that is the output of the SCWT. There were slight modifications to the calculation method to reduce the computation effort; wavelet coefficients that lie on the dyadic grid are simply copied to coefficients that do not lie on the grid.

The second vector is composed from the data values that lie on the dyadic grid. The 42 point vector spans four time intervals of the 18 point vector. Figure 1 illustrates how the 42 point vector is composed from the 18 point vector.

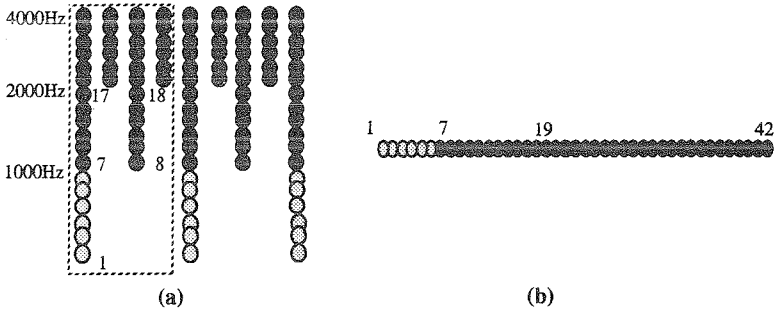

Figure 1: (a) Data on the dyadic grid.
(b) The 42 point vector composed from data on the dyadic grid.

PRINCIPAL COMPONENT ANALYSIS

PCA (Anderson, 1984) is a multivariate statistical technique, often used for data compression, which reduces or reconfigures feature space. It achieves this by producing linear weightings of the elements of the input feature vector which transform the feature space with the following characteristics:

- Each PCA vector explains a maximum of the variance of the input features not already modelled by a preceding PCA vector,
- Successive PCA vectors explain reducing proportions of the input feature variance,
- Each PCA vector is uncorrelated with preceding PCA vectors,
- The maximum number of PCA vectors is equal to the number of input features.

The calculation of PCA weightings is straight-forward. The PCA weightings are the eigenvectors of the covariance matrix of the input feature set. The eigenvalues indicate the proportion of variance explained by each PCA vector and thus provide a useful metric for assessing the transformation.

We apply PCA to the sets of time-grouped and time-synchronous wavelet parameters to obtain a transformed feature space with uncorrelated features and with a high percentage of the wavelet variance explained in a few transformed parameters. The calculation of the covariance matrix and of the eigenvectors takes only a few minutes for the complete training database described below. Thus, the use of PCA adds only a minor increase to the computational overhead of training and testing.

EXPERIMENTAL SETUP

The results presented here are based on the NIST TI-46 word database. This database contains 16 speakers. Each speaker provides 26 repetitions of each of 46 words. Ten samples of the word are used for training and the 16 remaining are used for testing. The database was down-sampled to 8kHz. The leading and trailing silences were removed from each utterance prior to performing the wavelet transform. We chose the "E-set" subtask (b, c, d, e, g, p, t, v, z) to assess the PCA transform because the difficulty in discriminating the initial consonant makes this a difficult speaker independent recognition task which is sensitive to the speech parameterisation.

The experiments described here used continuous density HMMs with 5 states and 5 weighted mixtures (Rabiner, 1989). HMMs were trained from the PCA transformed wavelets from both time-grouped and time-synchronous wavelet vectors. Varying numbers of PCA features were testing (representing a varying percentage of the underlying variance) to determine a suitable trade-off between vector size and recognition accuracy.

RESULTS

PCA Calculation

We first report results on the application of PCA to the wavelet features. Figure 2 shows the cumulative percentage of variance explained by PCA for time-grouped and frame-synchronous wavelet vectors. In Figure 2a, the first 7 PCA wavelet parameters based on frame-synchronous wavelets explain 90% of the variance and the first 11 explain 95% of the variance. In Figure 2b, for the time-grouped wavelets, the first 12 PCA parameters explain 90% of the variance and 21 parameters explain 95% of the variance. Clearly, PCA is able to "distil" the variance of wavelets into a relatively small number of features.

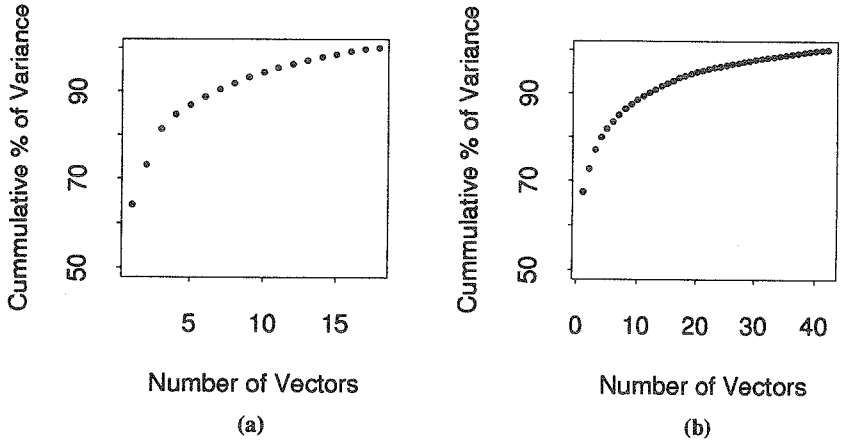

Figure 2: Cumulative % of Variance explained by PCA vectors based on
(a)Frame-Synchronous Wavelets and
(b) Time-Grouped Wavelets.

Interestingly, there is a superficial resemblance between the discrete cosine transform (DCT) used in MFCC calculation and the first few PCA weightings. There is little resemblance for the remaining vectors. It is not clear whether there is any particular significance to this result.

Recognition Results

Table 1 shows the percentage of variance explained (as in Figure 2), the data rate, and recognition accuracy for different numbers of PCA vectors for both the frame-synchronous and time-grouped wavelets. The first observation is that accuracy increases with the number of PCA vectors up to a point, and then drops off slightly. We offer the following explanation; the PCA transformation has captured the most important information in a small number of features which leaves "noise" in the remaining features. This noise reduces the effectiveness of the recognition modelling. The second observation is that a

relatively small number of PCA features is required to achieve satisfactory recognition accuracy. Four PCA features derived from forty two time-grouped wavelet parameters provide significantly better performance than both the untransformed wavelet values (with over ten-fold reduction in data rate) and the MFCC (with less than half the data rate).

# PCA Vectors	Frame-Synchronous Wavelet			Time-Grouped Wavelet		
	% Variance	Data Rate (Params/s)	Recognition Accuracy	% Variance	Data Rate (Params/s)	Recognition Accuracy
1	64.2	500	42.8	67.5	125	-
2	73.2	1000	60.0	72.8	250	-
4	84.7	2000	68.3	79.9	500	69.8
6	88.7	3000	69.5	83.6	750	70.1
8	91.9	4000	71.2	86.5	1000	71.2
10	94.4	5000	69.8	88.5	1250	72.9
12	96.3	6000	67.6	90.1	1500	69.3
18 / 42	100.0	9000	66.5	100.0	5250	-

Table 1: Percentage Variance, Data Rate and Recognition Accuracy for varying numbers of PCA vectors for both Frame-Synchronous and Time-Grouped Wavelets.

Table 2 provides comparative results for MFCC (Favero and King, 1993), untransformed frame-synchronous wavelets, untransformed time-grouped wavelets (Favero and Gurgen, 1994), and results for the best performed PCA transformed frame-synchronous wavelets and PCA transformed time-grouped wavelets from Table 1. The use of the PCA transformation has significantly improved recognition accuracy in comparison to both MFCC parameterisation and non-PCA transformed wavelet parameters. The application of PCA to the frame-synchronous and time-grouped wavelets has provided error-rate reductions of 14% and 19% respectively. The best performance gives a 25% error-rate from MFCC with a similar data rate.

Parameterisation	Vector Size	Frames/sec	Data Rate	Accuracy
MFCC	12	100	1200	63.9%
Frame-Synchronous Wavelets	18	500	9000	66.5%
Time-Grouped Wavelets	42	125	5250	66.6%
PCA Frame-Synchronous Wavelets	8	500	4000	71.2%
PCA Time-Grouped Wavelets	10	125	1250	72.9%

Table 2: Comparison of Parameterisation Methods

CONCLUSION

This work has shown clear benefits from the application of the PCA transformation to wavelet parameterisation. We have been able to reduce data rates by a factor of seven and simultaneously improved recognition accuracy on the E-set task. For frame-synchronous wavelets, the recognition

accuracy improved from 66.5% to 71.2% with the PCA transformation, and for time-grouped wavelets the equivalent improvement is from 66.6% to 72.9%. These results are well in excess of MFCC performance (63.95). The drop in the data rate has also provided substantial improvements in both the training and execution time for the HMMs.

It is not possible to determine whether the improvements in accuracy arise from more the compact parameterisation of the speech signal or because the PCA vectors are uncorrelated and therefore better modelled by the HMMs. Perhaps the improvements are due to a combination of both factors.

It should be possible to extend this work in a number of ways. The covariance matrix can be manipulated to vary the relative importance of each wavelet parameter; this is the equivalent of cepstral liftering. Linear discriminant analysis could be used as an alternative to PCA to provide optimal separation of phonemic classes. Temporal values (e.g. delta or delta-delta parameters) could be introduced to the PCA modelling to capture the dynamic characteristics of speech.

ACKNOWLEDGEMENTS

Both authors are supported by Telecom Research Laboratory Postgraduate Fellowships and by Australian Postgraduate Research Allowances.

REFERENCES

Anderson, T.W. (1984) *An Introduction to Multivariate Statistics: 2nd Edition*, John Wiley and Sons: New York.

Basile, P., Cutugno, F. and Maturi, P. (1992) *The Wavelet Transform and Phonetic Analysis*, Proc. of the Fourth Australian Intl. Conf. on Speech Science and Technology, Brisbane, pp 2-7.

Daubechies, I. (1992) *Ten Lectures on Wavelets*, Philadelphia.

Favero, R.F. and King, R.W. (1993) *Wavelet Parameterisation for Speech Recognition*, Intl. Conf. Signal Processing Applications and Technology, Santa Clara, Vol. 2 pp. 1444-1449.

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

Favero, R.F. and Gurgen, F. (1994) *Using Wavelet Dyadic Grids and Neural Networks for Speech Recognition*, ICSLP '94 pp 1539-1542.

Rabiner, L. (1989) *Tutorial on Hidden Markov Models*, Proceedings IEEE, Vol. 77(2), pp.257-285.

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