THE USE OF WAVELET TRANSFORMS IN PHONEME RECOGNITION

Beng T. TAN, Minyue Fu, Andrew Spray

Dept. of Electrical and Computer Engineering,
The University of Newcastle,
NSW 2308, Australia.

Phillip Dermody

National Acoustic Laboratories,
126 Greville Street,
NSW 2067, Australia.

ABSTRACT

This study investigates the usefulness of wavelet transforms in phoneme recognition. Both discrete wavelet transforms (DWT) and sampled continuous wavelet transforms (SCWT) are tested. The wavelet transform is used as a part of the front-end processor which extracts feature vectors for a speaker-independent HMM-based phoneme recognizer. The results are evaluated on a portion of TIMIT corpus consisting of 30293 phoneme tokens for training and 14489 phoneme tokens for testing. The test results suggest that SCWT gives considerably better recognition rate than DWT. On the other hand, the improvement of SCWT over Mel-scale cepstral coefficients appears to be marginal.

1. INTRODUCTION

The wavelet transform (WT) theory provides an alternative tool for short time analysis of quasi stationary signal, such as speech, as opposed to the traditional short-time Fourier transform (STFT). The WT has been applied widely in different speech analysis problems [16, 8, 9, 7, 3].

Scalograms produced by WT and Spectrograms by STFT have been visually compared [6, 11, 12, 9, 1]. It has been found that both the formant frequencies and harmonic structures of speech are well preserved in the scalogram. This suggests that WT may be suitable for speech analysis. In [5], a continuous wavelet transform (CWT) was used in an E-set alphabet speaker independent isolated word recognition system, and error reduction ranges from 1.6 % to 6.2 % were reported. The work in [10] uses a discrete wavelet transform (DWT) in a small vocabulary speaker-dependent isolated word recognition system. It was shown that DWT performs better than linear predictive coding (LPC) for unvoiced sounds. However, it is unclear whether WT can improve the recognition performance at a phonetic level. The purpose of the present study is to compare both CWT and DWT with Mel-scale cepstral coefficients and report their performances in a speaker independent phoneme recognition system.

2. WAVELET TRANSFORM

The wavelet transform is a non-parametric analysis tool which allows localizations in both the time and frequency domains. The main difference between STFT and WT is that STFT is a constant-bandwidth analysis method, whereas WT is a constant-Q analysis method which resembles auditory filters.

Wavelet coefficients are obtained by computing the correlation between each wavelet and the signal. The realizable form of continuous wavelet transform is called sampled CWT (SCWT), which is most widely used in speech signals analysis [6, 12, 11, 5, 1]. In SCWT, the mother wavelet is truncated in the continuous time ranging from $-\tau_\psi$ to τ_ψ. This wavelet is sampled with the sampling period given by

$$T_\psi = \frac{2\tau_\psi}{N_0}, \tag{1}$$

where N_0 is the number of samples which gives sufficient resolution at the smallest scale (highest frequency) in consideration.

The scaling of the sampled mother wavelet is accomplished by changing wavelet sampling period $T_{a\psi} = T_\psi / a$. The scaling factor, $a \geq 1$, can take any values as long as the resulting representation is not too sparse. The translation parameters is fixed at a constant b_0 to avoid irregular sampling. The SCWT is then defined as

$$SCWT_f(a,n) = \sum_{k=-\lceil \frac{\tau_\psi}{T_\psi} \rceil + nb_0}^{\lceil \frac{\tau_\psi}{T_\psi} \rceil + nb_0} f[k]\psi_a^*[k - nb_0] \tag{2}$$

where

$$\psi_a[k] = |a|^{\frac{-1}{2}} \psi(kT_{a\psi}), \quad -\lceil \frac{\tau_\psi}{T_\psi} \rceil \leq k \leq \lceil \frac{\tau_\psi}{T_\psi} \rceil \tag{3}$$

with its frequency response given by

$$\hat{\psi}_a(w) = |a|^{\frac{1}{2}} \hat{\psi}(\frac{aw}{T_\psi}) \tag{4}$$

The SCWT is implemented simply by linear filtering [12, 11]. It is common to discretize the scale parameter by choosing $a = a_0 2^{m/V}$ where $m \in \mathcal{Z}$ and V is the number of voices per octave.

The DWT is similarly defined as the SCWT, except that a and b are restricted to be on a dyadic grid, i.e., $a = 2^m, b = n2^m$ with $m, n \in \mathcal{Z}$. As the result, the DWT is much coarser than SCWT but it can be implemented very efficiently by fast wavelet transform (FWT) based on subband coding.

3. PHONEME RECOGNITION ENVIRONMENT

This section details the enviornment in which our phoneme recognition tests are done.

3.1. Database

Our phoneme recognition tests are evaluated on the prototype version (1988) of the TIMIT database. We use the DR1, DR2 and DR3 regions only. The training tokens consist of 30 females and 75 males, and the testing tokens consist of 13 females and 37 males. The "sa" sentences which are common to all speakers are not used to avoid possible bias towards certain phones. There are 840 sentences for training and 400 sentences for testing in total. The speech signals are sampled at 16 kHz.

For this study, the silent segments /#h/, /h#/, /epi/ and /pau/ are discarded since we are more interested in modeling phones rather than silence. The rest of 59 phones from the TIMIT phonetic set are used for modeling. Most speech recognition systems [14, 13, 15] select about 42 to 48 phonemes to model. This is done by grouping the allophones into one phone group. 15 allophones from TIMIT phonetic set are identified in [14]. Therefore, there are seven groups of allophones as shown in Table 1 where within-group confusions are not counted as errors. Thus, 59 phone models are built but there are effectively only 46 phonemes to disambiguate. In this sense, the phoneme recognition rate represents "accepted correct" recognition rate [15].

Set	Phone
1	er axr
2	m em
3	n nx en
4	ng eng
5	hh hv
6	pcl tcl kcl qcl bcl dcl gcl
7	el l

Table 1: Seven group of allophones

3.2. Baseline System

All the speech signals are preemphasized by a factor of 0.95 prior to parameterization. Our baseline system uses Mel-frequency cepstral coefficients which are computed using 40 triangular bandpass filters as described in [4]. Each analysis frame has a duration of 20 ms with a 10 ms overlap. Cepstral coefficients of order 12 are produced as the feature vectors.

The phoneme recognizer consists of 59 phone models. Each phone is modeled by a three state left-to-right HMM. The output probability distribution of each state is modeled by a mixture of three multivariate Gaussian density functions with a diagonal covariance matrix.

The initial estimate of HMM parameters is found by using the segmental K-mean algorithm. The Baum-Weltch re-estimation algorithm is then used to further enhance the initial estimate. During the re-estimation process, the floor value for the transition probability and the mixture coefficients is chosen to be 0.00001, and that for the diagonal elements of the covariance matrix is taken to be 0.01.

The same HMM system is used for all the tests (Mel-scale cepstral coefficients and wavelet transforms).

3.3. Wavelet Transforms

Window modulated wavelets have been widely used in speech analysis. Examples include Gaussian (Morlet wavelet) [12, 11], Hamming window [2], and Hanning window [1, 5]. It is important to choose a wavelet function which suits the application. For example, the Mexican hat wavelet is popular in vision analysis, but it is not suitable for speech analysis due to its flat frequency response which results in low formant resolution. We choose to use the Morlet wavelet to construct SCWT.

After normalization and by taking $\omega_0 = 5.5$, the Morlet wavelet function is given as,

$$\psi(t) = exp(-i\omega_0 t)exp(-t^2/2) \quad (5)$$

Morlet wavelets are shown in Figure 1. The mother wavelet

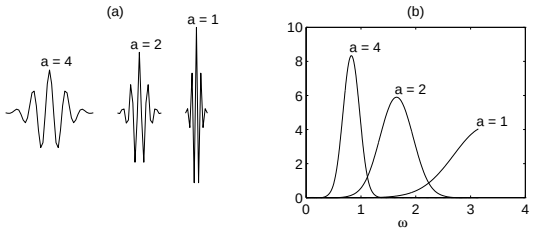

Figure 1: (a) The real part of Morlet wavelet (b) The frequency response of Morlet wavelet.

is a complex function and therefore the CWT coefficients are complex numbers. In our implementation, $N_0 = 10$ and $\tau_\psi = 3$. This Morlet wavelet has a constant Q factor of about 3.3087, which corresponds to 1/2.28 octave bandwidth (c.f. the critical band of the ear, i.e. 1/3 octave). Figure 2 shows a speech segment processed by SCWT with $a_0 = 1$, $V = 8$ and $m = 0, ..., 53$. b_0 is set to be 1 for all scales. Both harmonic and formant structure are preserved in the same plot. The output of SCWT is half-wave rectified and low-passed. It is then down sampled from 16 kHz to 100 Hz. Cepstral analysis is used to reduced the wavelet coefficients down to 12 cepstral coefficients which are then sent to the HMM system for phoneme recognition.

Our example for DWT is based on the Daubechies wavelet. This wavelet is one of the popular wavelets and has been used for speech recognition [10]. A Daubechies wavelet of order 8 is shown in Figure 3. As mentioned earlier, the

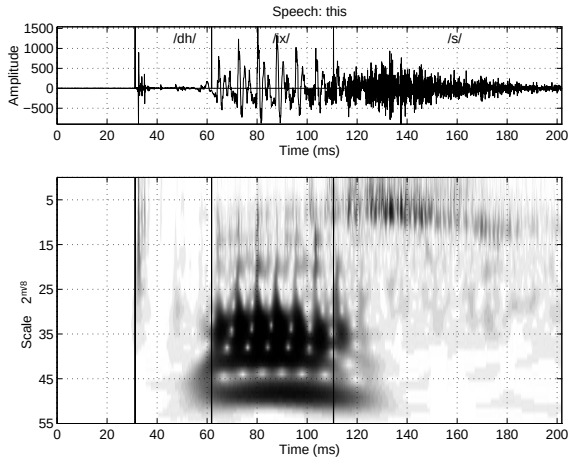

Figure 2: Example of SCWT (Morlet wavelet)

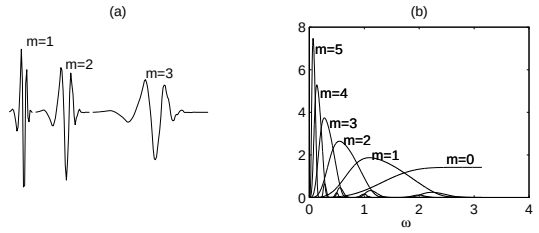

Figure 3: (a) The Daubechies wavelet of order 8 (b) The frequency response of Daubechies wavelet.

DWT is implemented using FWT and that a dyadic scale is used. Subsequently, not much information is retained after the decomposition of the sixth scale. A segment of FWT of speech signal based on Daubechies wavelet is shown in Figure 4. The speech is processed by FWT to produce 6 scale outputs covering 6 octaves. The two FWT coefficients with the largest magitudes at each scale, and they are updated every 8ms using non-overlapping time frames. Note that the number of samples within each frame is different at each scale. Therefore, 12 FWT coefficients are generated from each analysis frame. These FWT coefficients are used as the input to the HMM system. The FWT parameterization does show very sharp onsets.

4. RESULT AND DISCUSSION

The phoneme recognition results are shown in Figure 5. We offer some discussions below.

1) The SCWT gives a slight improvement in the recognition rate of the untrained tokens in comparison with the baseline system. However, the improvement is marginal. We suspect that the main reason for SCWT not being able to provide significant improvement is due to the fact that the many useful features produced by the scalogram are smoothed out after the frame rate reduction. A detailed analysis of our results

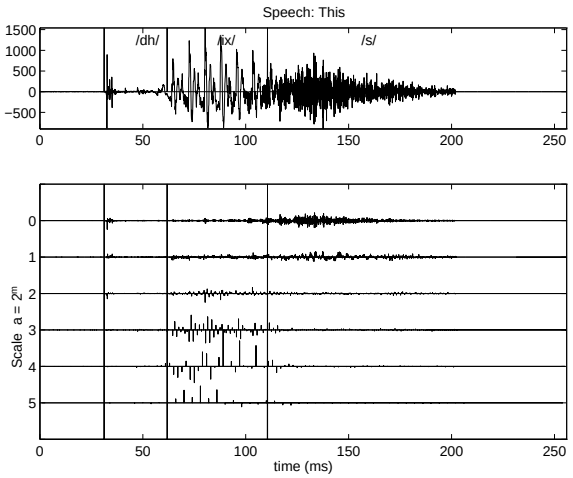

Figure 4: FWT (Daubechies wavelet of order 8) of speech segment "THIS"

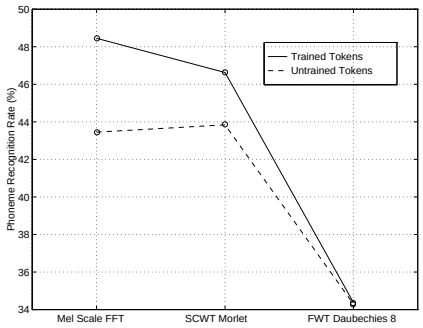

Figure 5: The phoneme recognition results.

(as shown in Figure 6) reflects that the SCWT shows more consistent improvement for most of the vowels and diphthongs than other sounds. This improvement could be due to the fact that the formant structures are well resolved by SCWT. We could have chosen to use a sampling rate higher than 100Hz for frame rate reduction. But further experiments show that 200Hz sampling rate does not give significant improvement. Choosing an even higher rate will make the recognition rate impractical due to the increased processing time.

2) The recognition rate for liquids degrades in the SCWT case. We believe that this phenomenon may be due to the fact that these sounds are characterized by a significant sloping of the resonance bars below 2kHz where the frequency resolution of the wavelet transform is not enough to capture this fast transition. Note that the slopes of this transition are higher than those in the fricative and the stop sounds.

3) The DWT performs very poorly as shown in the figure. We believe that this is caused by the coarse quantization of the DWT. Further, no improvement of the recognition rate for unvoiced speech is observed either. The dyadic decompo-

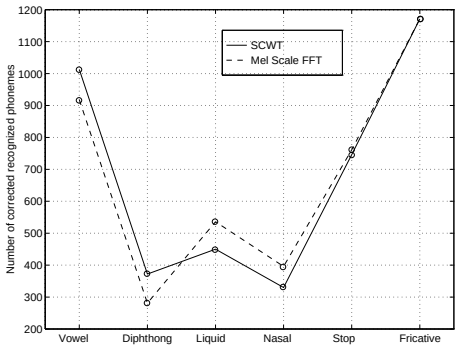

Figure 6: The recognition results of individual group of phonemes.

sition is suitable for speech coding but does not seem to be suitable for speech recognition. Another problem associated with FWT is that it is not time-shift invariant.

5. CONCLUSION

The main advantage of scalograms (SCWT) over spectrograms (STFT) is that the former can preserve both the harmonic structure and the formant structure of the speech signal, resembling analysis performed by the human ear. In particular, sharp onset points can be identified from the scalogram. Due to these features, SCWT seems to have potential application in phoneme speech recogntion.

The phoneme recognition results reported in this paper suggest that SCWT is a significant better choice than DWT for speech recognition. However, the improvement of SCWT over our Mel-scale cepstral coefficients appears to be very marginal. This observation may be unique to the particular post-processing we do to the SCWT coefficients. Work is now being undertaken to identify whether this is indeed the case. In particular, it is not clear how detailed features in the scalogram can be incorporated into the parameterization to give further improvement of the recognition rate, while at the same time keeping the computational time to be moderate. We do note that the compuational time for current implementation of SCWT is much higher than for DWT and also for Mel-scale cepstral coefficients.

6. REFERENCES

[1] P. Basile, F. Cutugno, P. Maturi, and A. Piccialli. The time-scale transform method as an instrument for phonetic analysis. In M. Cooke, S. Beet, and M. Crowford, editors, *Visual representations of speech signals*, chapter 13, pages 169–174. John Wiley & Sons Ltd., 1993.

[2] C. d'Alessandro. Auditory-based wavelet representation. In M. Cooke, S. Beet, and M. Crawford, editors, *Visual representations of speech signals*, chapter 8, pages 131–137. John Wiley & Sons Ltd., 1993.

[3] M. Davenport and H. Garudadri. A neural net acoustic phonetic feature extractor based on wavelets. *IEEE Pacific Rim Conf. on Communication Computers and Signal Processing*, pages 449–452, 1991.

[4] S. B. Davis and P. Mermelstein. Comparison of parametric representation for monosyllabic word recognition in continously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing*, 28(4):357–366, 1980.

[5] R. Favero and R. King. Wavelet parameterization for speech recognition. Preprint, 1993.

[6] A. Grossmann, R. Kronland-Martinet, and J. Morlet. Reading and understanding continuous wavelet transforms. In J. Combes, A. Grossmann, and P. Tchamitchian, editors, *Wavelets: time-frequency methods and phase space*, pages 2–20. Berlin: Springer-Verlag, 1989.

[7] T. Irino and H. Kawahara. Signal reconstruction from modified auditory wavelet transform. *IEEE Trans. Signal Processing*, 41(12):3549–3554, 1993.

[8] S. Kadambe and G. Boudreaux-Bartels. Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans. Inform. Theory*, 38(2):917–924, 1992.

[9] M. Kobayashi and M. Sakamoto. Wavelets analysis of acoustic signals. In *Japan SIAM Wavelet Seminars II*, chapter 3. 1993.

[10] M. Krishnan, C. Neophytou, and G. Prescott. Wavelet transform speech recognition using vector quantization, dynamic time wraping and articicial neural networks. Preprint, 1994.

[11] R. Kronland-Martinet. The wavelet transform for analysis, synthesis, and processing of speech and music sounds. *Computer Music J.*, 12(4):11–20, 1988.

[12] R. Kronland-Martinet, J. Morlet, and A. Grossmann. Analysis of sound patterns through wavelet transforms. *Int. J. Pattern Recog. Artificial Intell.*, 1(2):273–302, 1987.

[13] H. Lee, E. Giachin, L. Rabiner, P. Pieraccini, and A. Rosenberg. Improved acoustic modeling for large vocabulary continuous speech recognition. *Computer Speech and Language*, 6:103–127, 1992.

[14] K.-F. Lee and H.-W. Hon. Speaker-independent phone recognition using hidden Markov models. *IEEE Trans. Acoust., Speech, Signal Processing*, 37(11):1641–1648, 1989.

[15] M. Ostendorf and S. Roukos. A stochastic segment model for phoneme-based continuous speech recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, 37(12):1857–1869, 1989.

[16] B. T. Tan, R. Lang, H. Schroder, A. Spray, and P. Dermody. Applying wavelet analysis to speech segmentation and classification. In H. H. Szu, editor, *Wavelet Applications*, volume Proc. SPIE 2242, pages 750–761, 1994.