Using Wavelets to Acoustically Analyse Oral Stops

† Shunichi Ishihara and ‡ Catherine I. Watson

† Japan Centre (Asian Studies), the Australian National University, † Department of Computing, Macquarie University
and
‡ Department of Electrical & Computing Engineering, University of Auckland

Abstract

The superiority of Wavelet Transform (WT) has been reported in the field of speech technology. However, there has been very little systematic research investigating whether the WT would be a suitable tool for acoustic phonetics. In this study, therefore, it will be investigated whether WT analysis would provide any more useful phonetic information for speech events than traditional spectral analysis, such as Fast Fourier Transform (FFT). To investigate if there would be any advantages to using WT over traditional spectral representations in an acoustic tool, a qualitative comparison is conducted using the scalograms based on WT and the spectrograms based on FFT of oral stop consonants. It is demonstrated in this study that the scalograms indeed provide some useful phonetic information that the FFT based spectrograms fail to do.

1. Introduction

This study investigates whether or not Wavelet Transform (WT) analysis is an appropriate form of analysis in acoustic phonetics. The superiority of WT to Fast Fourier Transform (FFT) has been reported in various speech technology areas (Kadambe and Srinivasan 1994; Maes 1994; Wassner and Chollet 1996). The results of these studies imply that WT is indeed better able to extract and present acoustic features than FFT. However, this implication needs to be empirically tested. There are indeed some studies which have touched upon the usefulness of the WT in speech acoustics (Blacklock and Shadle 2003). Unfortunately, however, these are not systematic enough to empirically show the usefulness of the WT in speech acoustics. As a result of the lack of systematic research, there is very little validation that WT is an appropriate form of analysis in speech acoustics.

Visual examinations through spectro and sonographic representations based on the FFT are usually the first stage of the acoustic phonetic analysis before moving to a quantitative analysis based on spectro and sonographic measurements. Thus, to investigate if there would be any advantages to using WT over traditional spectral representations as an acoustic tool, a qualitative comparison is conducted using the scalograms based on the WT and the spectrograms based on the FFT in this study. This study is presenting the visual scalogram cues for a visual study.

English oral stops produced by both adults and children are used in this study to investigate the usefulness of WT in speech acoustics. Although stops are short in duration compared to other speech sounds, variant acoustic cues like other sounds are associated to them. In usual speech analysis, spectral information obtained by means of the FFT is used to represent different stop consonants (Blumstein and Stevens 1979; Blumstein and Stevens 1980; Kewley-Port 1983). However, the FFT has superiority in representing the frequency behavior of complex speech waves whereas it does not convey useful information in the time domain if the width of window is set to be wide. In stop consonants, the temporal information of spectral change as well as the spectral information at the burst are important for their classification (Kewley-Port 1983). Therefore, analysis using the WT may provide more useful phonetic information for oral stops than traditional analyses because the WT analysis attempts to solve the problem of the time–frequency resolution that the FFT inherits by decomposing a time series into time space and frequency space simultaneously (Daubechies 1992).

It has been reported in various studies on the temporal and acoustic properties of adults' and children's phonemic speech segments that there is a general pattern in the differences between adults and children (Ohde 1985; Lim and Watson 2002). In this study, therefore, it is also investigated whether any acoustic differences between adults' and children's production of oral stops are observed with special attention to the acoustic differences which the scalograms can identify but the FFT based spectrograms cannot.

1.1. Fourier and wavelet transforms

Short Term Fourier Transform (STFT) assumes that some portion of a non–stationary signal is stationary, and looks at the portion along the time scale so that it can provide the time–frequency representation of a signal. STFT divides signals into small units (or windows) and applies Fourier Transform (FT) to each window along the time scale. As can be understood from this, STFT analysis crucially depends on the selection of the window size. Moreover, once a window size has been chosen, the time–frequency resolution is fixed at all times and frequencies. In addition, the time–frequency resolution of STFT cannot be arbitrarily small due to the uncertainty theorem of Heisenberg (Bracewell 1986). Since STFT analyses a signal by means of a window of finite length (unlike FT), it results in a trade–off between time–resolution and frequency–resolution: namely the narrower the window, the better the time–resolution, but the poorer the frequency–resolution, which is also true of the reverse. These are all drawbacks of STFT resulting from the width of the window function.

WT can be adapted as an alternative approach to STFT to overcome some of the above mentioned problems of

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after abstract only review

the time–frequency resolution that STFT inheritably has. Those problems can be solved in WT by varying the window size according to the frequencies of a signal: namely short windows at high frequencies and long windows at low frequencies.

For the WT, the windows (daughter wavelets) can be obtained by either expanding or compressing the analysis wavelet (the mother wavelet), $\psi(t)$, which is defined in (1):

$$\psi a,\tau = \frac{1}{\sqrt{a}}\psi\left(\frac{a}{t-\tau}\right) \quad (1)$$

where a is the scale parameter, τ is the translation parameter, and $a^{-\frac{1}{2}}$ is for energy normalisation across the different scales. The scale parameter governs the expansion or compression of the mother wavelet, whilst the translation parameter, τ, governs the movement of the mother wavelet along the time axis. Thus, the WT for a signal $x(t)$ is defined in (2):

$$Wa,\tau = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty}\psi\left(\frac{a}{t-\tau}\right)x(t)dt \quad (2)$$

where $W(a,\tau)$ are the wavelet coefficients, and for a mother wavelet at a scale, a, and a translation, τ.

In wavelet analysis the scale parameter is used instead of frequency, it relates to the size of the wavelet and is defined as the inverse of frequency.

1.2. Spectral features of oral stop consonants

Some acoustic cues have been proposed for the identification of the place of articulation in oral stops. These cues are: burst spectra (Fant 1973; Blumstein and Stevens 1979); onglide formant transitions of the following vowel (Lehiste and Peterson 1961; Fant 1973; Kewley-Port 1982) and so on. In this subsection, the received knowledge about the burst spectra for stop classification is explained.

Several studies reported that a high degree of place of articulation separation can be achieved from burst frequencies (Fant 1973; Cassidy and Harrington 1995). Based on a 25.6 ms window from the burst onset, Blumstein and Stevens (1979, 1980) defined three acoustic templates to capture the salient place differences. These templates are: *diffuse–falling* for bilabials (falling spectrum with increasing frequency, and no significant peaks); *diffuse–rising* for alveolars (rising spectrum with increasing frequency, and no prominent peak); and *compact* for velars (a prominent peak in the mid–frequency region). They reported that 85% of the stops could be correctly identified in terms of place of articulation on the basis of these templates. Emphasising the importance of temporal properties associated with stop release gestures, Kewley–Port (1983) defined a set of three time–varing features for stop consonants: *Tilt of the spectrum at the burst onset*; *Late onset of low frequency energy* and *Mid–frequency peaks extending over time*. In matching these features, she reported that average correct identification of stop place was 88%.

Harrington and Cassidy (1999, p88) summaries the relationship between the burst spectra shape and the place of articulation in that 1) 'the burst spectra [p] and [b] are characterised by an even distribution of energy throughout the spectrum that is either flat or falling with increasing frequency'; 2) 'the burst spectra of [t] and [d] also lack a significant concentration of energy in a particular frequency region, but in this case the spectrum rises with increasing frequency in the 2 to 5kHz range' and 3) 'velar stops are more difficult to characterise because their well known context–dependent variation has such a marked effect on the burst spectrum'. However, it is often claimed that a velar consonant has a prominent spectral peak in the mid–frequency range (approximately 1200Hz–3500Hz in 0 to 5000Hz scale) (Fant 1973; Blumstein and Stevens 1980).

1.3. Differences in spectral features between adults and children

Lim and Watson (2002) acoustically compared adult and child stops using the same databases used in the current study. They obtained stop spectra through a series of 128 point FFTs taken across the burst, with each FFT slice overlapping by 50%. Although they say that the overall spectral shapes of the children's stops are quite similar to the adults', and the adults' spectral patterns conformed to those reported in previous studies, they also reported that:

> 'the children's bilabial stops are overall higher in energy than the adults', with their voiced /b/ comparatively having even greater energy in the region above 4500Hz. The spectra for the alveolar stops are very similar between the two speaker groups, though the children do have more energy in the region above 7000Hz... The voiceless velar /k/, on the other hand, have a less compact spectral shape for both adults and children (p138).'

2. Methodology

2.1. Database and materials

The oral stops were obtained from the children speech database (hereafter children database) compiled by Cassidy and Watson (1998) and the Otago speech database (hereafter adults database) (Sinclair and Watson 1995) mainly because these databases are compiled using the same wordlist. The children database includes speech samples collected from 4 boys and 4 girls (C1–C8) aged between 7 and 11 years, all being native speakers of Australian English (AE). Although 21 native speakers of New Zealand English (NZE) participated in the adult database, only the data from 4 men and 4 women (A1–A8) aged between 16 and 33 years old was selected for this study.

The adults and the children databases consist of a set of citation–form productions of 129 different words from each speaker. Those words given in Table 1 which contains one of /p, b, t, d, k, g/ phonemes as the onset oral stop were selected for this study from these two databases. Those target words given in Table 1 have a CVC syllable structure. Although the two databases were based on two different varieties of English, no difference has been reported between word initial stops in AE and NZE (Lim and Watson 2002). The vowel context for the bilabial and alveolar stops is /a, i/ and for velar stops is /a/. These two databases are from two different dialects of English, yet the vowel spaces between

Accepted after abstract only review

AE and NZE are quite similar, and there is little difference between AE and NZE particularly in those vowels used for this study (Watson et al. 1998).

Table 1: *Target words.*

	/a/	/i/
/p, b/	*part, barb*	*peat, beat*
/t, d/	*tart, dart*	*teeth*
/k, g/	*card, guard*	

2.2. Recording, digitisation and labelling

The recording was conducted for the children database in a sound–treated studio at the Speech, Hearing and Language Research Centre, Macquarie University as described in Cassidy and Watson (1998). The speech was recorded onto digital audio–tape (DAT), sampled at 20 kHz and quantised to a 16–bit number. The adult speech was recorded in a quiet room, and sampled at 22.05 kHz and quantised to a 16–bit number. The target words of these databases were presented randomly to the speakers, a word at a time, to avoid the listing effect. The child data consisted of one token for each target word while for the adults, there were three tokens each. Speech samples were segmented and labelled phonetically by trained phoneticians (Cassidy and Watson (1998) for the children database; and Lim and Watson (2002) for the adults database). The labelling process was performed using the hierarchical speech data management system — EMU (Cassidy and Harrington 1996). The criteria for labelling conformed to those defined in Croot and Taylor (1995).

2.3. Frequency and wavelet analysis

The spectrogram is a traditional tool for acoustic phonetics, it is obtained as a consequence of a frequency analysis using the FFT. The spectrograms for this study were produced from purpose written functions in `R`. Spectrograms from two difference analysis windows were investigated, 512 and 128. It is well known that 'frequencies in the upper part of the spectrum are attenuated at a rate of approximately 6dB per octave' due to the 'combination of glottal source spectrum, which slopes at -12dB per octave, and the radiation effect due to the lips, which causes a spectral boost of +6dB per octave (producing a net trend of -6dB/octave) (Harrington and Cassidy 1999, p168)'. In this study, therefore, -6dB/octave trend was compensated for by a preemphasis factor of +6dB per octave. The dynamic range of spectrum was 35–40dB.

In this study the wavelet analysis was done using a modified version of the `cwt` function from the `Rwave` package which runs on `R`. The Morlet wavelet was the mother function. The window size of the WT was 256 samples. This yielded a scalogram with 8 octaves, the number of scale parameters within each octave was set to 16. The equivalent of a spectrogram in wavelet analysis is the scalogram, this shows how the energy of the scale parameters very over time. The scalogram was also produced by the modified `cwt` function.

To match the non–linear scale of the WT, a perceptual scale was used for the FFT based spectrograms. The perceptual scale used for the Y–axis of the FFT based spectrograms is the ERB scale (the equivalent rectangular bandwidth of the auditory filter) that is more directly based on the shape of the auditory filter (Moore and Glasberg 1990). The ERB scale is defined as the number of ERBs below each frequency as can be seen in (3)

$$ERBf = 21.4log(0.00437)f \quad (3)$$

Fig. 1 contains two examples of FFT based spectrograms differing in window size (128 and 512) and one scalogram. Please note that the corresponding frequency scale (Hz) is also given in the Y-axis as well as non–linear scales (ERB and octave).

3. Qualitative Results

In this section we identify features visual cues in the scalograms which potentially show differences in adult and child stop productions.

3.1. Scalogram visual cues for bilabial stops

Fig. 1 contains two FFT spectrograms (a, b) and one scalogram (c) for the [p^{h}i] portion of a typical adult *peat* token. Regarding the bilabial stops produced by adults, Lim and Watson (2002, p138) who used the same databases as the current study, reported that 'the bilabial stops have a flat or falling spectrum with higher spectral energy in the 1000–2500 Hz region.' Conforming to Lim and Watson's report and other previous studies, although the spectral energy is quite diffused at the burst, a spectral energy peak—which is indicated by solid arrows—can also be identified around 3000 Hz from the scalogram and the spectrograms in Fig. 1.

However, if one sets the window size of FFT as 128 (refer to Fig. 1a) to improve the time resolution, the burst spectral peak is less precisely located from the spectrogram.

Scalograms are superior in accurate temporal measurements. Fig. 2 contains a 512 FFT spectrogram (a) and a scalogram (b) for the [ba] portion of a typical adult *barb* token. Using a wide window to enhance frequency resolution makes it very difficult to precisely identify the onset locations of the stop burst and the vowel phonation from the wide–band spectrogram (refer to Fig. 2a). As can be seen from Fig. 2b, on the other hand, the onset locations of the stop burst and the vowel phonation, which are indicated by an arrow, can be precisely located in the scalogram while keeping good frequency resolution.

Another point that needs to be mentioned in this section is that a very strong and sustained low frequency energy (the frequency below approximately 1000Hz) accompanies the burst of bilabials (refer to the dotted arrows of Fig. 1). Although this low frequency energy can be observed from the scalogram as well as the spectrograms given in Fig. 1, what can be better observed from the scalogram is that the energy peak of this very low frequency does not coincide with the high frequency energy at the burst. That is, the peak of this very low frequency energy exists not at the burst onset, but after the burst onset. This point cannot be

Accepted after abstract only review

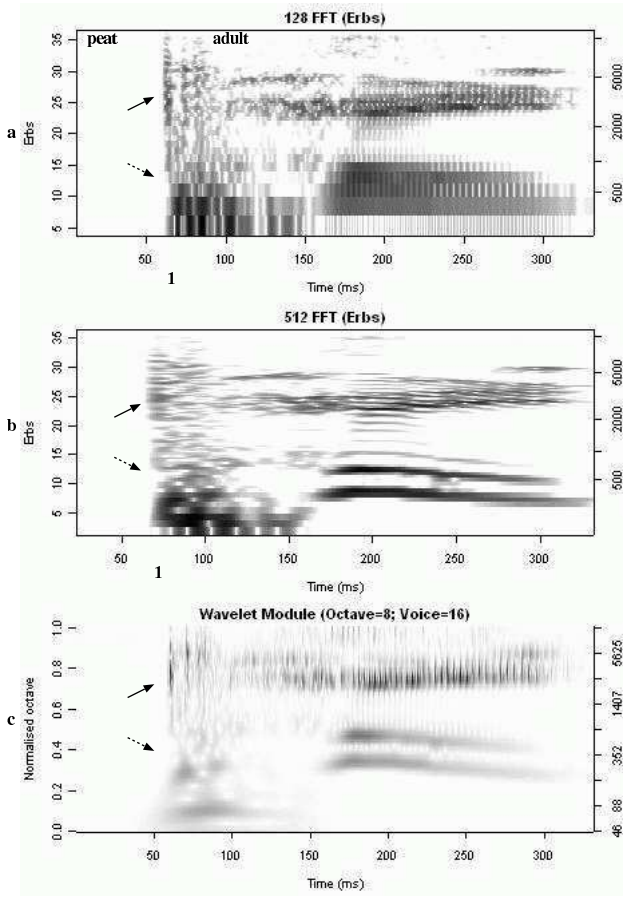

Figure 1: *FFT (128 and 512 point windows) spectrograms and scalogram of the [p^{h}i] of 'peat' uttered by A4. 1 is the onset of the stop burst.*

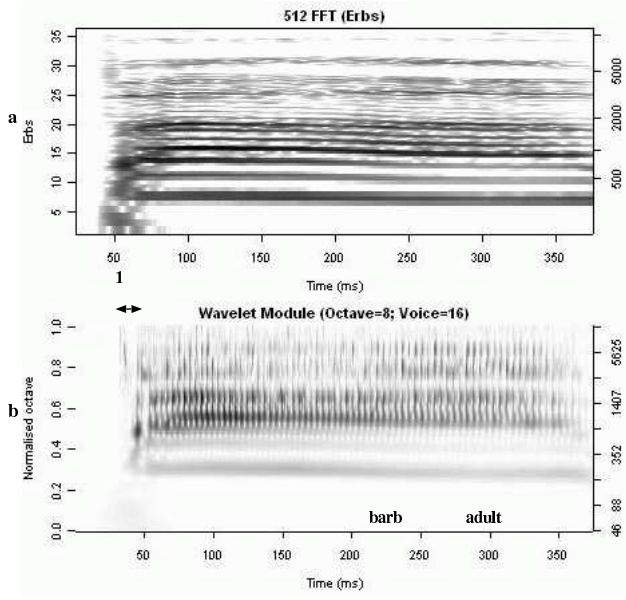

Figure 2: *512 FFT spectrogram and scalogram of the [ba] of 'barb' uttered by A3. 1 is the onset of the stop burst.*

clearly observed if the window of FFT is too wide (i.e. Fig. 1b).

In fact, the existence of this low frequency energy appears to be one of the main differences between the bilabial stops produced by adults and those by children. In many cases, this very low frequency energy is not observed in the bilabial stops of children, or even if it is observed, unlike adults, its peak energy coincides with other higher frequency energy at the burst and the energy is very weak, as indicated by an arrow in Fig. 3.

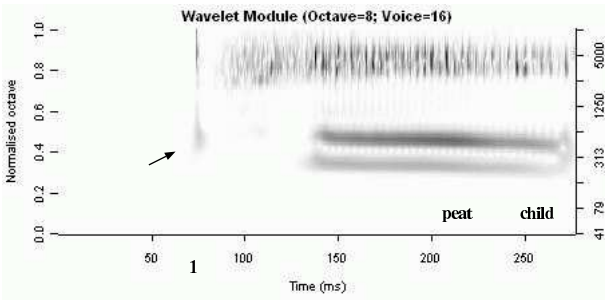

Figure 3: *Scalogram of the [p^{h}i] of 'peat' uttered by C2. 1 is the onset of the stop burst.*

As has been demonstrated above, one needs to examine both narrow and wide band spectrograms to acquire the same amount of frequency–temporal information as corresponding scalograms can provide.

3.2. Scalogram visual cues for alveolar stops

As Harrington and Cassidy (1999, p88) summarise the typical spectra shape of alveolar stops, a spectrum energy rise can be observed with increasing frequency in the 2000 and 5000 Hz in many tokens of adult alveolar stops. This point can be well observed in Fig. 4 that contains two FFT spectrograms (a, b) and one scalogram (c) of the [t^{h}a] portion of a typical adult *tart* token.

The 512 FFT spectrogram and the scalogram of Fig. 4 clearly exhibit that the spectral energy starts increasing from approximately 1000 Hz peaking around 4000 Hz—which is indicated by black arrows—as the energy peak at the burst point. However, due to its poor frequency resolution, the 128 FFT spectrogram of Fig. 4a does not show a rising spectral shape at the burst as clearly as the 512 FFT counterpart and the scalogram do, and as a result, it fails to accurately locate the energy peak at the burst (refer to the dotted arrow of Fig. 4a). Also note though that the start time of this spectral energy peak can clearly be seen in the scalogram, but not in the spectrogram from the 512 FFT.

Regarding the differences between adults and children in alveolar stops, Lim and Watson (2002, p138) reported that '[t]he children have more energy in the region above 7000 Hz' than the adults. Although it was initially suspected that it may be difficult to observe this subtle difference between adults and children—that the latter has more energy peak in the higher frequency region than the former—from the scalograms because non–linear scales are used in the Y–axis, this subtle difference can be comfortably observed from the scalograms. Fig. 5 contains a scalogram

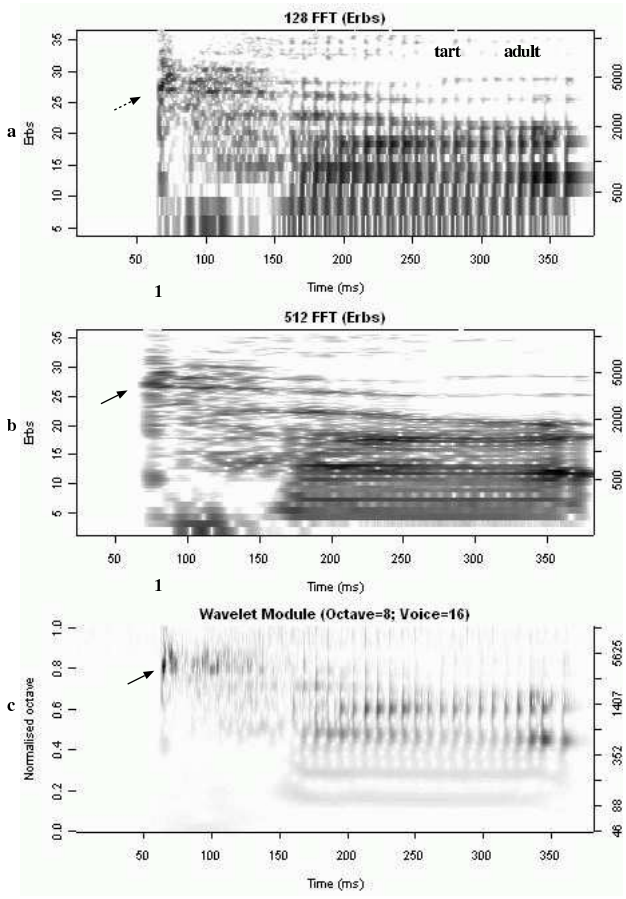

Figure 4: *FFT (128 and 512 point windows) spectrograms and scalogram of the [t^ha] of 'tart' uttered by A6. 1 is the onset of the stop burst.*

of the [t^hi] portion of a *teeth* produced by a child. By comparing Figs. 4c and 5, it is apparent that the frequency energy is concentrated higher in the children than the adults.

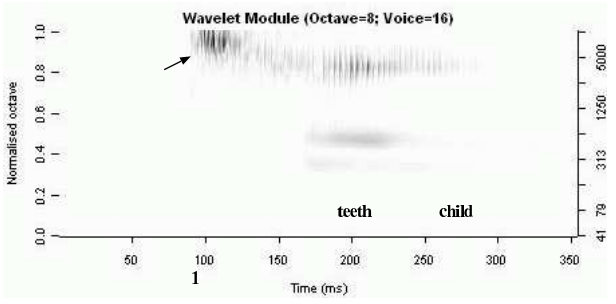

Figure 5: *Scalogram of the [t^hi] of 'teeth' uttered by C3. 1 is the onset of the stop burst.*

3.3. Scalogram visual cues for velar stops

As previous studies report, velar consonants are typically accompanied with a prominent burst spectral peak in the mid–frequency range of the 0 to 5000 Hz scale. Fig. 6 contains two scalograms of the [k^ha] portion of *card* uttered by two different adult speakers, respectively. Although many of the velar stop tokens show a compact burst spectral peak in this mid–frequency range as can be seen from Fig. 6a, many others have another energy peak above this mid–frequency range or a diffused spectral energy shape as can be seen from Fig. 6b. The same observation is also true of children's velar stops.

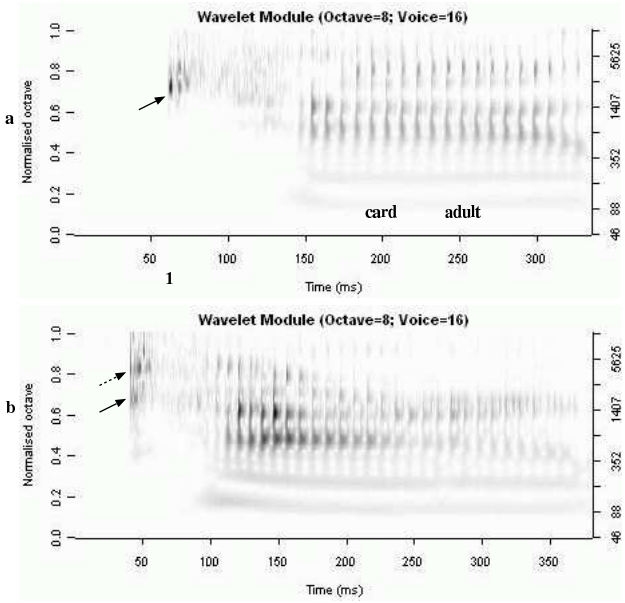

Figure 6: *Two scalograms of the [k^ha] of 'card' uttered by A4 and A5, respectively. Solid arrows are for mid–frequency spectral peak and a dotted arrow for the higher spectral peak than the mid–frequency.*

Although there is a difference in overall spectral shape at the burst point between bilabial and velar stops, the frequency region where the spectral energy peak is observed is very close: (1000–2500Hz for bilabial stops and the mid–frequency range of the 0 to 5000 Hz scale for velar stops). If you compare the stop portions of the bilabials in Fig. 1 and the velars in Fig. 6, it is noticeable that a prominent energy peak can be observed around 3000Hz in both the bilabial and velar stops. However, this energy peak observed around 3000Hz attenuates quickly in the bilabial (refer to Fig. 1c) while it is sustained longer in the velar stop (refer to Fig. 6). It is said that the burst duratIon Of Bilabials is shorter than that of velars (Kewley-Port 1982). This short duration of bilabial burst and the quick attenuation of prominent burst frequency energy can also be observed in Fig. 7b, which includes the scalogram of an adult *part*. Again, this point is very difficult to observe if the window size is set wide (i.e. 512 window points) to enhance the frequency resolution (refer to Figs. 1b, 2a, 3, 7a). This strong and long sustained mid–frequency energy of velar stops is considered to be what Kewley–Port (1983) tried to capture using the temporal feature of 'mid–frequency peaks extending over time'. The difference between bilabial and velar stops in terms of the temporal feature of 'mid–frequency peaks extending over time' can be well observed in scalograms while keeping good frequency resolution.

4. Conclusion

This study has demonstrated that the scalogram is a potentially useful acoustic phonetic tool. There are some

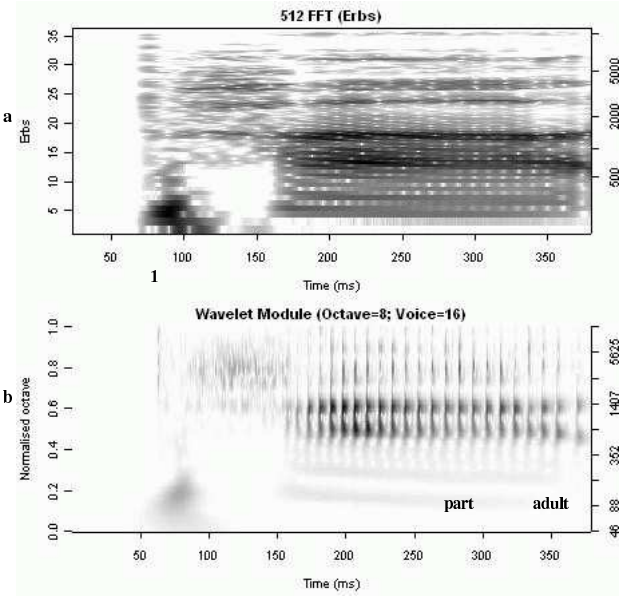

Figure 7: *512 FFT spectrogram and scalogram of the [p^ha] of 'part' uttered by A3. 1 is the onset of the stop burst.*

acoustic differences between adult and child stop productions which can be seen more clearly on a scalogram than a spectrogram. The next step need in this study is quantify these differences.

References

Blacklock, S. and C. Shadle (2003). Spectral moments and alternative methods of characterizing fricatives. *Journal of the Acoustical Society of America 113*(4), 2119.

Blumstein, S. and K. Stevens (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of Acoustical Society of America 67*, 648–662.

Blumstein, S. E. and K. N. Stevens (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of Acoustical Society of America 66*(4), 1001–1017.

Bracewell, R. N. (1986). *Fourier Transform and Its Applications*. MacGraw-Hill.

Cassidy, S. and J. Harrington (1995). The place of articulation distinction in voiced oral stops: Evidence from burst spectra and formant transition. *Phonetica 52*, 263–284.

Cassidy, S. and J. Harrington (1996). Emu: An enhanced hierarchical speech data management system. In *Proceeding of the Sixth International Conference on Speech Science and Technology*, pp. 361–366.

Cassidy, S. and C. Watson (1998). Dynamic features in children's vowels. In *Proceedings of the fifth International Conference on Spoken Language Processing*, pp. 959–962.

Croot, K. and B. Taylor (1995). *Criteria for Acoustic-Phonetic Segmentation and Word Labelling in the Australian National Database of Spoken Language*. http://www.shlrc.mq.edu.au/criteria.html.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia, Pa.: Society for Industrial and Applied Mathematics.

Fant, G. (1973). *Speech Sounds and Features*. Cambridge, MA: MIT Press.

Harrington, J. and S. Cassidy (1999). *Techniques in Speech Acoustics*. Kluwer Academic Publisher.

Kadambe, S. and P. Srinivasan (1994). Application of adaptive wavelets for speech coding. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pp. 632–635.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant vowel syllables. *Journal of Acoustical Society of America 72*, 379–389.

Kewley-Port, D. (1983). Time-varing features as correlates of place of articulation in stop consonants. *Journal of Acoustical Society of America 73*, 322–335.

Lehiste, I. and G. Peterson (1961). Transitions, glides and diphthongs. *Journal of Acoustical Society of America 33*, 268–277.

Lim, L. and C. Watson (2002). An acoustic comparisions of adult and child stops. In *Proceesings of the Ninth Australian International Conference on Speech Science and Technology*, pp. 136–141.

Maes, S. (1994). Nonlinear techniques for parameter extraction from quasi-continuous wavelet transform with application to speech. In *Proceedings of SPIE - The International Society for Optical Engineering*, Volume 2093, pp. 8–19.

Moore, B. and B. Glasberg (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research 47*, 103–138.

Ohde, R. N. (1985). Fundamental frequency correlates of stop consonant voicing and vowel quality in the speech of preadolescent children. *Journal of Acoustical Society of America 78*(5), 1554–1561.

Sinclair, S. and C. Watson (1995). The development of the Otago speech database. In *Proceedings of the Second New Zealand International Conference on Artificial Neural Networks and Expert Systems*, pp. 298–301.

Wassner, H. and G. Chollet (1996). New cepstral representation using wavelet analysis and spectral transformation for robust speech recognition. In *Proceedings of the Fourth International Conference on Spoken Language Processing*.

Watson, C., J. Harrington, and Z. Evans (1998). An acoustic comparison between New Zealand and Australian English vowels. *Australian Journal of Linguistics 18*(2), 185–207.

Accepted after abstract only review