# A wavelet- and neural network-based voice system for a smart wheelchair control

**Article** *in* Journal of the Franklin Institute · February 2011

Impact Factor: 2.4 · DOI: 10.1016/j.jfranklin.2009.02.005



**2 authors:**

M. Al-Rousan
Jordan University of Science and Technology
**43** PUBLICATIONS **554** CITATIONS
SEE PROFILE

Khaled Assaleh
American University of Sharjah
**107** PUBLICATIONS **1,144** CITATIONS
SEE PROFILE

# A wavelet- and neural network-based voice interface system for wheelchair control

Q.P. Ha*, T.H. Tran and G. Dissanayake

Faculty of Engineering,
ARC Centre of Excellence for Autonomous Systems (CAS),
University of Technology, Sydney,
PO Box 123, Broadway, NSW 2007, Australia
E-mail: quangha@eng.uts.edu.au E-mail: ttran@eng.uts.edu.au
E-mail: gdissa@eng.uts.edu.au
*Corresponding author

**Abstract:** Voice control has long been considered as a natural mechanism to assist powered wheelchair users. However, one implementation difficulty is that a voice input system may fail to recognise a user's voice. Indeed, speech activated interface between human and autonomous/semi-autonomous systems requires accurate detection and recognition. In this area pitch and end-point detection is of vital importance. This paper presents a new method for pitch detection based on the continuous wavelet transform phase. The proposed technique can serve as an accurate pitch detector, and also can offer an efficient solution to the end-point detection problem. The extracted features from a user's speech are then used to train a neural network for speech recognition. Experimental results are provided for the detection of pitch periods and end points and the recognition of a number of commands of male and female users. Laboratory tests are reported for the proposed voice control wheelchair system.

**Keywords:** pitch; end-point detection; continuous wavelet transform; neural network; wheelchair.

**Reference** to this paper should be made as follows: Ha, Q.P., Tran, T.H. and Dissanayake, G. (2005) 'A wavelet- and neural network-based voice interface system for wheelchair control', *Int. J. Intelligent Systems Technologies and Applications*, Vol. 1, Nos. 1/2, pp.49–65.

**Biographical notes:** Quang Ha received a BE degree in Electrical Engineering from Ho Chi Minh City University of Technology, Vietnam; a PhD degree in Engineering Science from Moscow Power Engineering Institute, Russia, and a PhD degree in Electrical Engineering from the University of Tasmania, Australia, in 1983, 1992, and 1997, respectively. He is currently a University Reader at the University of Technology, Sydney, Australia. His research interests include robust control and estimation, robotics, and artificial intelligence applications.

Thanh Hung Tran obtained a BE degree from Can Tho University, Vietnam, and a MSc degree from Ho Chi Minh City University of Technology, Vietnam, both in Electronic Engineering in 1996 and 2000, respectively. He was a Lecturer at Can Tho University from 1996 to 2003. He is currently a PhD student at the University of Technology, Sydney, Australia. His research interests include speech recognition, wavelets and applications, and robotics.

Gamini Dissanayake graduated in Mecahnical/Production Engineering from the University of Peradeniya, Sri Lanka. He received his MSc in Machine Tool Technology and PhD degree in Mechanical Engineering (Robotics) from the University of Birmingham, England in 1981 and 1985, respectively. He is now the Professor of Mechanical and Mechatronic Engineering at the University of Technology, Sydney, Australia. His current research interests are in the areas of localisation and map building for mobile robots, navigation systems, dynamics and control of mechanical systems, cargo handling, optimisation and path planning.

## 1 Introduction

Several researchers have considered using human voice to control powered wheelchairs, see, e.g., Simpson and Levine (2002) and the references listed therein. Naturally, a wheelchair voice control system should operate reliably for a large number of users, reduce the physical requirements; and if avoiding the need to move on one or more road extremities, should assist a user in maintaining well the chair position. However, the voice's limited bandwidth makes it difficult to adjust frequently the wheelchair's velocity, and also a voice input system may fail to identify a speaker. Thus, voice interface has yet become commercially viable for wheelchair control; rather its use is normally suggested in combination with a navigation assistance system for obstacle identification and avoidance in the wheelchair's path (Levine et al., 1999).

Human-robot interface plays a very important role in autonomous/semi-autonomous operations that involve interactions with people. These interactions must possess a setting that is easy to participate, interesting and intuitive for ordinary users (Prodanov et al., 2002). In wheelchair control, compared with other means for human-machine interface such as head movement (Taylor et al., 2002), verbal communication remains the most natural (Tran et al., 2004). The area of human-robot voice communication covers many speech research areas such as speech recognition, speech synthesis, speech identification and verification (Prodanov et al., 2002; Takahashi et al; 1998, Nam et al., 1998). Human-robot voice enabled interface, although still in its infancy, has some successful applications in tour-guide robots (Thrun et al., 1999). As noted in Simpson and Levine (2002), a wheelchair voice interface should be user-safe and user-friendly. The latter can be achieved by using just a small number of commands that are simple, consistent and intuitive. The requirement on safety implies a certain level of robustness in handling such commands that may be misinterpreted or perturbed by extraneous noise.

For voice recognition, it is essential to extract those features that are invariant with regard to a speaker while maintaining the uniqueness in order to prevent an impostor. The periodicity of voiced speech known as pitch is considered a key feature that can be used to identify reliably the speaker (Hess, 1983). A pitch period is thus an essential parameter (Nam et al., 1998; Hess, 1983; Zhijin and Jie, 2000) in accurate voice detection and speaker identification. Estimating pitch periods in speech processing is difficult because pitch frequencies can vary from 60 Hz to 500 Hz and the pitch period of the same person may vary depending on the emotional state, accents, and other perceptual variables of that person (Obaidat et al., 1996, Jing and Changchun, 2002). There are a few methods available for pitch period estimation (Kadambe and

Boudreaux-Bartels, 1992; Kader, 2000; Jing and Changchun, 2002). Classical methods, based on the autocorrelation function, average magnitude difference function, and spectrum, are insensitive to non-stationary variations in pitch periods over the segment length, and hence unsuitable for low pitched and high pitched speakers (Kadambe and Boudreaux-Bartels, 1992). Recently, methods based on the discrete wavelet transform have been developed and shown to be suitable for a wide range of speakers (Kader, 2000; Zhijin and Jie, 2000; Jing and Changchun, 2002). As commented in Shah et al. (2002), these methods do not perform well in determining the pitch period under severe noise conditions, where a wheelchair user's speech utterance is quite often in a background of noise. For voice control of such systems as a wheelchair, there exists the need for an accurate method for the estimation of the pitch period and the location of speech end points as well.

In this paper, two essential parts of a voice interface system for wheelchair control will be presented. The first part is for endpoint detection and feature extraction. A new detection method, based on the phase of the continuous wavelet transform (CWT), will be developed. Firstly, the relationship between the CWT phase and the pitch phase is established. An effective algorithm for pitch detection is then proposed, making use of the pitch period parameter. The algorithm is applied to detect starting and ending points of monosyllable words having continuous speech waves. In the second part, the features extracted, namely the voice frequencies and pitch periods, will be used to train a neural network (NN) for the recognition of a number of monosyllable-word voice commands. The results are compared with those using features extracted by the short time Fourier transform (STFT). The proposed pitch detection method, possessing a reliable performance, is applied to voice control of a wheelchair. System configuration and experimental results are provided.

The paper is organised as follows. After the introduction, Section 2 presents the new pitch detection method using the CWT phase. Section 3 provides the detection results obtained by the proposed method and discusses its potential application to voice control of a wheelchair. The NN design, training and recognition results are given in Section 4. The wheelchair hardware configuration, the voice recognition system block diagram, and experimental results are described in Section 5. Finally, Section 6 concludes the paper.

## 2 Pitch detection using the CWT phase

In speech processing, the pitch period is an important parameter in many applications such as speech compress coding, analysis and synthesis, speech segment and automatic monosyllable-word speech recognition. In our proposed wheelchair voice control system, the pitch period will be used in the end-point detector and as an extracted feature for NN training and voice recognition.

For speech and image processing, the wavelet transform, developed as a branch of applied mathematics in the late 1980s, has become a popular tool (Daubechies, 1990). A set of wavelets is generated from the mother wavelet, $\psi(t)$, by

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{1}$$

where the scale $a > 1$ ($a < 1$) is for dilation (contraction) and $b$ is the time step by which wavelets are moved over the signal to be analysed. Wavelets have been successfully used for speech analysis, pitch detection and speech recognition (Favero, 1994). Because of their flexible frequency resolution, wavelets can deal well with variations in speech signals. Speech signals are generally non-stationary and comprising many frequency components. A segment of a speech signal can be mathematically represented, to a given accuracy, in the Hilbert space as the sum of exponential functions by:

$$x(t) = \sum_{i=1}^{N} X_i(t) e^{j\phi_i(t)}, \tag{2}$$

where $X_i(t)$ and $\phi_i(t)$ are, respectively, the instantaneous amplitude and phase of the $i$th frequency component, $i = 1, 2, \ldots, N$ (suppose that there are $N$ frequency components). The instantaneous frequency is defined as

$$\omega_i(t) = \frac{\mathrm{d}\phi_i(t)}{\mathrm{d}t}. \tag{3}$$

The continuous wavelet transform (CWT) of a speech signal $x(t)$ can be written as follows (Vetterli and Kovacevic, 1995):

$$\mathrm{CWT}_x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \psi^* \left( \frac{t-b}{a} \right) x(t) \mathrm{d}t.$$

By incorporating equation (2), one has:

$$\mathrm{CWT}_x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \sum_{i=1}^{N} X_i(t) e^{j\omega_i(t)t} \, \psi^* \left( \frac{t-b}{a} \right) \mathrm{d}t. \tag{4a}$$

or, from the linear property of CWT (Vetterli and Kovacevic, 1995),

$$\mathrm{CWT}_x(a,b) = \frac{1}{\sqrt{a}} \sum_{i=1}^{N} \int_{-\infty}^{\infty} X_i(t) e^{j\omega_i(t)t} \, \psi^* \left( \frac{t-b}{a} \right) \mathrm{d}t. \tag{4b}$$

The wavelet function, $\psi((t-b)/a)$, limits the speech signal to a short time interval $\Delta t$ around $t = b$ and as this signal is slowly time-varying, one can assume that in $\Delta t$ the instantaneous amplitude $X_i(t)$ and the instantaneous frequency $\omega_i(t)$ are constants $X_i(t) = X_i(b)$ and $\omega_i(t) = \omega_i(b)$. Equation (4b) can be rewritten as

$$\mathrm{CWT}_x(a,b) = \frac{1}{\sqrt{a}} \sum_{i=1}^{N} X_i(b) \int_{-\infty}^{\infty} \psi^* \left( \frac{t-b}{a} \right) e^{j\omega_i(b)t} \mathrm{d}t$$

$$= \frac{1}{\sqrt{a}} \sum_{i=1}^{N} X_i(b) \int_{-\infty}^{\infty} \left[ \psi \left( \frac{t-b}{a} \right) e^{-j\omega_i(b)t} \right]^* \mathrm{d}t,$$

$$\mathrm{CWT}_x(a,b) = \frac{1}{\sqrt{a}} \sum_{i=1}^{N} X_i(b) \left[ \int_{-\infty}^{\infty} \psi \left( \frac{t-b}{a} \right) e^{-j\omega_i(b)t} \mathrm{d}t \right]^*. \tag{5}$$

It can be seen that $\int_{-\infty}^{\infty} \psi((t-b)/a) e^{-j\omega_i(b)t} dt$ itself is the Fourier transform of the wavelet function, $\psi((t-b)/a)$, at instantaneous frequency $\omega_i(b)$, or

$$\psi\left(\frac{t-b}{a}\right) \xleftrightarrow{\text{FT}} a \times \Psi(a\omega) e^{-j\omega b} \tag{6}$$

where $\Psi(a\omega)$ is the Fourier transform of the mother wavelet $\psi(t)$. Substituting (6) into (5) yields

$$\text{CWT}_x(a,b) = \sqrt{a} \times \sum_{i=1}^{N} X_i(b)\, \Psi^*(a\omega_i(b))\, e^{j\omega_i(b)b} \tag{7}$$

$$= \sqrt{a} \times \sum_{i=1}^{N} X_i(b) \Psi^*(a\omega_i(b)) e^{j\phi_i(b)} \tag{8}$$

The scale $a_i$ corresponding to the instantaneous frequency $\omega_i$ is defined as

$$a_i = \frac{\omega_0}{\omega_i}, \tag{9}$$

where $\omega_0$ is the central frequency of the mother wavelet function.

The smallest frequency component in the speech signal, $\omega_1$, is the fundamental frequency or pitch frequency. At the scale value $a_1$ corresponding to this frequency, the band-pass interval of $\Psi(a_1\omega)$ is very narrow because the frequency localisation of the wavelet function at low frequency is very good (Vetterli and Kovacevic, 1995). This band-pass interval is considered to be narrow enough to contain only the fundamental component, $\omega_1$, as shown in Figure 1. Therefore, equation (9) can be rewritten as:

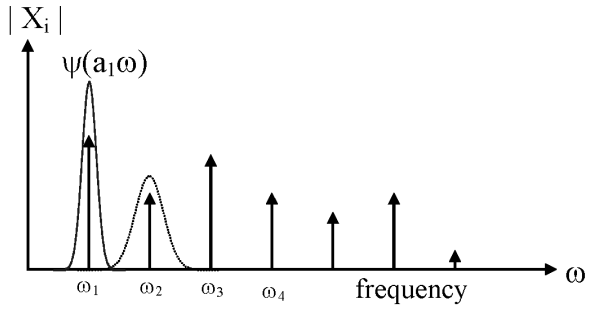
$$\text{CWT}_x(a_1,b) = \sqrt{a_1} X_1(b)\, \Psi^*\, (a_1\omega_1(b)) \times e^{j\phi_1(b)}, \tag{10}$$

where the transform module and phase are determined respectively as:

$$\left|\text{CWT}_x(a_1,b)\right| = \sqrt{a_1} \times X_1(b) \Psi^*(a_1\omega_1(b)), \tag{11a}$$

$$\angle \text{CWT}_x(a_1,b) = \phi_1(b). \tag{11b}$$

**Figure 1** Frequency localisation of wavelet function at pitch frequency



The following proposition is obtained from equation (11b) (Tran et al., 2004):

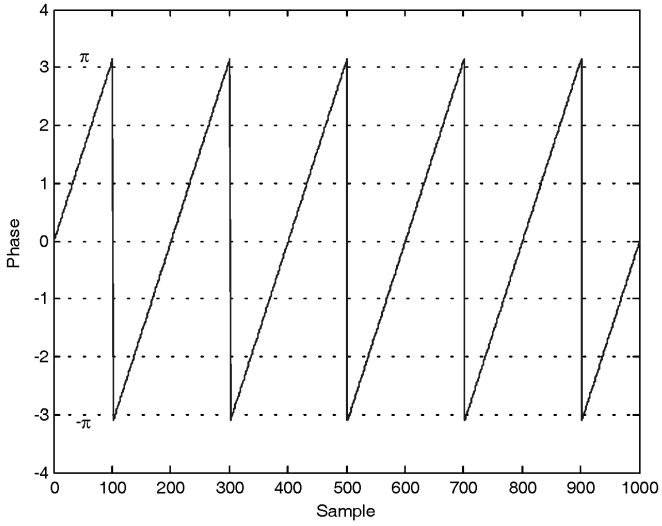
**Proposition:** *At a fixed scale $a_1$ corresponding to the pitch frequency of a speech signal, the phase of the wavelet transform, $CWT_x(a_1, t)$, with time step b varying in the existing time interval of the speech signal, is approximately equal to the phase of the pitch frequency signal*

$$\angle \mathrm{CWT}_x(a_1, t) = \phi_1(t)\,. \tag{12}$$

**Remark 1:** The phase of $\mathrm{CWT}_x(a_1, t)$ is also a periodic signal with the same period as the pitch period.

Indeed, the signal with fundamental frequency $\omega_1$ is a periodic signal with period $T = 2\pi/\omega_1$, so the phase angle $\phi_1(t)$ is also periodic with period $T$, as shown typically in Figure 2.

**Figure 2** Phase of one frequency component




**Remark 2:** Those points where the phase angle changes suddenly from $\pi$ to $-\pi$ can be used to mark periodic points of the pitch period in a speech signal.

Remark 2 serves as the basis of the proposed method using the phase of CWT to extract pitch periods from a speech signal. Figure 3 suggests an algorithm for this. The most important step is to determine the scale value $a_1$ of the speech pitch frequency from equation (9).

Firstly, a 128 ms segment of the speech signal corresponding to the highest energy part is multiplied with a Kaiser window (Owens, 1993) to smooth out the signal. The fast Fourier transform (FFT) is then used to calculate the signal spectrum. The first main peak of the spectral result is the pitch frequency, extracted by local maxima (Hess, 1983). Figure 4 illustrates the detection of the pitch frequency of an adult male signal 'go'.

**Figure 3** Algorithm for pitch detection using phase of CWT

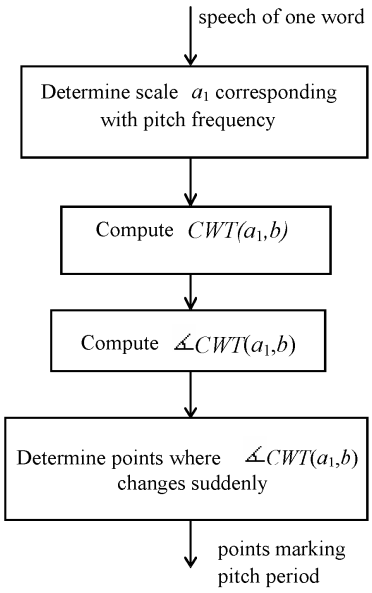
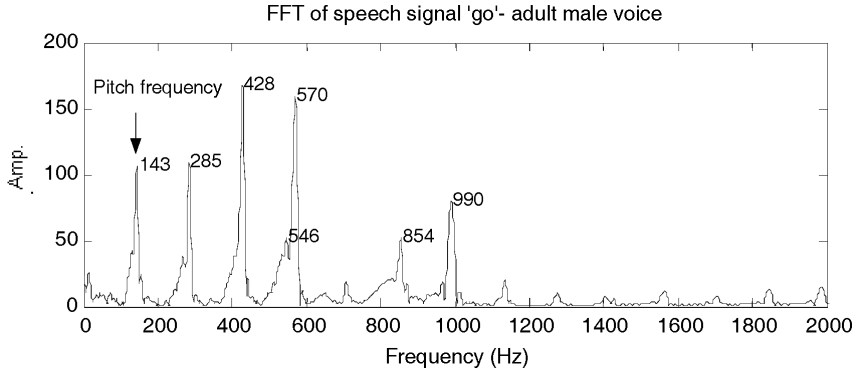



**Figure 4** Pitch frequency estimation




The next steps will be to compute the CWT and its phase at scale $a_1$, and then to determine those points at which the phase angle changes abruptly. In practice, points where $\mathrm{CWT}_x(a_1, t)$ has sudden changes in phase always correspond to main valleys of a speech waveform. These points correspond to the start points of the fundamental component period of each speech signal.

As a result from the above algorithm one can obtain for speech signals those points marking the start and the end of each period. Given the start-point and the end-point (next start point) of a period, one is now able to extract the pitch period or frequency of the signal in consideration.

**Remark 3:** The above algorithm can still be applicable when a pitch frequency cannot be obtained due to a very noisy environment provided information of the pitch frequency average value is known approximately from previous estimations.

## 3 Detection results and discussion

To increase reliability, ease learning and improve consistency of the proposed voice control system for wheelchairs, our list of voice command was intentionally limited to a minimal set of commands that are sufficient to control a wheelchair (Simpson and Levine, 2002). They include monosyllable speeches ‘go’, ‘back’, ‘left’, ‘right’, and ‘stop’. For the purpose of illustration this section details the results obtained when detecting the pitch period and end-points of the speaker-dependent voice command ‘go’ – a simple speech that can be used to control a wheelchair to go forward.

### *3.1 Pitch detection*

A sampling rate of 8 KHz by a PC sound card is used to record speech. The results of the proposed method for pitch detection of speech using the CWT phase are presented in Figures 5–8.

**Figure 5** Results of speech ‘go’, adult male voice

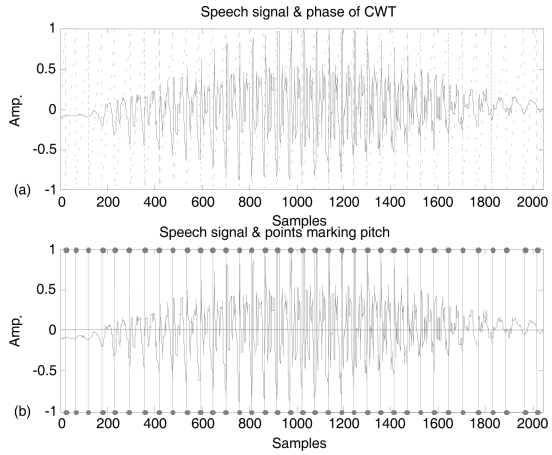



**Figure 6** Zoom in of results of speech ‘go’ – head part

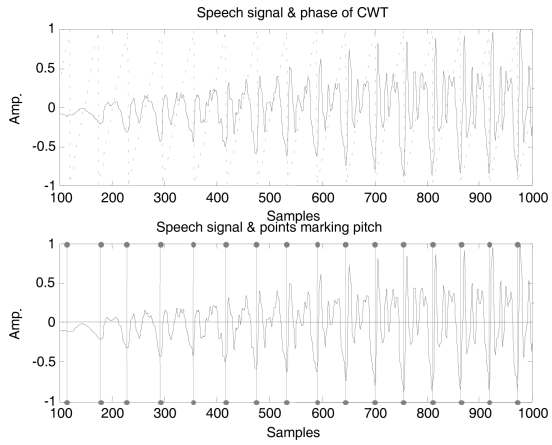


**Figure 7** Zoom in of results of speech 'go' – middle part

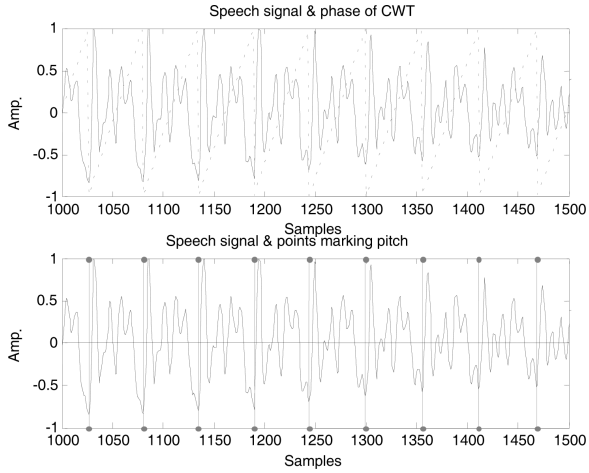
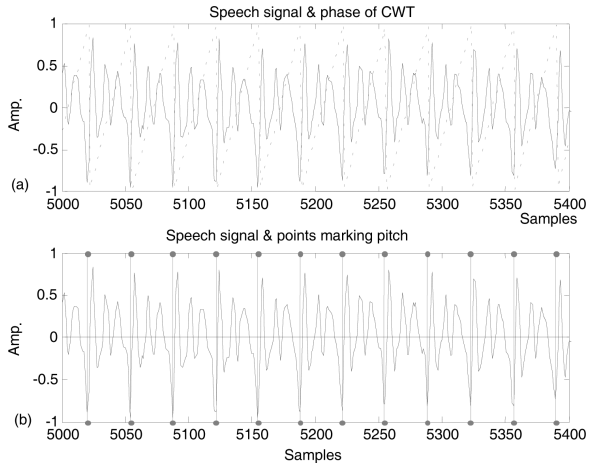



**Figure 8** Results of speech 'go'– adult female voice




After recording an adult male voice command, the amplitude and phase of the corresponding CWT are shown in Figure 5(a), and the signal together with points marking the pitch period in Figure 5(b).

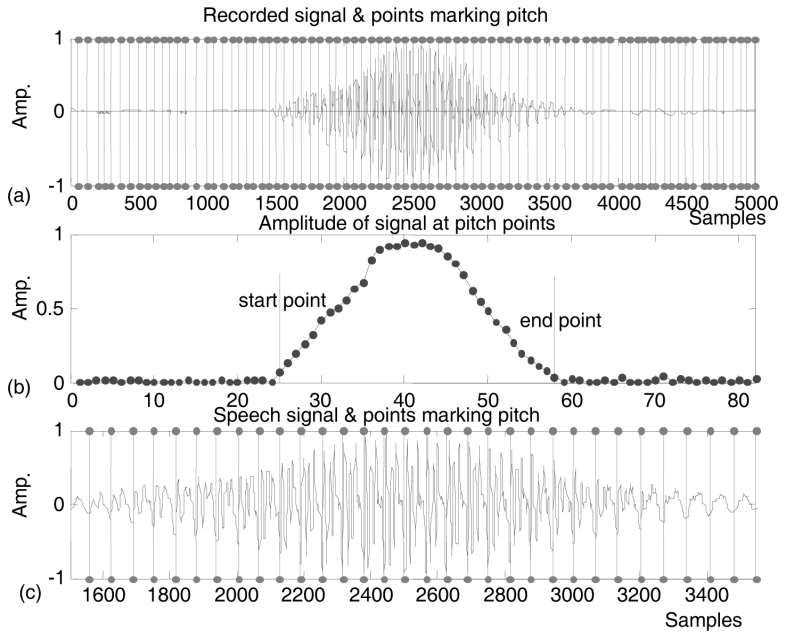
To have a better look at the head and middle parts of the CWT responses, Figure 5 is zoomed in for the first 1000 samples and the next 500 samples, as depicted, respectively in Figures 6 and 7. In these figures, the signal amplitude and the CWT phase angle are normalised in the interval [– 1, 1]. It can be seen that the CWT phase is a saw-tooth waveform whose falling edges exactly locate the main negative peaks (main valleys) of the speech waveform (Figures 5(a), 6(a), and 7(a)). The points marking pitch periods, which are based on the falling edges of the CWT phase, represent therefore the exact location of the main valleys. These start points of each period are shown by vertical lines in Figures 5(b), 6(b), and 7(b).

To indicate that the proposed algorithm can be applied to a wide range of users, a similar result for an adult female voice is shown in Figure 8.

### *3.2 End-point detection*

Extracting exactly speech signals is very important in a speech recognition system, especially in a noisy environment. The proposed algorithm for pitch detection can be used to detect start-points and end-points of speech signals. Results of the end-point detection are shown in Figure 9. The recorded speech signal detected using pitch periods is shown in Figure 9(a). The middle part of the signal together with its pitch period marking points is presented in Figure 9(c). As pitch points always locate main peaks of a speech signal, its amplitude at these points is much larger than that in silence or of noise (Figure 9(b)). An empirical threshold of 0.1 (normalised value) is therefore used to determine the start and end points for the speech 'go' as shown in Figure 9(c).

**Figure 9** End-point detection




The proposed pitch detection method is then applied to recognise, with the aid of a neural network, the above-mentioned set of simple speaker-dependent monosyllable commands. It is interesting to note here that the method can be, moreover, very useful for speech synthesis in human machine voice interface. The neural network based voice recognition system will be detailed in the following section.

## 4 Neural network-based speech recognition

A speech recognition system for human-machine interactions should be able to learn the characterising features extracted from a user's voice. As the system involves a small command vocabulary, a resilient back-propagation neural network (Demuth and Beale, 2003) is selected for fast training. Here, using the proposed pitch detection technique, frequencies and pitch periods are obtained for neural network training.

### *4.1 Feature extraction*

After speech being recorded and end-point detected, monosyllable speech signals are analysed in the scale-time domain with CWT, as shown in Figure 10 for the speech 'go'. Here local maxima are used to extract frequency-time or scale-time information of the speech signal (Figures 10(c) and (d)). The obtained results are two matrices of frequency (or scale) and amplitude values with respect to time. These matrices are re-sampled to reduce the data size. Ten points in each frequency components are extracted. Frequency and amplitude values at these points will be used as the training sets for the neural network. The results of the recognition process are comparable with those when the NN is trained off-line with features extracted by using STFT. For the reference purpose, the frequency extraction using STFT is also presented as shown in Figure 11.

**Figure 10** Frequency extraction using CWT

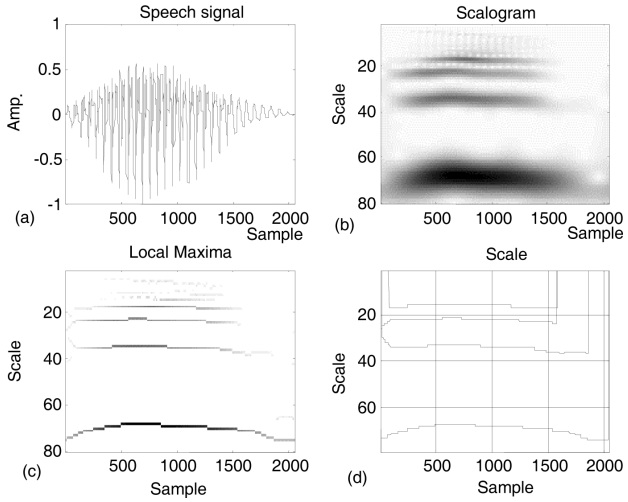
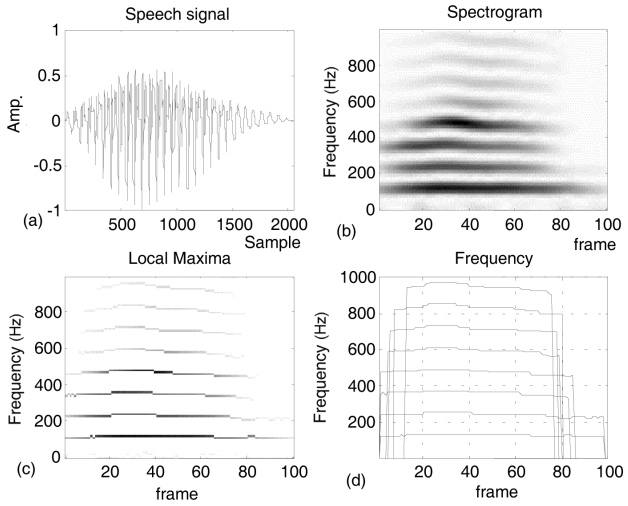
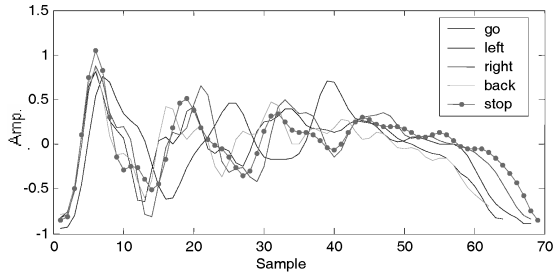



**Figure 11** Frequency extraction using STFT



Notably, the pitch period waveforms of speech signals extracted by the algorithm proposed in this paper can also serve as a valuable input to a speech synthesis system for robot-human interactions via voice (Zhang et al., 2002). Figure 12 shows the normalised amplitudes of the speech waveforms taken in 70 samples for five speech signals mentioned above after being filtered by a 1500 Hz low-pass filter to reduce high-frequency influences. The waveform shape can be used to represent, in approximately one pitch period, characterising information of each speech spoken by a particular user.

**Figure 12** Speech waveforms in one pitch period




## *4.2 Neural network training*

A two-layer feed-forward network with 75 '*tansig*' neurons in the hidden layer and 5 '*purelin*' neurons (Demuth and Beale, 2003) in the output layer corresponding to five signals is used in this paper. The NN has 100 inputs for the frequency feature points extracted by CWT and 50 for the pitch period features. The resilient back propagation method is used for training. For each speech command, 10 datasets are obtained for the NN training purpose. The training process for five control outputs exhibits a fast convergence as shown in Figure 13.

**Figure 13** Results of NN training

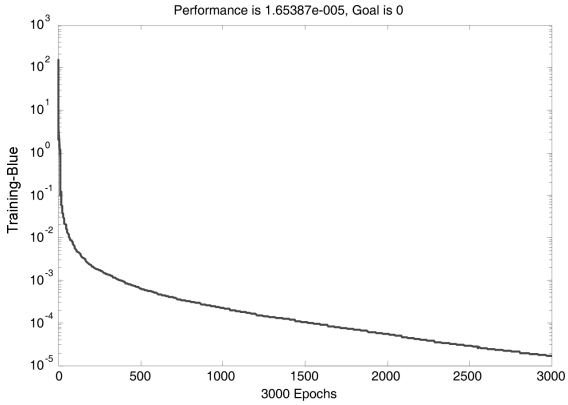



Table 1 shows recognition results using input data extracted by the proposed technique. Each voice command was spoken five times. The NN outputs were almost the same with the voice commands.

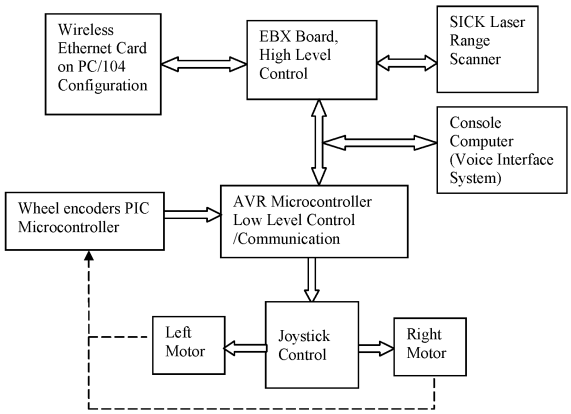
**Table 1** Recognition results

| *Input speech* | *Output* | | | | | *Recognised command* |
|---|---|---|---|---|---|---|
| | *y1* | *y2* | *y3* | *y4* | *y5* | |
| 'go' | 0.9967 | 0.0000 | – 0.0000 | 0.0033 | 0.0000 | 'go' |
| 'go' | 0.9987 | 0.0000 | – 0.0000 | 0.0013 | 0.0000 | 'go' |
| 'go' | 1.0000 | 0.0000 | – 0.0000 | 0.0000 | 0.0000 | 'go' |
| 'go' | 1.0000 | 0.0000 | – 0.0000 | 0.0000 | 0.0000 | 'go' |
| 'go' | 0.9993 | 0.0000 | – 0.0000 | 0.0007 | 0.0000 | 'go' |
| 'back' | 0.0015 | 0.9951 | 0.0034 | 0.0000 | 0.0000 | 'back' |
| 'back' | 0.0000 | 1.0000 | – 0.0000 | 0.0000 | 0.0000 | 'back' |
| 'back' | 0.0000 | 1.0000 | 0.0000 | – 0.0000 | 0.0000 | 'back' |
| 'back' | 0.0003 | 0.9997 | 0.0000 | 0.0000 | 0.0000 | 'back' |
| 'back' | – 0.0000 | 1.0000 | – 0.0000 | 0.0000 | 0.0000 | 'back' |
| 'left' | – 0.0000 | 0.0000 | 1.0000 | 0.0000 | 0.0000 | 'left' |
| 'left' | 0.0000 | 0.0000 | 1.0000 | – 0.0000 | 0.0000 | 'left' |
| 'left' | – 0.0000 | – 0.0000 | 1.0000 | 0.0000 | 0.0000 | 'left' |
| 'left' | 0.0000 | – 0.0000 | 1.0000 | 0.0000 | 0.0000 | 'left' |
| 'left' | 0.0001 | – 0.0000 | 1.0000 | – 0.0001 | 0.0000 | 'left' |
| 'right' | – 0.0000 | 0.0000 | 0.0000 | 1.0000 | – 0.0000 | 'right' |
| 'right' | – 0.0000 | 0.0000 | 0.0000 | 1.0000 | – 0.0000 | 'right' |
| 'right' | – 0.0000 | 0.0000 | 0.0000 | 1.0000 | – 0.0000 | 'right' |
| 'right' | – 0.0000 | 0.0000 | 0.0000 | 1.0000 | – 0.0000 | 'right' |
| 'right' | 0.0997 | 0.0000 | 0.0000 | 0.9003 | – 0.0000 | 'right' |
| 'stop' | 0.0000 | 0.0000 | – 0.0000 | 0.0000 | 1.0000 | 'stop' |
| 'stop' | 0.0000 | 0.0000 | – 0.0000 | 0.0000 | 1.0000 | 'stop' |
| 'stop' | 0.0000 | 0.0000 | – 0.0000 | 0.0000 | 1.0000 | 'stop' |
| 'stop' | 0.0000 | 0.0000 | – 0.0000 | 0.0000 | 1.0000 | 'stop' |
| 'stop' | 0.0000 | 0.0000 | – 0.0000 | 0.0000 | 1.0000 | 'stop' |

## 5 Wheelchair speech recognition system

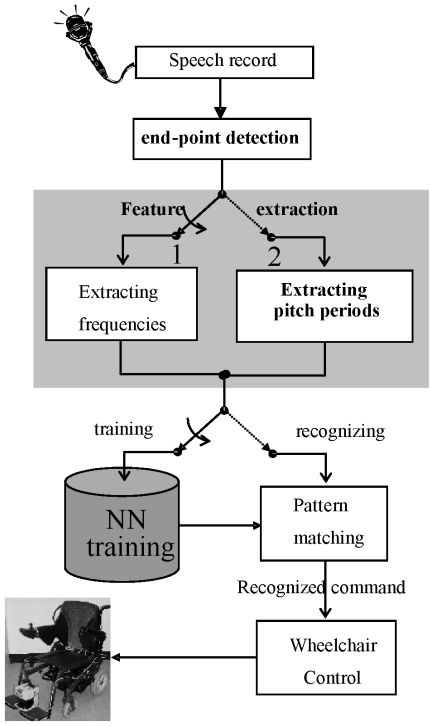
A semi-autonomous wheelchair, developed at the CAS/UTS, is used for testing. The configuration of the powered wheelchair hardware is shown in Figure 14. The platform has two levels of control. At the low level, an AVR microcontroller is used for the control of the joystick and for the communication from the encoder microcontroller. At the height level, an embedded system is used to interface with the low level and a SICK laser range sensor for navigation. The Player-P2OS interface is used to communicate between the two control levels. A console computer is used at the moment for voice control. Note that this voice interface system can be implemented at the high level in combination with a navigation assistance system.

**Figure 14** Wheelchair system configuration




A block diagram for the speech recognition system implemented on the wheelchair is shown in Figure 15. This is a monosyllable, speaker-dependent system with a small vocabulary. A user shall first record his/her voice commands given in the recognised vocabulary. Using the recorded signal, containing also silence and noise, the proposed system shall be able to detect end-points and to extract frequencies and pitch periods as important features for the NN training sets. After training, the user shall be able to control the wheelchair in real time.

**Figure 15** Wheelchair wavelet-based pitch detection speech recognition system



The system has been tested successfully with several different users in laboratory conditions. A photograph extracted from a video-clip is shown in Figure 16. The wheelchair responded smoothly and accurately to the voice commands with a rise time less than 0.5 seconds, mainly due to the electro-mechanical inertia of the powered wheelchair itself. Note that in order for a user to operate the wheelchair, that person should have his/her commands recorded first, and then allow for training the NN using his/her extracted voice features. The system, with command speeches recorded and voice features trained for a particular user, may not be able to recognise accurately the voice commands of another user. This remark can be used to trigger any protection mechanism against impostors.

**Figure 16** Voice control wheelchair

Future research may be directed to the integration of the proposed voice interface with a robotic localisation and mapping system for navigation and the improvement of the recognition accuracy in a noisy environment.

## 6 Conclusion

We have presented a new method for speech pitch detection using the CWT phase. The algorithm can also serve as an efficient end-point detector for speech signals. Important features extracted from the proposed method can find promising applications in speech processing in human-robot voice interface. In this paper, they are used as the training sets for a NN-based speech recognition system of a wheelchair that is controlled with a small vocabulary of monosyllable voice commands. Experimental wavelet-based pitch detection and NN training results are included. These results indicate that the wheelchair can be operated reliably with five voice commands. Laboratory tests are reported for a powered wheelchair, controlled by using our speech recognition method. In combination with a navigation assistance system, the proposed voice interface system can be useful for wheelchair control.

## Acknowledgement

This work is supported, in part, by the ARC Centre of Excellence programme, funded by the Australian Research Council (ARC) and the New South Wales State Government.

## References

Daubechies, I. (1990) 'The wavelet transform, time-frequency localization and signal analysis', *IEEE Transactions on Information Theory*, Vol. 36, pp.961–1005.

Demuth, H. and Beale, M. (2003) *MATLAB Neural network Toolbox*, The MathWorks.

Favero, R.F. (1994) 'Compound wavelets: wavelets for speech recognition', *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, France, 25–28th October, pp.600–603.

Hess, W. (1983) *Pitch Determination of Speech Signals: Algorithm and Devices*, Springer-Verlag.

Jing, L. and Changchun, B. (2002) 'A pitch detector based on the dyadic wavelet transform and the autocorrelation function', *Proceedings of the 6th International Conference on Signal Processing*, Vol. 1, China, 26–30th August, pp.414–417.

Kadambe, S. and Boudreaux-Bartels, G.F. (1992) 'Application of the wavelet transform for pitch detection of speech signals', *IEEE Transactions on Information Theory*, Vol. 38, pp.917–924.

Kader, N.A. (2000) 'Pitch detection algorithm using a wavelet correlation model', *Proceedings of the 17th National Radio Science Conference*, Egypt, 22–24 February, pp.C33/1–C33/8.

Levine, S.P., Bell, D.A., Jaros, L.A., Simpson, R.C., Koren, Y. and Borenstein, J. (1999) 'The navchair assistive wheelchair navigation system', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 7, pp.443–451.

Nam, H., Kim, H.S., Kwon, Y. and Yang, S.I. (1998) 'Speaker verification system using hybrid model with pitch detection by wavelets', *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, USA, 6–9 October, pp.153–156.

Obaidat, M.S., Lee, T., Zhang, E., Khalid, G. and Nelson, D. (1996) 'Wavelet algorithm for the estimation of pitch period of speech signal', *Proceedings of the 3rd IEEE International Conference on Electronics, Circuits, and Systems*, Vol. 1, Greece, 13–16 October, pp.471–474.

Owens, F.J. (1993) *Signal processing of speech*, Macmillan.

Prodanov, P.J., Drygajlo, A., Ramel, G., Meisser, M. and Siegwart, R. (2002) 'Voice enabled interface for interactive tour-guide robots', *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 2, Switzerland, October, pp.1332–1337.

Shah, A., Ramachandran, R.P. and Lewis, M.A. (2002) 'Robust pitch estimation using an event-based adaptive Gausssian derivative fileter', *Proceedings of the 6th International Conference on Circuits and Systems*, Vol. 2, USA, May, pp.843–846.

Simpson, R.C. and Levine, S.P. (2002) 'Voice control of a powered wheelchair', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 10, pp.122–125.

Takahashi, T., Nakanishi, S., Kuno, Y. and Shirai, Y. (1998) 'Human-robot interface by verbal and nonverbal behaviors', *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 2, Canada, October, pp.924–929.

Taylor, P., Nguyen, H. and Craig, A. (2002) 'Head movement recognition for power wheelchair control', *Engineering and Physical Sciences in Medicine 2002*, New Zealand, November, p.135.

Thrun, S., Bennewitz, M., Burgard, W., Cremers, A.B., Dellaert, F., Fox, D., Hahnel, D., Rosenberg, C., Roy, N., Schulte, J. and Schulz, D. (1999) 'Minerva: a second generation museum tour-guide robot', *Proceedings of the IEEE International Conference on Robotics and Automation*, Vol. 3, USA, May, pp.1999–2005.

Tran, T.H., Ha, Q.P. and Dissanayake, G. (2004) 'New wavelet-based pitch detection method for human-robot voice interface', *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sendai Japan, pp.527–532.

Vetterli, M. and Kovacevic, J. (1995) *Wavelets and Subband Coding*, Prentice Hall.

Zhang, W., Xu, G. and Wang, Y. (2002) 'Pitch estimation based on circular AMDF', *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, USA, May, pp.341–344.

Zhijin, X. and Jie, W. (2000) 'A new pitch detection method', *Proceedings of the 5th International Conference on Signal Processing*, Vol. 2, China, August, pp.747–751.