The Application of Bionic Wavelet Transform to Speech Signal Processing in Cochlear Implants Using Neural Network Simulations

Jun Yao, *Associate Member, IEEE,* and Yuan-Ting Zhang*, *Senior Member, IEEE*

***Abstract*—Cochlear implants (CIs) restore partial hearing to people with severe to profound sensorineural deafness; but there is still a marked performance gap in speech recognition between those who have received cochlear implant and people with a normal hearing capability. One of the factors that may lead to this performance gap is the inadequate signal processing method used in CIs. This paper investigates the application of an improved signal-processing method called bionic wavelet transform (BWT). This method is based upon the auditory model and allows for signal processing. Comparing the neural network simulations on the same experimental materials processed by wavelet transform (WT) and BWT, the application of BWT to speech signal processing in CI has a number of advantages, including: improvement in recognition rates for both consonants and vowels, reduction of the number of required channels, reduction of the average stimulation duration for words, and high noise tolerance. Consonant recognition results in 15 normal hearing subjects show that the BWT produces significantly better performance than the WT ($t = -4.36276$, $p = 0.00065$). The BWT has great potential to reduce the performance gap between CI listeners and people with a normal hearing capability in the future.**

***Index Terms*—Bionic wavelet transform, cochlear implants, neural networks, speech signal processing.**

I. Introduction

AROUND 10% of the population in developed countries suffer from hearing impairment [1]. Cochlear implants (CI) have been developed to improve the listening capability for people whose auditory sensors (the hair cells in the cochlea) are not functional. By the year 2000, over 30 000 adults and children around the world had benefited from the use of CI [2]. These devices permit an average word recognition score of 70% [3]. However, a significant performance gap in speech recognition still remains between CI listeners and people with a normal hearing capability, especially in an environment with substantial background noise [4]. Filling this gap is a challenge.

The performance gap between CI listeners and people with a normal hearing capability can be attributed to many factors, one of which is speech signal processing. In the early 1970s, William House and his associates developed the first single-channel implant, which was capable of conveying gross temporal fluctuations as well as some frequency information [5], [6]. The transmitted frequency information, however, was limited and insufficient for speech recognition [7]. In order to improve frequency information, multichannel implant that provides electrical stimulation at multiple sites in the cochlea by using an array of electrodes, was first introduced in the 1980s [8]–[10]. Multichannel implants are now widely used. Some of the state-of-the-art multichannel processors are described below. The Clarion processor developed by the Advanced Bionics Corporation, supports both the simultaneous analog stimulation (SAS) approach and the continuous interleaved sampling (CIS) approach. In the SAS approach, the acoustic signal is processed through seven filters, compressed and then delivered simultaneously to seven-electrode pairs [11]. The CIS approach, on the other hand, is a non-simultaneous pulsatile scheme, where each of the outputs of a bank of eight bandpass filters (BPFs) goes through an envelope detector that extracts the sound temporal waveform envelope, which is used to modulate biphasic pulses [11]. The Nucleus Spectra 24 device, developed by Cochlear Pty. Ltd., Australia, includes advanced combination encoders (ACE), the spectral peak (SPEAK), and CIS strategies. SPEAK strategy encodes the spectral maximum information [12], where the incoming signal is sent to a bank of 20 filters with center frequencies ranging from 250 to 10 kHz, and the six to ten filters with the largest amplitude are continuously selected as the outputs at each cycle. ACE combines the best feature of SPEAK and CIS in a flexible speech coding strategy. The sound is divided into a maximum of 22 frequency bands, and 6–20 bands with the highest energy are selected as the outputs.

Currently, these signal-processing strategies are typically based on BPFs, and are implemented using separate hardware and microprocessors. Consequently, with these techniques it is hard to control the signal processing performance because there are too many signal-processing units involved [13]. Recently, K. B. Nie, *et al.*, and C. P. Behrenbech, independently experimented with new signal processing strategies based on wavelet transform (WT) in order to simplify and minimize the hardware, because WT inherently supports all the features of signal processing provided by separate hardware and microprocessors but in fewer units [13], [14].

All the above approaches use traditional linear filters to span the signal in time-frequency domain (WT also can be viewed as a linear BPF bank), and may not be as effective as a normal human ear, which analyzes the signal by an active and nonlinear system. Research into the physiology of the human ear

Manuscript received March 28, 2001; revised June 3, 2002. *Asterisk indicates corresponding author.*

J. Yao is with the Department of Electronic Engineering, Chinese University of Hong Kong, N. T., Shatin, Hong Kong.

*Y.-T. Zhang is with the Department of Electronic Engineering, Chinese University of Hong Kong, N. T., Shatin, Hong Kong (e-mail: ytzhang@ee.cuhk.edu.hk).

Digital Object Identifier 10.1109/TBME.2002.804590

0018-9294/02$17.00 © 2002 IEEE

shows that the active and nonlinear mechanisms in the ear are necessary for maintaining its high sensitivity and frequency selectivity, and that the loss of active and nonlinear mechanisms results in moderate to profound hearing loss [15], [16]. The drawbacks with the existing linear signal-processing strategies and the limitations that they place on the CI performance motivated us to investigate the potential application of bionic wavelet transform (BWT), an active and nonlinear method stemming from the active auditory model, in CI.

We compare the performance of BWT and WT in CI application in three specific areas: 1) recognition rate; 2) number of required channels; and 3) tolerance of noise. Recognition rate directly reflects the accuracy of a subject's understanding of the speech signal. There are several standardized tests designed for CI performance evaluation. Among them, closed-set vowel and consonant tests are aimed at assessing the subject's ability to resolve spectral and temporal information [17]. For the same purpose, the vowel and consonant tests are used in our experiments.

The question about how many channels are needed to obtain high levels of speech understanding was raised regarding multichannel stimulation. In principle, the larger the number of channels, the finer the frequency resolution for coding. Frequency coding is constrained, however, by the spread of excitation associated with electrical stimulation. To overcome this contradiction, many efforts have been made to design an optimum configuration of electrode stimulation [17]. In our opinion, the tradeoff between the spread of excitation and frequency resolution is based upon the balance of the time and frequency resolutions; and it can be alleviated by using an optimum signal processing strategy.

In noisy surroundings, CI listeners show marked differences to people with a normal hearing capability. Noise reduction with CI is difficult to achieve by way of current BPF techniques, since the noise is usually mixed with the speech signal in the whole frequency band.

To test the BWT performance on these three aspects, we conducted four different experiments, whose methods and results are presented in Section III. Before this, we briefly introduce BWT in Section II. In Section III, we discuss our experimental results and conclude that BWT has potential to reduce the performance gap between CI listeners and people with a normal hearing capability in the future.

II. Bionic Wavelet Transform

The BWT is developed by incorporating the active cochlear mechanism into the WT, resulting in a biologically-based model and an adaptive time-frequency analysis. If the signal to be analyzed is $f(t)$, the BWT is defined as

$$(BWT_T f)(\tau, a) = \langle f, h_T \rangle = \frac{1}{T\sqrt{a}} \int f(t)\tilde{h}^* \times \left(\frac{t-\tau}{aT}\right) \exp\left(-j\omega_0 \left(\frac{t-\tau}{a}\right)\right) dt \quad (1)$$

where a is the scale, τ is the time shift, and $*$ and $\langle , \rangle$ denote the complex conjugate and inner product, respectively. In (1), we set

$$h_T(t) = \frac{1}{T\sqrt{a}} \tilde{h}\left(\frac{t}{T}\right) \exp(j\omega_0 t) \quad (2)$$

which acts as the mother wavelet of BWT. In (1) and (2), $\tilde{h}_T(t)$ is the envelope function that relates to the mother wavelet function of WT, $h(t)$, by

$$\tilde{h}(t) = h(t) \exp(-j\omega_0 t) \quad (3)$$

where $\omega_0 = 2\pi f_0$, and f_0 is the center frequency of $h(t)$; and T function in BWT simulates the active mechanism of the cochlea, which is expressed by

$$T(\tau + \Delta\tau) = \left(1 - G_1 \frac{\mathrm{BWT}_s}{\mathrm{BWT}_s + |\mathrm{BWT}_I(\tau, a)|}\right)^{-1} \times \left(1 + G_2 \left|\frac{\partial \mathrm{BWT}_I(\tau, a)}{\partial(t)}\right|\right)^{-1} \quad (4)$$

where G_1 and G_2 are the active factors, BWT_s is the saturation constant, and BWT_I is the inertial item, which is defined as the mean value of $\mathrm{BWT}(\tau)$, $\mathrm{BWT}(\tau - \Delta\tau)$, and $\mathrm{BWT}(\tau - 2\Delta\tau)$.

Equation (4) stems from the active cochlear model developed based on the observations of otoacoustic emissions, an acoustic signal generated by active cochlea [18]–[20]. With (4), it has been shown that the active cochlear model can successfully generate biologically realistic otoacoustic emissions [18]. Further, according to the analogy between signal processing mechanism of cochlear model and of WT [21], the same T-function was introduced into WT. With this introduction, the adaptive adjustment of resolutions has been achieved in two dimensions, i.e., even in a fixed center frequency; the resolutions in the time and frequency domains are still variable. Experimental results show that BWT achieves better tradeoff between time and frequency resolutions than WT [21].

In implementation, BWT coefficients can be easily calculated based on corresponding WT coefficients by

$$(\mathrm{BWT}_T f)(a, \tau) = K(WTf)(a, \tau) \quad (5)$$

where K is a factor depending on T. Especially, for the real Morlet function $h(t) = e^{-(t/T_0)^2}$, which is used as the mother function in our experiments, where $\tilde{h}(t) = e^{-(t/T_0)^2}$, K is equal to

$$\frac{\int_{-\infty}^{+\infty} e^{-t^2}\, dt}{\sqrt{(T/T_0)^2 + 1}}$$

and approximately equal to $1.7725/\sqrt{(T/T_0)^2 + 1}$ [22].

III. Experiments and Results

A. Material and Protocol of Experiments

The experimental materials consist of seven consonants, /b p d t g k j/ in the contexts /aCa/, /iCi/, and /uCu/ and 11 vowels in the words: "heed, hid, hayed, head, had, hod, hud, hood, hoed, who'd, heard," which were spoken by a single female speaker in a nonsoundproof room. Sound was recorded by a personal computer with a Crystal PnP sound card at the sampling frequency of 22 050 Hz and the resolution of 16 bits. These materials be-

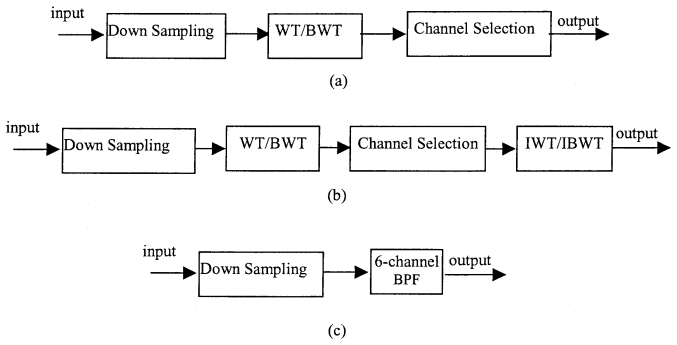

Fig. 1. Experimental protocols. In both protocols (a) and (b), signals were first down-sampled to a sampling frequency of 11 025 Hz, and then processed by WT or BWT with 22 channels in total. After time-frequency decomposing, in every 4-ms interval, the energy of the signal at each channel was estimated by computing the RMS energy of the outputs. Signals in the M ($M \leq 22$) channels, which have maximum RMS energy, were selected as the outputs of WT or BWT analysis. In the protocol (b), the signal was further reconstructed by the IWT or IBWT. In protocol (c), the down-sampled signal was bandpassed into six frequency bands using sixth-order Butterworth filters.

long to a subset of Iowa consonant and vowel test. Part or all of the materials were used in following experiments.

In the experiments, we used three protocols that are shown in Fig. 1(a)–(c), respectively. In both protocols (a) and (b), signals were first down-sampled to a sampling frequency of 11 025 Hz, and then processed by WT or BWT with 22 channels in total. Each of the center frequencies of the 22 channels for both WT and BWT is $f_i = f_0/q^n$, where $f_0 = 15\,165.4$ Hz, $q = 1.1623$, and $i = 1, 2, \ldots, 22$. The calculation step size $\Delta\tau$ is set as 1/11 025 s. After time-frequency decomposing, i.e., WT or BWT processing, in every 4-ms interval, the energy of the signal at each channel was estimated by computing the root mean-square (RMS) energy of the outputs, where RMS energy of signal x_k is defined as

$$\sqrt{\left(\sum_{k=0}^{K-1} x_k^2\right) \Big/ K}.$$

Signals in the M ($M \leq 22$) channels, which have maximum RMS energy, were selected as the outputs of WT or BWT analysis. No compression or amplification was performed to affect the consonant-to-vowel ratios. In the protocol (b), the signal was further reconstructed by the inverse transforms of WT or BWT (IWT/IBWT). In order to compare the performance of BWT to current Fourier based methods, the protocol (c) was used. In protocol (c), the down-sampled signal was bandpassed into six frequency bands using sixth-order Butterworth filters. The center frequency and the 3-dB bandwidth of each filter were listed in Table I, which are the same as those used in six-channel CIS strategy.

TABLE I
THE CENTER FREQUENCY AND 3-dB BANDWIDTH OF DIFFERENT CHANNELS

Channel Number	Center Frequency (Hz)	3-dB Bandwidth
1	393	187
2	639	304
3	1037	493
4	1685	801
5	2736	1301
6	4443	2113

B. Experiment 1

1) Goals and Method: This experiment aimed at comparing the recognition rates of the same speech materials processed by BWT, WT, and BPFs, respectively. All of the materials (consonants and vowels) were used in this experiment. The original signals were processed according to the three previous mentioned protocols, and the outputs were directly sent to a two-layer feed-forward back-propagation neural network that had (frame length) $\times$ (the number of channels) input nodes, 15$\sim$150 hidden nodes, and one output node. In this experiment, the number of selected channels was six; and the frame length was 44. Signals in one frame and six channels were entered into the network together. When moving the frame window, there were ten overlapping data.

With a fixed number of the hidden nodes, all of the materials were trained and tested. The training targets and the standard outputs for testing were automatically set by a program, in which we compared the RMS energy of each frame to a pre-set threshold. If it was lower than the threshold, this frame was defined as a nonspeech frame; otherwise, it was defined as a speech frame. For consonants, the speech frames in the context /vCv/, where /v/ represents the vowel, and /C/ represents the consonant, were further divided into two categories: 1) context-vowel frames including all the frames for the first /v/; and 2) consonant frames including all the frames for /Cv/. Zero, two, and seven continuous integers from 4 to 10, in turn, were set as the standard outputs of nonspeech frames, context-vowel frames, and the seven consonant frames: /tv kv dv bv pv gv jv/, respectively. For vowels, the standard outputs of the speech frames in the words: "had, hid, hod, hud, head, heed, hoed, hood, who'd, hayed, heard" were set as 11 continuous integers from 2 to 12 in turn. The resilient back-propagation training method with the same training parameters was used to train the network. The maximum training loop was 100. Then, the network was used to recognize the same materials as those used in the training session but in a different order. Finally, the recognition rates for all the frames, speech frames, and consonant frames, which were defined as (the number of frames correctly recognized and belong to the specific category)/(the total number of the frames in that category), were calculated.

The size of the network also influences network's performance. To reduce this influence, we kept increasing the size of the network by increasing the number of hidden nodes, and repeated the training and testing procedures for vowels and consonants. We then calculated the saturation recognition rates by fitting the first order exponential function

$$y = y_0 + Ae^{(-(n-n_0/\tilde{n}))} \tag{6}$$

to the recognition rates obtained by BWT, WT, and BPF, respectively. In (6), A is a coefficient; n is the number of the hidden nodes; $n_0 = 15$; y is the recognition rate; and y_0 is defined as the saturation recognition rate.

2) Results: The recognition rates of the consonants in the contexts /aCa/, /iCi/, and /uCu/, and of the vowels versus the

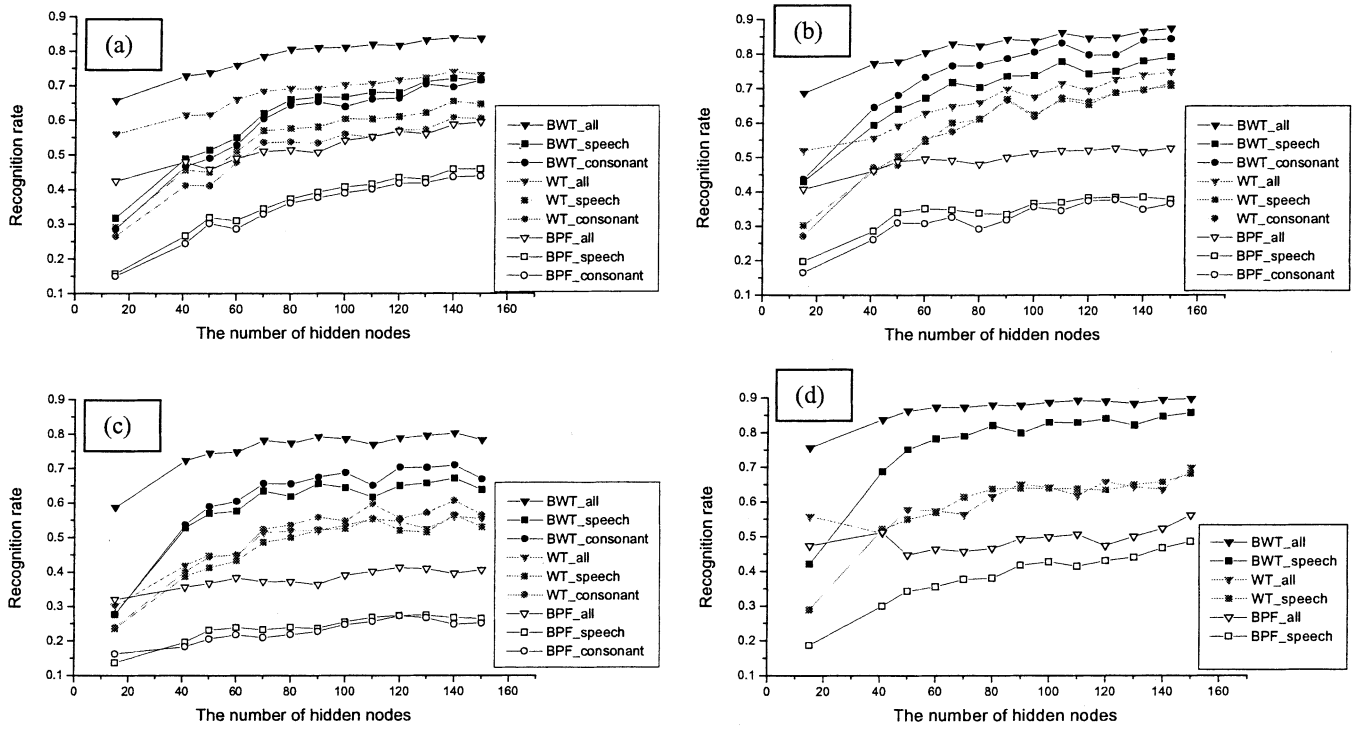

Fig. 2. The recognition rates of the consonants in the contexts: (a) /aCa/, (b) /iCi/, (c) /uCu/, and (d) the vowels versus the number of hidden nodes. The triangles, squares, and circles are the recognition rates for all the frames, speech frames, and consonant frames, respectively. The black, gray, and white symbols represent the results of BWT, WT, and BPF, respectively.

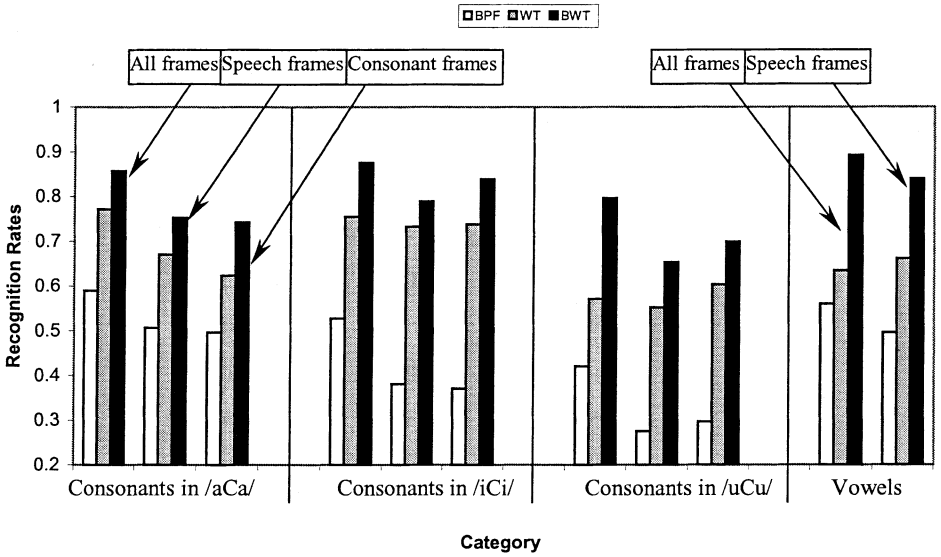

Fig. 3. Saturation recognition rates of consonants and vowels. The black, gray, and white symbols represent the results of BWT, WT, and BPF, respectively.

number of hidden nodes are plotted in Fig. 2(a)–(d), respectively. In Fig. 2, the triangles, squares, and circles represent the recognition rates for all the frames, speech frames, and consonant frames, respectively. And the black, gray, and white symbols represent the results obtained by BWT, WT, and BPF, respectively. From Fig. 2, it is clear that the recognition rates for both consonants and vowels increase with the increasing number of the hidden nodes, and they show asymptotic performance when the number of hidden nodes is large enough.

The saturation recognition rates of the consonants and vowels are shown in Fig. 3, in which the black, gray, and white bars are results of BWT, WT, and BPF, respectively. From Fig. 3, we can see that the recognition rates of the materials processed by BWT are highest.

An example of training performances of the network with 41 hidden nodes for consonants in the contexts: /aCa/, /iCi/, and /uCu/, and for vowels is shown in Fig. 4(a)–(d), respectively. In Fig. 4, the solid and dotted lines are the RMS errors versus

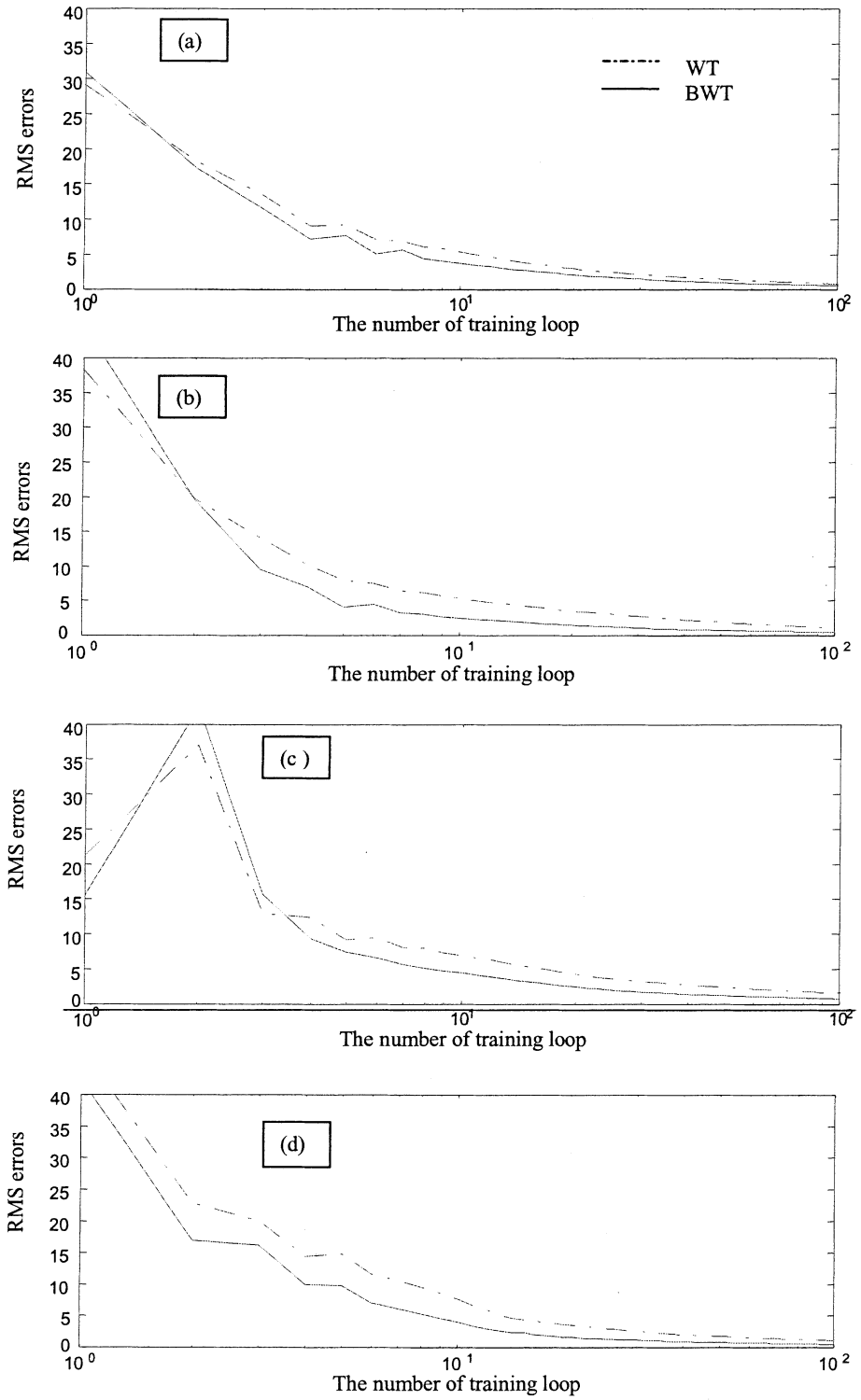

Fig. 4. Training performances of the network with 41 hidden nodes for consonants in the contexts: (a) /aCa/, (b) /iCi/, (c) /uCu/, and (d) for vowels. The solid and dotted lines are the RMS errors for the signals processed by BWT and WT, respectively.

the number of training loops for signals processed by BWT and WT, respectively, where the RMS error is defined as

$$\sqrt{\sum_k (y_k - \overline{y}_k)^2/(K-1)}$$

where y_k and $\overline{y}_k$ are reconstructed and original signals, respectively. Fig. 4 shows that when the input signals were decomposed by BWT, the network converges faster and converges at a lower RMS error level.

Another result of this experiment is that the time resolution of BWT for both consonants and vowels is higher than that of WT. Example outputs from the network with 41 hidden nodes for the seven consonants in the context /iCi/ and for the 11 vowels are shown in Fig. 5(a) and (b), respectively. The network outputs

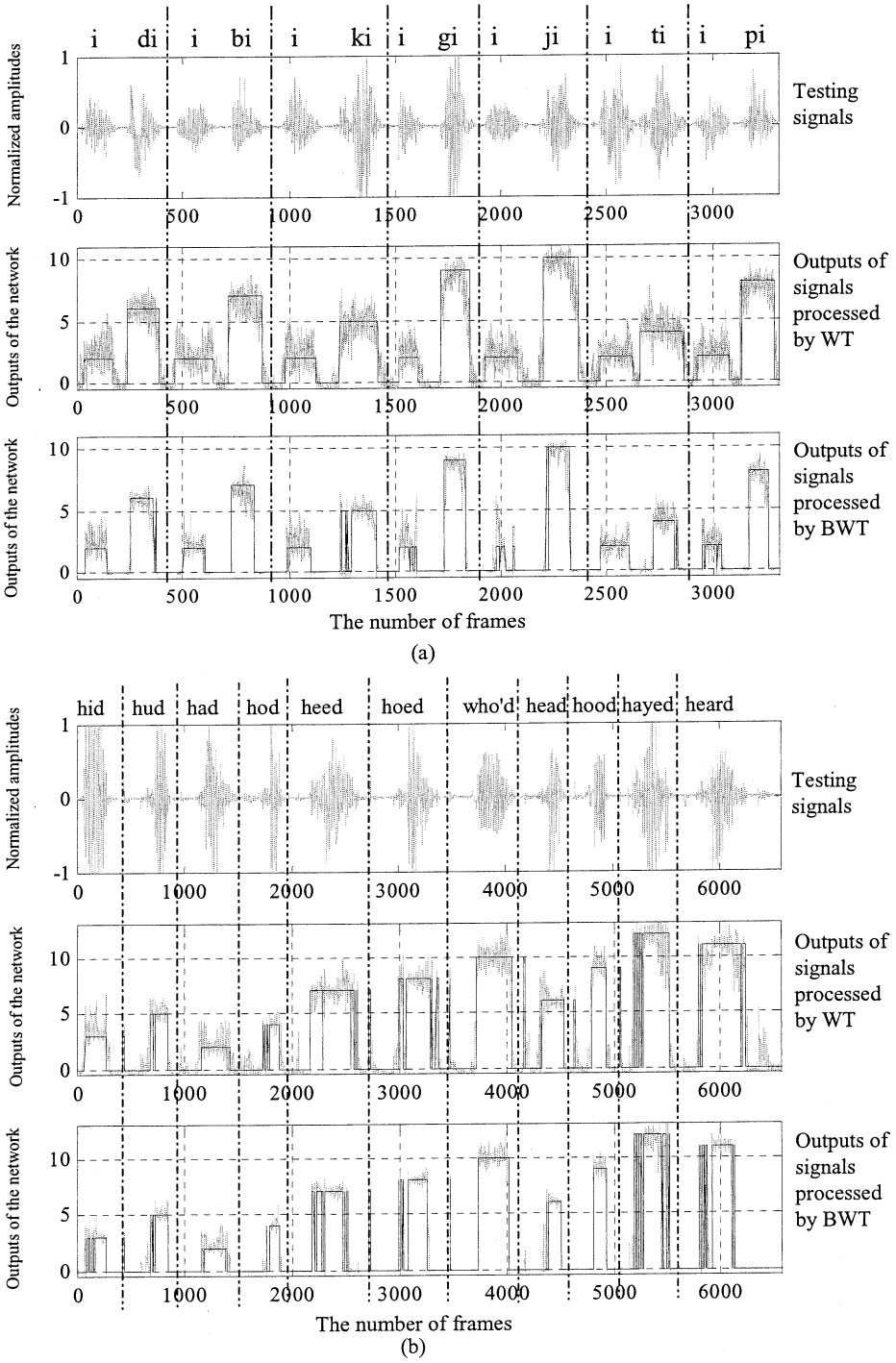

Fig. 5. Outputs of the network for (a) the consonants in the context /iCi/ and (b) vowels versus the number of frames. The length of frame is 44; and the sampling frequency is 11 025 Hz. In (a) and (b), the first subplots are the testing signals; and the second and third subplots are the outputs from the network for the signals processed by WT and BWT, respectively. In the second and third subplots, the black lines are the standard outputs, and the gray lines are the outputs of the networks.

for the consonants in the contexts /aCa/ and /uCu/ are not shown here, because they look similar to the outputs in the context /iCi/. In both Fig. 5(a) and (b), the first subplots are the testing signals; and the second and third subplots are the outputs from the network with the input signals decomposed by WT and BWT, respectively. In the second and third subplots, the black lines are the standard outputs automatically obtained by the program described in the Section III-B, and the gray lines are the outputs of the networks for the corresponding frames. By counting the number of speech frames in Fig. 5, we can see that in the results of BWT, the number of the speech frames for both vowels and consonants is smaller than that in the results of WT. When the number of hidden nodes is 41, we observed that the total number of speech frames for the seven consonants in all the three contexts for BWT is only about 77.00% of that for WT, and that for the 11 vowels is 71.92%.

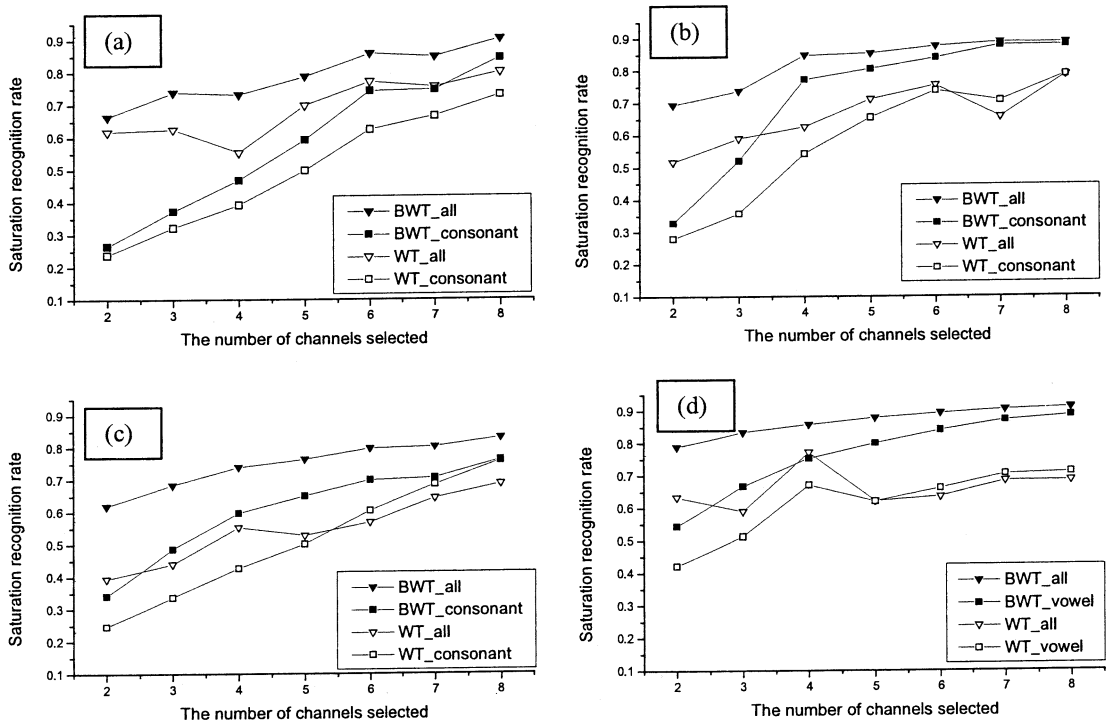

Fig. 6. Saturation recognition rates verse the number of selected channels for the seven consonants in: (a) /aCa/, (b) /iCi/, (c) /uCu/, and (d) for the 11 vowels. The triangles and squares are the recognition rates for all the frames and consonant frames, respectively. And the black and white symbols represent the results of BWT and WT, respectively.

TABLE II
THE NUMBER OF REQUIRED CHANNELS FOR 70% RECOGNITION RATE

	/aCa/		/iCi/		/uCu/		/hvd/	
	All-frames	Consonants	All-frames	Consonants	All-frames	Consonants	All-frames	Vowels
BWT	3	6	3	4	4	6	2	4
WT	6	8	5	6	>8	8	>8	7

C. Experiment 2

1) Goals and Method: This experiment tested whether using BWT to replace WT can reduce the number of required channels. All the previous mentioned materials were included as the experimental materials. In this experiment, according to the protocol (a), we tested the speech recognition rates as a function of the number of channels selected. The simulation was conducted on the same neural network. In the experiment, besides increasing the number of hidden nodes, we also kept increasing the number of selected channels from two to eight. For each fixed number of channels selected, we calculated the saturation recognition rate according to (6).

2) Results: The results of the saturation recognition rates versus the number of channels selected for the seven consonants in the contexts /aCa/, /iCi/, and /uCu/ and for the 11 vowels are plotted in Fig. 6(a)–(d), respectively. In Fig. 6, the triangles and squares signify the saturation recognition rates for all the frames and consonant/vowel frames, respectively. The black and white symbols represent the results of BWT and WT, respectively. Fig. 6 implies that regardless of the number of channels selected (from two to eight), the saturation recognition rates obtained by BWT is always higher than that obtained by WT.

Table II shows that the number of channels required for 70% correct phoneme recognition is less for BWT than for WT. Considering the worst cases, which are for the consonants in the context /aCa/ and /uCu/, at least two channels can be saved by using BWT to replace WT.

D. Experiment 3

1) Goals and Method: This experiment compared the noise tolerance of BWT and WT. The seven consonants in the context /iCi/ were the testing materials. In this experiment, a uniform distribution noise at different signal-to-noise ratio (SNR) level was added to the original signals. In this experiment, SNR is defined as

$$10\log\left(\frac{\sum_k \overline{x}_k^2}{\sum_k n_k^2}\right)$$

where $\overline{x}_k$ is the original signal recorded by the PC, which is viewed to be free of noise, and n_k is the noise. The same network recognized the contaminated signals. Equation (6) was fitted to the recognition rates. Besides calculating the saturation recognition rate y_0, we also calculated the critical number of the hidden

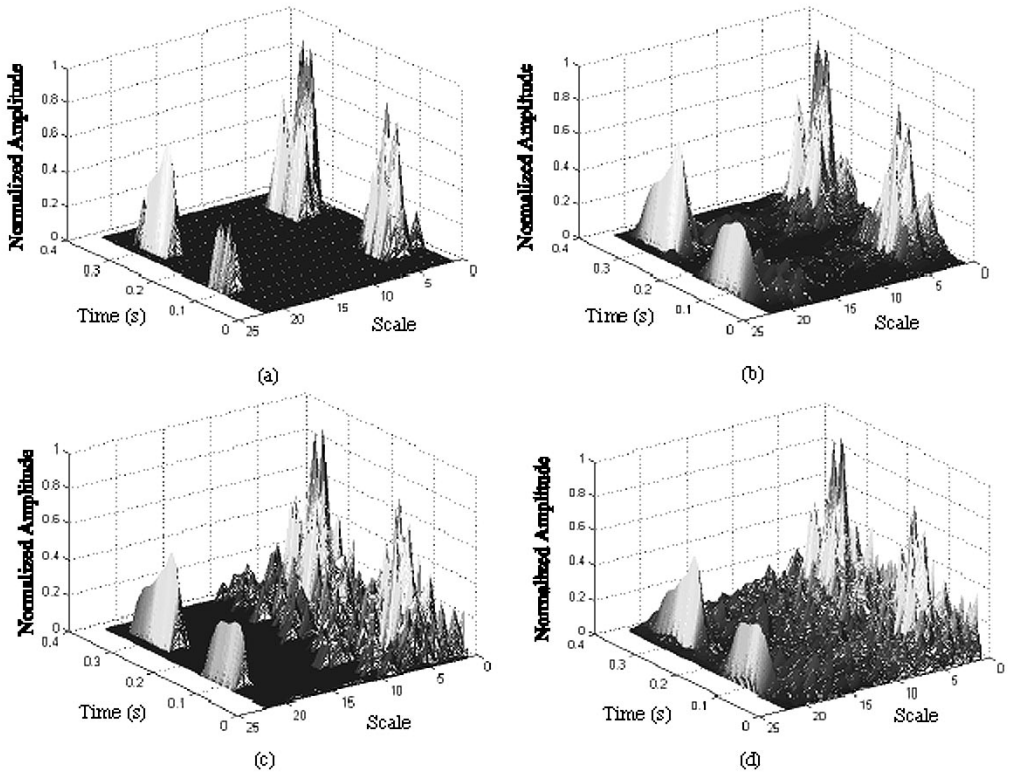

Fig. 7. Time-frequency representation results of /ipi/ with the SNR equal to $+\infty$ processed by (a) BWT and (b) WT, respectively, and with the SNR equal to 0.46 dB processed by (c) BWT and (d) WT, respectively.

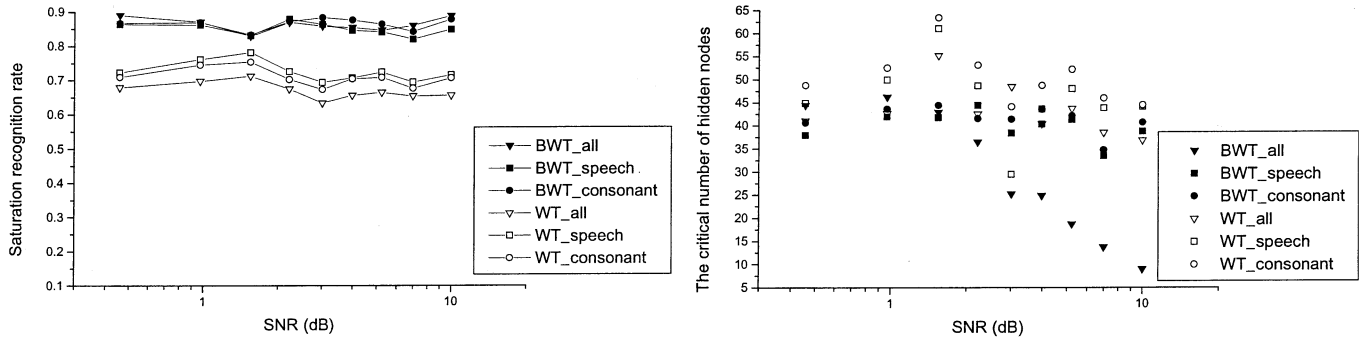

Fig. 8. Performance of network on (a) saturation recognition rates and (b) the critical numbers of the hidden nodes versus SNR. The triangles, squares, and circles are the recognition rates for all the frames, speech frames, and consonant frames, respectively. The black and white symbols represent the results of BWT and WT, respectively.

nodes, $\hat{n}$, which corresponds to asymptotic performance of the network, where $\hat{n} = -\tilde{n}\ln[(y_0/A)((1/\sqrt{2})-1)] + n_0$. We supposed that for a simple signal pattern in the time-frequency domain, less number of hidden nodes is required for the network to achieve the asymptotic performance, and higher recognition rate can be achieved.

2) Results: Fig. 7(a)–(d) shows the time-frequency representation results of /ipi/ with SNR equal to $+\infty$ and 0.46 dB decomposed by BWT and WT, respectively. Comparing Fig. 7(c) and (d) for /ipi/ with SNR equal to 0.46 dB, the most significant difference is found in the low-frequency range, where in the representation of WT, the original signals are embedded in the noise; however, there is still very clean signal in the representation result of BWT.

Fig. 8(a) and (b) shows the saturation recognition rates and the critical numbers of the hidden nodes versus SNR, respectively. In Fig. 8, the triangles, squares, and circles are the results for all the frames, speech frames, and consonant frames, respectively. Fig. 8(a) shows that the saturation recognition rates obtained by BWT (filled symbols) are always higher than those obtained by WT (open symbols). And from Fig. 8(b), we can see that the critical numbers for BWT in each category are always smaller than those for WT, with the exception of the critical number with SNR equal to 3.0 dB for speech frames.

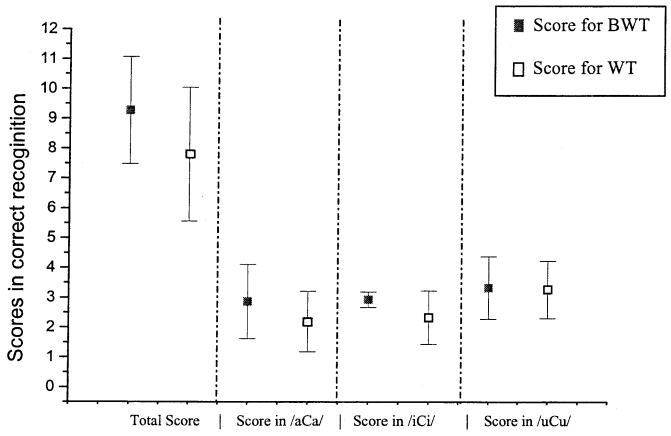

Fig. 9. The results scored in correctly recognized number and their deviations. The solid and open squares are the total scores and the scores in the contexts /aCa/, /iCi/, and /uCu/, for BWT and WT, respectively.

E. Experiment 4

1) Goals and Method: This experiment tested whether there was significant improvement in speech intelligibility between BWT and WT processing schemes.

Fifteen graduate students in the Chinese University of Hong Kong with normal hearing volunteered as subjects; and five consonants, /t k d p g/, in the three contexts were used as the testing materials. In order to avoid the performance gap caused by the distortion introduced by the nonlinearity of BWT, the protocol Fig. 1(b) was used, where the signal was reconstructed by inverse transforms, and the number of channels selected was four.

In each test, the subjects were first given a training session. In this session, the subjects listened to the testing materials to gain familiarity with them. Then in the testing session, they were required to write down the consonants that they had heard. The testing materials were played at about 23 words/min, continuously and randomly, on a PC equipped with SoundMAX Integrated Digital Audio soundcard. The subjects listened to the consonants through an earphone at a comfortable level in a lab that was not sound proof. The testing materials could be played a second time if necessary.

2) Results: The results, scored as correctly recognized number of consonants, along with standard deviations are shown in Fig. 9. Since there are five consonants in each context, the highest score in each context is five. The total score is the sum of the scores in the three contexts. In Fig. 9, the solid and open squares are the scores for BWT and WT, respectively. Fig. 9 shows that in all the three contexts, the scores for BWT are higher than those for WT.

The two population paired t-test was performed to evaluate whether there was a significant difference between BWT and WT. Our results show that there is significant difference between the methods BWT and WT in the contexts /aCa/ and /iCi/, where the t-test results were ($t = -2.65$, $p = 0.019$) and ($t = -0.236$, $p = 0.033$), respectively. While in the context /uCu/, no significant difference is found ($t = -0.43$, $p = 0.67$). The t-test result of total scores shows that when using BWT to process the signal, the performance is significantly improved ($t = -4.36$, $p = 0.000\,65$).

IV. Discussion and Conclusion

In this paper, a new time-frequency analysis method, BWT, was applied to speech processing in cochlear implants. BWT is generated by introducing a new parameter T into the WT [21]. From the point of view of signal processing, BWT can be viewed as a special subset of adaptive WT (AWT), or an adaptive BPF. The distinguishing characteristic of BWT is that it does not use an entropy function as a criterion, but uses the active control mechanism stemming from the auditory system to adjust the wavelet function. One reason for carrying out wavelet function adjustment according to the active control mechanism in a bio-system is that it is hard to know which entropy function of the bio-system is used [21]. The performance of BWT depends highly on the T-function. In this paper, the T-function stems from the auditory model for healthy cochlea; therefore, it is especially appropriate for speech signal processing in cochlear implants.

A. Recognition Rate

We used a series of neural networks, each of which had a different number of hidden nodes, to detect the saturation recognition rates of consonants and vowels that were pre-processed by BWT and WT, respectively. This method is different from the other methods in which recognition rates were obtained based on CI subjects [13], [23]–[25]. Our results suggest a probable optimal performance. However, these results cannot directly be compared with those obtained from the implant subjects. We decided to use neural networks since the recognition rate obtained by a CI subject can be attributed to many other confounding factors besides the signal processing method, including surviving ganglion cells, electrode insertion depth, and the status of the subjects, etc. [26]. With a neural network, however, the judgment of recognition rates is objective. When keeping the training and testing materials and methods as the same, the only other factor that will influence the recognition rate besides the signal processing method is the network itself. In our experiments, we calculated the saturation recognition rates to eliminate the impact on the performance caused by the size of the neural network. The other reason for the usage of a neural network is that this work is a preliminary investigation on the application of BWT to speech signal processing in CI. Further experiments on CI subjects are in our research plan. The preliminary experimental results on saturation recognition rates, as shown in Figs. 2 and 3, indicate that the BWT processing improves the recognition rate for both consonants and vowels compared with the WT and BPF schemes. The training performance of the network, as shown in Fig. 4, further shows that it is easier to train a neural network to recognize the signals processed by BWT than by WT.

B. Contradiction Between Time and Frequency Resolution

One reason for the improvement in recognition rates is attributed to the better tradeoff between time and frequency resolutions achieved by BWT than that achieved by WT. As shown in Fig. 5(a) and (b), the number of speech frames of signals processed by BWT is smaller than that processed by WT. This im-

plies that a higher time resolution is achieved by BWT, possibly leading to a reduction of the stimulation duration of the electrodes. Furthermore, Table II shows that in order to keep a 70% recognition rate, a fewer number of channels are required for BWT than that for WT. The reduction of required channel number and average stimulation duration could alleviate the contradiction between the spread of excitation and the fine frequency-coding requirement.

C. Tolerance of Noise

Another reason for the improvement of recognition rates is that BWT is more tolerant of noise than WT. As shown in Fig. 8, independent of the SNR level, the most significant improvement is found in the recognition rates for all the frames, which indicates that the improvement of recognition rate can be attributed to the increase in recognition rate for nonspeech frames. This is consistent to the result found in experiment 3 that BWT can reduce the low-frequency noise as shown in Fig. 7.

Fig. 8 shows that no significant reduction in recognition rates was found for both BWT and WT with an increase in the size of the network. Further, in Fig. 8(b), the critical number of hidden nodes for BWT is smaller than that for WT. Combining Fig. 8(a) and (b), we can see that with a smaller network, BWT achieves better saturation recognition rates than WT. Therefore, BWT is more tolerant of noise than WT.

D. Subject Performance Versus Neural Network Performance

Although the neural network performance is objective, the working mechanisms of a neural network are simple and different from that of the human brain. In order to predict the speech intelligibility obtained by BWT and WT for a CI subject, we further conducted a panel judging on consonants by normal hearing subjects. The usage of normal hearing subjects is, again, to eliminate the influence caused by the other confounding factors of CI subjects. Our experimental results shown in Fig. 9 indicate that there is improvement of sound intelligibility of consonants in all the three contexts. Comparing the results shown in Figs. 6 and 9, we can see that the perceptual results with four channels selected are similar to the results obtained by the neural network when the number of channels selected is no less than 8. In both of them, significant improvement was found for the consonants in the contexts /aCa/ and /iCi/, and no significant improvement was found in the context /uCu/. This consistency between subject's performance and network performance implies that the prediction of recognition rates by neural network may be meaningful.

However, an inconsistency between the results of panel judgment and the network performance was also found. In the panel judgment, the highest scores are obtained on the consonants in the context /uCu/, while in the performance of the network, the highest recognition rates were obtained in the context /iCi/. The reason for this inconsistency may include many factors, including the different working mechanism between networks and human brain. To verify BWT performance in CI application, experiments on CI subjects will be conducted in our future studies.

E. Conclusions

The BWT is a bio-model-based signal processing method, whose performance is highly dependent on the T-function stemming from the auditory model. Results obtained from the present study demonstrate the following.

1) Based on neural network simulations, a better recognition rate in processing the speech signals can be achieved by BWT in comparison with WT and BPF.
2) BWT has better tradeoff between time and frequency resolutions, which would result in the reduction of the number of channels required for CIs and the reduction of the stimulation duration required for a word. This property of BWT might lead to alleviate the conflict between spread of excitation and fine frequency coding in CIs.
3) BWT is more tolerant of the noise than WT.

In order to test the performance of BWT in speech intelligibility, we have further conducted a penal judging experiment on consonants with 15 normal hearing subjects. The results of the experiment show that there is a significant improvement on recognition rates using BWT. These experimental results, supported by those obtained from our neural network simulations, show that the BWT has a potential to reduce the performance gap between CI listeners and people with a normal hearing capability. Given this result, the BWT approach to processing speech signal will be beneficial and further investigation on CI subjects is warranted.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for many useful suggestions. They would also like to thank D. Ison for English corrections and editing.

REFERENCES

[1] S. U. Ay, F. G. Zeng, and B. J. Sheu, "Hearing with bionic ears: Speech processing strategies for cochlear implant devices," *Circuits Devices*, pp. 18–23, May 1997.
[2] "The Bionic Ear Institute Annual Report—1999," The Bionic Ear Institute, Melbourne, Australia, 1999.
[3] J. P. Rauchecher, "Making brain circuits listen," *Science*, vol. 285, pp. 1686–1687, 1999.
[4] F. G. Zeng and J. J. Galvin, III, "Amplitude mapping and phoneme recognition in cochlear implant," *Ear Hearing*, vol. 20, pp. 60–74, 1999.
[5] W. House and J. Urban, "Long term results of electrode implantation and electronic stimulation of the cochea in man," *Ann. Otol. Rhinol. Laryngol.*, vol. 82, pp. 504–517, 1973.
[6] W. House and K. Berliner, "Cochlear implants: Progress and perspectives," *Ann. Otol. Rhinol. Laryngol.*, vol. Sipple. 91, pp. 1–124, 1982.
[7] P. C. Loizou, "Signal-processing techniques for cochlear implants," *IEEE Eng. Med. Biol. Mag.*, pp. 34–46, May 1999.
[8] G. M. Clark, "The University of Melbourne-Nucleus multi-electrode cochlear implant," *Adv. Oto-Rhino-Laryngol.*, vol. 38, pp. 1–189, 1987.
[9] D. Eddington, "Speech discrimination in deaf subjects with cochlear implants," *J. Acoust. Soc. Amer.*, vol. 102, pp. 2403–2411, 1980.
[10] D. Eddington, W. Dobelle, D. Brachman, M. Mladevosky, and J. Parkin, "Auditory prosthesis research using multiple intracochlear stimulation in man," *Ann. Otol. Rhinol. Larynhol.*, vol. 87, pp. 1–39, 1978.
[11] D. K. Kessler, "The Clarion multi-strategy cochlear implant," *Ann. Otol. Rhinol. Laryngol.*, vol. 108, pp. 8–16, 1999.
[12] P. Seligman and H. McDermott, "Architecture of the Spectra 22 speech processor," *Ann. Otol. Rhinol. Laryngol.*, vol. 104, pp. 139–141, 1995.
[13] C. P. Behrenbruch, "SNR improvement, filtering and spectral equalization in cochlear implants using wavelet techniques," in *Proc. 2nd Int. Conf. Bioelectromagnetism*, 1998, pp. 61–62.

[14] K. Nie, N. Lan, and S. Gao, "Implementation of CIS speech processing strategy for cohlear implants by using wavelet transform," in *Proc. ICSP'98*, pp. 1395–1398.
[15] A. Ryan and P. Dallos, "Effects of absence of cochlear outer hair cells on behavioral auditory threshold," *Nature*, vol. 253, pp. 44–46, 1975.
[16] M. C. Liberman and L. W. Dodds, "Single-neuron labeling and chronic cochlear pathology. III. Stereocilia damage and alterations of threshold tuning curves," *Hear. Res.*, vol. 16, pp. 55–74, 1984.
[17] P. C. Loizou, "Introduction to cochlear implants," *IEEE Eng. Med. Biol. Mag.*, pp. 32–41, Jan. 1999.
[18] J. Yao and Y. T. Zhang, "Modeling of equivalent nonlinear resistance and capacitance of active Cochlea," presented at the EMBC'2001, Istanbul, Turkey.
[19] J. Yao, Y. T. Zhang, and L. Zheng, "A preliminary report on the modeling of otoacoustic emissions," presented at the 3rd Int. Workshop Biosignal Interpretation, Chicago, IL, 1999.
[20] J. Yao and Y. T. Zhang, "A study on inhomogineous characteristics of Cochlea for spontaneous otoacoustic emissions," presented at the IEEE EMBS-BMES, Atlanta, GA, 1999.
[21] ——, "Bionic wavelet transform: A new time-frequency method based on an auditory model," *IEEE Trans. Biomed. Eng.*, vol. 48, pp. 856–863, Aug. 2001.
[22] ——, "A fast algorithm for continuous bionic wavelet transform," presented at the M2VIP, Hong Kong, 2001.
[23] B. Wilson, D. Lawson, and M. Zerbi, "Advances in coding strategies for cochlear implants," *Adv. Otolaryngol. Head Neck Surg.*, vol. 9, pp. 105–129, 1995.
[24] E. Wallenberger and R. Battmer, "Comparative speech recognition results in eight subjects using two different coding strategies with the Nucleus 22 channel cochlear implant," *Br. J. Audiol.*, vol. 25, pp. 371–380, 1991.
[25] M. Skineer, G. Clark, L. Whitford, P. Seligman, and S. Staller, "Evaluation of a new spectral peak coding strategy for the Nucleus 22 channel cochlear implant system," *Amer. J. Otol.*, vol. 15, Suppl. 2, pp. 15–27, 1994.
[26] P. Loizou, C. M. Dorman, O. Poroy, and T. Spahr, "Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution," *J. Acoust. Soc. Amer.*, vol. 108, pp. 2377–2388, 2000.

Jun Yao (S'99–A'01) was born in China, 1972. She majored in biomedical engineering and received the B.E. degree in 1995 from the Chongqing University, China, M.E. degree in 1998 from the Southeast University, China, and Ph.D. degree from the Chinese University of Hong Kong, Hong Kong.

She now works at the Rehabilitation Institute of Chicago, Northwestern University. Her current research interests include system modeling, signal processing especially on time-frequency analysis, artificial neural networks, and cochlear implants.

Yuan-Ting Zhang (M'90–SM'93) received his M.Sc. degree from Shan-Dong University, China, in 1981 and his Ph.D. from the University of New Brunswick, Fredericton, NB, Canada, in 1990.

He was a Research Associate and Adjunct Assistant Professor at the University of Calgary between 1989 and 1994. He is currently a Professor and Director of Joint Research Centre for Biomedical Engineering at the Chinese University of Hong Kong (CUHK). At CUHK, he has developed and teaches courses including biomedical modeling, medical instruments and sensors, and telemedicine techniques and applications. His research activities have focused on the development of bio-model-based signal processing techniques to improve the performance of medical devices and biosensors, particularly for telemedicine. His work has been published in several books, over 20 scholarly journals, and many international conference proceedings.

Dr. Zhang served the Technical Program Chair of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS). He was the Chairman of Biomedical Division of Hong Kong Institution of Engineers in 1996/1997 and 2000/2001, an AdCom member of IEEE-EMBS in 1999, and the Vice-President of the IEEE-EMBS in 2000 and 2001. He serves currently as an Associate Editor of IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, an Editorial Board member for the Book Series of Biomedical Engineering published by Wiley and IEEE Press, and an Associate Editor of IEEE TRANSACTIONS ON MOBILE COMPUTING.