UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Institute of Sound and Vibration Research

Wavelet Filter Banks for Cochlear Implants

by

Siriporn Dachasilaruk

Thesis for the degree of Doctor of Philosophy

December 2014

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Institute of Sound and Vibration Research

Doctor of Philosophy

Thesis for the degree of Doctor of Philosophy

WAVELET FILTER BANKS FOR COCHLEAR IMPLANTS

Siriporn Dachasilaruk

Cochlear implant (CI) users regularly perform as well as normal-hearing (NH) listeners in quiet conditions. However, CI users have reduced speech perception in noise. CI users suffer more in terms of speech intelligibility than NH listeners in the same noisy environment. Speech coding strategies with noise reduction algorithms for CI devices play an important role, allowing CI users to benefit more from their implants. This thesis investigates a wavelet packet-based speech coding strategy with envelope-based noise reduction algorithms to enhance speech intelligibility in noisy conditions.

The advantages of wavelet packet transforms (WPTs), in terms of time-frequency analysis, the sparseness property, and low computational complexity, might make WPT appropriate for speech coding and denoising in CI devices. In cases with an optimal set of parameters for a wavelet packet-based speech coding strategy, the 23- and 64-band WPTs with sym8 and frame length of 8 ms were found to be more suitable than other combinations for this strategy. These parameters can optimise speech intelligibility to benefit CI users. However, both the standard ACE strategy and the wavelet packet-based strategy provided almost the same results in either quiet or noisy conditions.

Cases using envelope-based denoising techniques in a wavelet packet-based strategy, namely time-adaptive wavelet thresholding (TAWT) and time-frequency spectral subtraction (TFSS) were developed and evaluated by objective and subjective intelligibility measures and compared to ideal binary masking (IdBM) as a baseline for denoising performance. IdBM can restore intelligibility to nearly the same level as NH listeners in all noisy conditions. Both TAWT and TFSS showed slight intelligibility improvements in some noisy conditions. This may result from noise estimation in denoising techniques. Noise level may be under- or overestimated, and this results in distortion in enhanced speech and difficult in speech discrimination.

Both objective and subjective intelligibility measures can predict the trend of the performance of different denoising techniques for CI users. However, NH listeners can achieve better intelligibility at higher SNR levels without noise reduction, since they are less sensitive to noise but more sensitive to speech distortion when compared to CI listeners. Therefore, denoising techniques may work well for CI users in further investigations.

Contents

List of tables

List of figures

DECLARATION OF AUTHORSHIP

I, SIRIPORN DACHASILARUK

declare that the thesis entitled

WAVELET FILTER BANKS FOR COCHLEAR IMPLANTS

and the works presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as conference papers and listed in Appendix A:

Signed: …….

Date:……..

Acknowledgements

My thesis would have been impossible without the support of my supervisors, a number of colleagues and staff from Institute of Sound and Vibration (ISVR), and others.

First of all, I am very grateful to my supervisors, Dr Stefan Bleeck and Professor Paul White, who provided excellent guidance and patience in listening to and reading my work. Dr Stefan Bleeck constantly encouraged and guided me with research methodology, ideas and organization. Professor Paul White provided valuable experience and insight in signal processing and acoustic research, as well as relaxed supervision.

I would like to thank Professor Shouyan Wang and Professor Mark Lutman for inspiring this research in many ways. Thanks to Professor David Simpson, Dr Carl Verschuur, and Dr Richard Turner for reviews and useful comments on my PhD research.

I would like to extend my gratitude to many great colleagues and a number of staff members for their friendship and support and for sharing knowledge and experiences, including Dr Hongmei Hu, Dr Gouping Li, Dr Seon Man Kim, Dr. Al Mander, Dr Konda Mammon, Dr Nuthnapa Triepaischajonsak, Dr Nuttarut Pananandа, Chokchai Yachusri, Khemapat Tontiwattanakul, Vorrapath Kokaew, Kathryn Earl, and ISVR technicians and support staff. I would also like to thank others around the world for their help during my time in Southampton.

My special thanks go to my parents and my relatives for their understanding during my PhD study abroad. Finally, I would very much like to thank the Thai Government Scholarship programme for giving me the opportunity to study for a PhD in the United Kingdom.

List of symbols

$L^2(\mathbb{R})$	Finite energy functions
$\mathbb{Z}$	Integers
$\mathbb{R}$	Real numbers
f_c	Centre frequency
f_s	Sampling rate
f	Frequency in Hertz
Δf	Bandwidth
f_l	Lower frequency
f_u	Upper frequency
$\phi(t)$	Scaling function
$\psi(t)$	Wavelet function
c_j	Approximation coefficients at level *j*
d_j	Detail coefficients at level *j*
w	Wavelet coefficients
h	Coefficients of lowpass filter for wavelet decomposition
g	Coefficients of highpass filter for wavelet decomposition
$\tilde{h}$	Coefficients of lowpass filter for wavelet reconstruction
$\tilde{g}$	Coefficients of highpass filter for wavelet reconstruction
J	Number of levels of wavelet decomposition and reconstruction
j	Level of wavelet decomposition and reconstruction
K	Number of wavelet coefficients in each subband
L	Filter length
M	Number of samples per frame
*	Convolution operation
$\downarrow 2$	Downsampling by 2
$\uparrow 2$	Upsampling by 2
ξ	A *priori* SNR

γ	A *posteriori* SNR
β, α	Weighting factors
$G(,)$	Gain function
$T_S(,)$	Soft-thresholding gain function
λ	Threshold values
$\wedge$	Estimated parameters

List of abbreviations

ABF	Adaptive beamforming
ACE	Advanced combination encoder
AGC	Automatic gain control
AI	Articulation index
AI-ST	Short-term articulation index
ANOVA	Analysis of variance
BKB	Bamford-Kawal-Bench
BM	Basilar membrane
BPF	Bandpass filter
BWT	Bionic wavelet transform
CA	Compressed analogue
CASA	Computational auditory scene analysis
CB	Critical bandwidth
CF	Characteristic frequency
CI	Cochlear implant
CIS	Continuous interleaved sampling
CSII	Coherence-based speech intelligibility index
CWT	Continuous wavelet transform
ℂWT	Complex wavelet transform
DAU	Dau Auditory model
DFT	Discrete Fourier transform
DWT	Discrete wavelet transform
ERB	Equivalent rectangular bandwidth
FAME	Frequency amplitude modulation encoding
FT	Fourier transform
FFT	Fast Fourier transform
FIR	Finite impulse response
fwsSNR	Frequency-weighted segmental SNR
HA	Hearing aid
HI	Hearing impaired
HL	Hearing level

ITFS	Ideal time frequency segregation
IdBM	Ideal binary masking
ICA	Independent component analysis
IIR	Infinite impulse response
LF	Leakage factor
MAD	Median of the absolute deviation
MAP	Speech processor program
Mel	Melody
MMSE	Minimum-mean square error
MPEAK	Multi-peaks
MRA	Multiresolution analysis
MSE	Mean-square error
MTF	Modulation transfer function
NCM	Normalised covariance metric
NH	Normal-hearing listener
NSEC	Normalised subband envelope correlation
PACE	Psychoacoustic advanced combination encoder
PCA	Principle components analysis
PESQ	Perceptual evaluation of speech quality
SNR	Signal-to-noise ratio
SPEAK	Spectral peak
SPL	Sound pressure level
SRT	Speech reception threshold
ST	Scala tympani
STI	Speech transmission index
STFT	Short-time Fourier transform
STOI	Short-time objective intelligibility measure
SWT	Stationary wavelet transform
T-F	Time-frequency
TAWT	Time-adaptive wavelet thresholding
TEO	Teager energy operator
TFSS	Time-frequency spectral subtraction
VAD	Voice activity detection

WPT	Wavelet packets transform
WT	Wavelet transform

Chapter 1: Introduction

1.1 Contribution to knowledge

Most cochlear implant (CI) users perform well in quiet listening conditions, and many can achieve more than 80% speech recognition scores. However, speech recognition scores are significantly degraded in noisy listening conditions. Enhancing speech intelligibility for CI users in noise is a major goal for improving CI systems. The speech coding strategies in cochlear devices play an extremely important role and can influence the overall performance of the CI device in order to greatly benefit CI users' communicative potential (Loizou, 1998).

Generally, the greater the level of background noise, the lower the capability in terms of speech intelligiblity. Since the speech signal contains highly redundant information, if some parts of speech signal are masked by noise in a moderately noisy environments, other parts of speech may still contain useful information, and speech intelligiblity will be sufficiently maintained for normal-hearing (NH) listeners (Kokkinakis et al., 2012). However, speech intelligiblity is poor in noise for CI users, at least, because of the limited number of electrodes, the spectral mismatch from the frequency-to-electrode allocation, and the interaction between electrodes (Stickney et al., 2004). There is poorer performance with nonstationary noise (e.g. babble noise) than stationary noise (e.g. speech-shaped noise) for both NH listeners and CI users (Qin and Oxenham, 2003; Stickney et al., 2004).

Some noise reduction algorithms have been proposed for CI users, for both multi-microphone noise reduction (Vanhoesel and Clark, 1995; Wouters and Vanden Berghe, 2001; Spriet et al., 2007) and single-microphone noise reduction. Algorithms for multi-microphone noise reduction include adaptive beamforming algorithms (Vanhoesel and Clark, 1995; Wouters and Vanden Berghe, 2001; Spriet et al., 2007) and two-microphone spectral subtraction (Kallel et al., 2012). However, multi-microphone noise reduction is undesirable, cosmetically unappealing, and computationally complex.

Single-microphone noise reduction algorithms can be divided into those using a pre-processing approach and those adopting an envelope-based approach. Pre-

processing approaches include spectral subtraction (Yang and Fu, 2005; Verschuur et al., 2006) and the subspace method (Loizou et al., 2005), and can bring benefits for stationary noise, but these are not guaranteed for nonstationary noises. The envelope-based approach can enhance in noisy speech by using envelope-weighting or envelope-selection in each channel. However, some techniques involve more complicated procedures (Li, 2008), and others require prior knowledge of the clean speech and noise information before both are mixed (Hu and Loizou, 2008). They are not suitable for real-time implementations in real-world situations.

The past decades have seen the rapid development of wavelet analysis in many applications (Peng and Chu, 2004; Mallat, 2009). The wavelet transform of most real-world signals tends to be dominated by a few large coefficients (Donoho and Johnstone, 1994), which constitutes the so-called sparseness property. This sparsity of wavelet representation is essential to the performance of noise reduction and data compression. In addition, wavelet transforms have proven to be successful for the detection and estimation of signals.

Wavelet thresholding is a powerful method for noise reduction. The concept of this method is based on thresholding the wavelet coefficients towards zero. Since noise is spread out over all the wavelet coefficients, the sparse representation allows the replacement of noisy coefficients by zero. Wavelet thresholding has been widely applied in the area of speech enhancement, including classical wavelet thresholding (Pinter, 1996; Chen and Wang, 2004), modified wavelet thresholding (Sheikhzadeh, 2001; Ghanbari and Karami-Mollaei, 2006) and combined with other noise reduction algorithms (Hu and Loizou, 2004; Shao and Chang, 2007).

The sparseness property of wavelet transform can reduce the unnecessary information in speech signals to improve in the efficiency of speech coding strategies without impacting the speech intelligibility and quality. This may be particularly attractive for CI system due to the limitations of frequency resolution (the number of electrodes) and temporal resolution (stimulation rate).

A few studies have introduced CI speech coding strategies based on wavelet packets (Behrenbruch and Lithgow, 1998; Nogueira et al., 2006; Gopalakrishna et al., 2010b). Wavelet packets can be easily adapted to approximate the critical bands of the human auditory system in CI design. The wavelet packet-based speech coding strategy

has been successfully produced for real-time implementations. In addition, this strategy yields lower spectral leakage, higher stimulation rate, lower computational complexity (Gopalakrishna et al., 2010b), and better speech intelligibility performance than the ACE strategy for CI users (Nogueira et al., 2006). However, there is still considerable work to be done in the investigation of the utility of wavelet packets for enhancing intelligibility for CI users in noisy listening conditions.

The research question for this work is to determine whether a new speech coding strategy developed by using wavelet packets with noise reduction algorithms, can enhance speech intelligibility for CI users in noisy environments. The main goal of the proposed work is to attempt find the optimal parameters for a wavelet packet-based speech coding strategy, and to evaluate the noise reduction algorithms (i.e. time-frequency spectral subtraction (TFSS) and time-adaptive wavelet thresholding (TAWT)) in the wavelet packet-based speech coding strategy in terms of different types of noise (i.e. speech-shaped noise and babble noise) and different signal-to-noise ratio (SNR) levels (i.e. 0, 5 and 10 dB).

1.2 Cochlear implants

A cochlear implant (CI) is an electronic prosthesis device. It is implanted into the inner ear in order to transmit electrical stimuli to the auditory nerve, restoring partial hearing for individuals with severe and profound hearing losses. CIs are introduced in 1984. Currently, CIs have successfully restored hearing in more than 324,200 hearing-impaired people world-wide (NIDCD, 2014). Using the latest CI, the majority of CI users can score above 80% correct on high context sentences, even without visual cues (Wilson and Dorman, 2008b). Some CI users can communicate without any signing or lip-reading, and some can communicate over the telephone (Loizou, 1998).

Recently, various signal-processing techniques for CI processors have been developed to mimic the function of a healthy cochlea. An understanding of the auditory system and speech perception is essential for speech coding design in CI systems. Comprehensive reviews and accounts of the history of speech-processing strategies for CI systems can be found in a variety of literature (Loizou, 1998; Loizou, 1999; Zeng,

2004; Loizou, 2006; Fan-Gang et al., 2008; Wilson and Dorman, 2008b; Wilson and Dorman, 2008a).

1.2.1 The components of cochlear implants

The main components in all modern cochlear implant systems are illustrated in Figure 1.1 (Fan-Gang et al., 2008). They consist of an ear hook and a microphone (1) to pick up sounds, a battery case and a behind-the-ear external processor (2) to transform the sound into a set of electrical stimuli for the implanted electrode, a radio frequency (RF) transmitter (3) which encodes the set of electrical stimuli into a RF signal and sends it to the antenna inside a headpiece, the internal receiver (4) placed under the skin behind the ear which receives and decodes a RF signal, the stimulator (5) containing an active electronic circuit which converts the signal into electrical currents, sending them along a cable (6) to the electrode array (7). The electrode array stimulates neurons of the auditory nerve (8) connected to the central nervous system, where the electrical impulses are interpreted as sound.

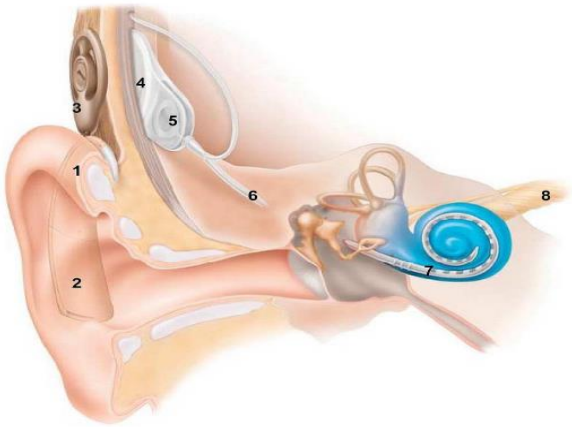


Figure 1.1 Cochlear implant systems (Fan-Gang et al., 2008).

1.2.2 The cochlear function

The auditory system consists of the outer, middle and inner ear. The outer ear consists of the pinna, which captures sound energy and conducts it directly to the ear drum. The middle ear transforms the sound waves into mechanical vibrations and transmits these vibrations to the fluids of the cochlea in the inner ear. The cochlea converts the mechanical signal into neural activity. The neural activity is transmitted to the central auditory system through the auditory nerve of the inner ear, and is translated into the perception of sound.

The hair cells of normal-hearing (NH) persons are activated according to the displacement of the basal membrane (BM). The bending of the hair cells releases an electrochemical substance to directly stimulate the neurons of the auditory nerve in the inner ear. These neurons communicate with the central nervous system and transmit information about speech signals to the brain. Hearing loss results from the destruction of the hair cells in the cochlea, as well as from age-related degeneration.

The hair cells of hearing-impaired (HI) persons may be damaged by certain diseases (e.g. Meniere's disease or meningitis), drug treatments, congenital disorders, and other causes (Loizou, 1998). The largely or completely damaged hair cells lead to a hearing impairment. The greater the number of hair cells that are damaged, the more the person's hearing is impaired. These damaged hair cells define where these neurons cannot transmit auditory information from the BM to the central nervous system.

The concept of a cochlear prosthesis is to bypass the damaged hair cells by stimulating the remaining neurons directly with electrical pulses. The electrical stimulation is directly transmitted through electrodes. The electrode is inserted and placed into the scala tympani (ST) close to the base of the cochlea. Different positions for an electrode array can stimulate different subpopulations of neurons. The neurons at different positions along the length of the cochlea respond to different frequencies of sound. Stimulating an electrode array at the base of the cochlea is consistent with high-frequency sound information, while stimulating an electrode array at the apex of the cochlea represents low-frequency information. When neurons are stimulated, they fire and propagate electrical or neural impulses to the central nervous system.

1.2.3 Speech perception

When a speech sound enters the human ear as a composite waveform, people can perceive and understand the context of speech. Speech perception is considered to be formed from the basic units of speech or language (i.e. phonemes), the smallest unit of sound that is used to form meaningful distinctions between utterances. A combination of phonemes is called a syllable. Words are formed from a combination of syllables. Therefore, if phonemes are changed, the meanings of words are also changed. Speech sounds in the English language (Loizou, 2007) are generally classified into two broad types: vowels and consonants. Vowels are also called monophthongs (single voiced sounds), and a related class of sounds is the diphthong (two voiced sound). Consonants can be divided into six classes: semivowels, whispers, nasals, stops, fricatives and affricates.

Each class of English language sound has unique characteristics, with acoustic cues that can be easily discriminated from other classes. Acoustic cues are essential in accurate phoneme identification (Loizou, 2007). Any acoustic cues that are masked by noise can affect specific phoneme identification, which is reflected in speech intelligibility. Generally, vowels often have low frequencies and relatively high energy. Many consonants have higher frequencies and less energy than vowels and diphthongs (French and Steinberg, 1947). Therefore, the low-energy consonants (e.g. stop) are masked by various noises more easily than the high-energy vowels and semivowels (Chen and Loizou, 2010).

Although NH listeners can recognise speech even with high SNR levels, speech recognition among HI listeners and CI users is much more susceptible to noise (Fu et al., 1998). NH listeners can benefit from the use of redundant information in speech, whereas CI users have perceptual difficulties because of the limitations of frequency resolution, temporal resolution and the amplitude of speech signals which can be transmitted by CI devices (Fu et al., 1998; van Schijndel et al., 2001; Kokkinakis et al., 2012).

Most CI users require much higher SNR than NH listeners, approximately 10–25 dB, to achieve a similar level of speech intelligibility performance in noise (Qazi et al., 2012). In one study, CI users' SNR levels varied between 10 and 15 dB for stationary noises and were equal to 25 dB for nonstationary noise (Kokkinakis et al., 2012). In

other words, speech perception performance in nonstationary noise (i.e. babble noise) is poorer than in stationary noise (i.e speech-shaped noise) for both NH listeners and CI users (Qin and Oxenham, 2003; Stickney et al., 2004).

Speech sounds contain a wide range of frequencies. The spectrum of a speech signal consists of the fundamental frequency (F0), defined as the lowest frequency of a periodic signal. The F0 is perceived by the human ear as pitch. The F0 is the first harmonic, and the other harmonics occur at integer multiples of the F0. The frequency range of F0 is approximately 60–150 Hz for males and 200–400 Hz for females and children (Loizou, 2007). The peaks of the spectral envelope in a speech signal are referred to as formants. The formants can be represented from the spectral envelope and not from the magnitude spectrum. On the other hand, the harmonics can be represented from the magnitude spectrum and not from the spectral envelope. Therefore, the formant frequencies may or may not coincide with one of the harmonics (Loizou, 2007).

The formants provide speech information (Shannon et al., 1995). The first three formants (i.e. F1, F2, and F3), in the frequency range of 0.1 to 4 kHz, contain sufficient information for speech perception (Loizou, 1998). Additionally, information around F1 and F2 is sufficient for the most vowel identification (McDermott, 1998). The formant frequencies were utilised for the first design of CI processors, and they proved their utility in increasing the average scores in speech perception tests.

Several studies investigated the effect of frequency resolution in CI simulation on speech perception with NH listeners. Speech perception in quiet conditions with greater than 90% correct for sentences could be achieved using four bands in a frequency range between 0 and 4 kHz (Shannon et al., 1995). However, speech perception in noise requires a greater number of frequency bands than in quiet (approximately six to eight bands) in order to discriminate the difference between speech and noise (Dorman et al., 1998). The performance with processed speech with twenty-four bands in a frequency range of 0 to 6 kHz was considerably poorer than for unprocessed speech (or natural speech) in both stationary and nonstationary noises (Qin and Oxenham, 2003). Unprocessed speech in a frequency range between 0 and 10 kHz maintained high levels of speech perception (approximately 80 % correct), whereas processed speech with four bands was close to the floor effect at 0 dB SNR (Stickney et al., 2004). Fu and Shannon (1998) found that speech perception by some Nucleus CI users with four bands in a

frequency range of 0 to 6 kHz was similar to that for NH listeners with the same processing strategy in quiet and noise conditions.

Some studies explored the effect of temporal resolution (or stimulation rate) on speech perception. Temporal resolution is associated with electrical pulses transmitted to each electrode. A high stimulation rate can better represent the temporal envelope of speech signals, and it can reasonably expect to provide high speech perception. However, the consistent advantages of higher stimulation rates from several studies have not been established yet (Loizou, 2006), and it is still unknown how to identify the optimal stimulation rate for CI individuals. Some studies (Loizou et al., 2000) found higher rates (e.g. 2100 pulses per second (pps)) in Med-El devices provided more benefits to speech perception than lower rates (< 800 pps). In contrast, other studies (Cochlear, 2007) found no significant effects of higher rates on speech perception. Low to moderate rates (e.g. 900 and 1200 pps) provided better speech perception than the higher rates (e.g. 1800, 2400 and 3500 pps). This is most likely to be because of differences in speech coding strategies, speech materials, specific parameters of electrodes, and neuron survival.

NH individuals naturally raise their voices in noisy environments, which effectively increase the SNR. A study by Firszt et al. (2004) found that the sentence recognition performance at 60 and 50 dB SPL showed a difference of approximately 15% correct for seventy-eight CI users with different CI devices. The performance at 60 dB SPL in quiet and in noise showed a difference of approximately 30% correct. The performance was poorer in noise (e.g. 60 dB SPL and 8 dB SNR) when compared with listening at a softer conversation level (e.g. 50 dB SPL) in quiet conditions.

1.3 Speech coding strategies for cochlear implants

Speech coding strategies have been used to describe techniques that process speech sounds in CI systems. The speech coding strategy is the brain of the CI system (Zeng, 2004). It plays a very important role and affects the overall performance of cochlear devices to benefit CI users in terms of more effective communication (Loizou, 1998).

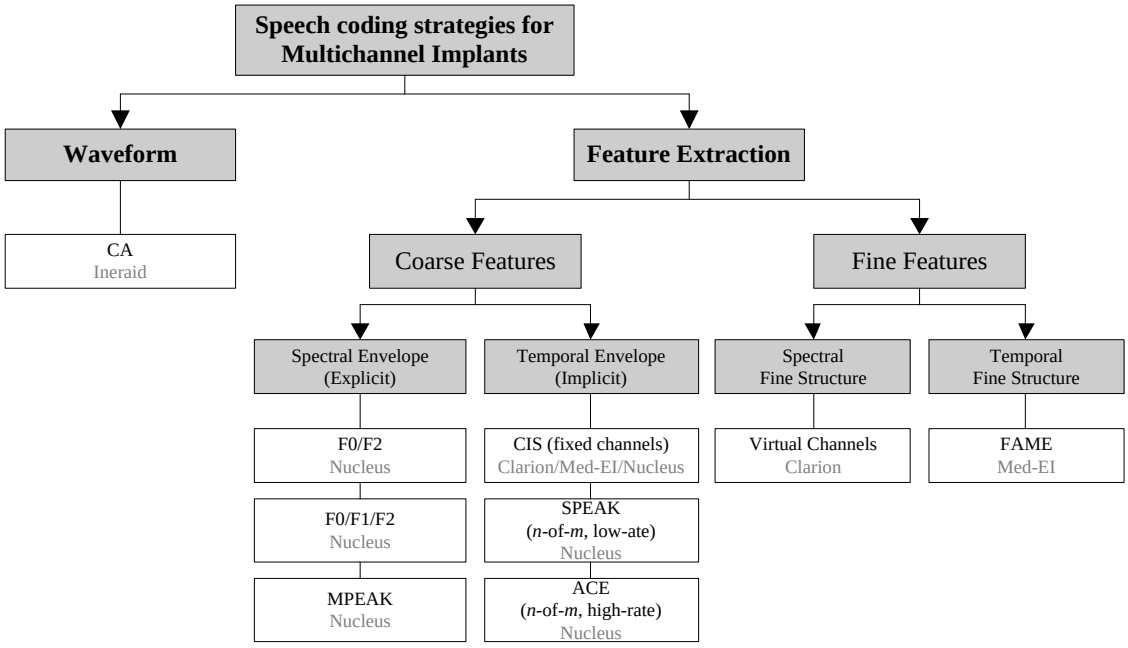


Figure 1.2 Classification of speech coding strategies for multichannel implants. Adapted from Fan-Gang et al. (2008).

Speech coding strategies have been developed over the past two decades. The strategies decompose speech sounds into multiple frequency channels to mimic the healthy cochlea (Loizou, 1999). The perceptually important information contained in speech sound needs to be preserved to facilitate hearing ability and to improve speech intelligibility. This important information is transmitted to the brain by electrical stimulation relating to the amplitudes and frequencies of speech signals. The amplitude of the stimulus current controls the loudness of speech sounds. The different pitches relate to the positions in the cochlea that are being stimulated. Low pitch is perceived when electrodes near the apical part of the cochlea are stimulated, while high pitch is perceived when electrodes near the basal part are stimulated (Loizou, 1998).

There are currently three major manufacturers for CI devices approved by the Food and Drug Administration in the United States: the Nucleus device (Cochlear Corporation, Australia), the Clarion device (Advanced Bionic Corporation, USA), and the Med-El device (Med-El Corporation, Austria). The CI manufacturers offer several speech coding strategies to CI users. All speech coding strategies for multichannel

implants (Figure 1.2) can be classified into two main categories: waveform and feature-extraction strategies (Loizou, 1998; Fan-Gang et al., 2008). The waveform strategy represents waveforms of each frequency band (Loizou, 1999), while the feature-extraction strategy (Loizou, 1998; Fan-Gang et al., 2008) represents the dominant features of speech signals in each frequency band.

1.3.1 Waveform strategy

An example of a waveform strategy is the compressed-analogue (CA) approach (Loizou, 1999), which was an early strategy for CIs. The concept of the CA approach is that the speech signal is first compressed using an automatic gain control (AGC), and then it is filtered into four frequency bands with centre frequencies at 0.5, 1, 2, and 3.4 kHz. The filtered waveforms are amplified using adjustable gain controls, and then they are delivered directly to four electrodes. The CA doesn't extract any features of speech signals in each frequency band, but it delivers the full waveform to different electrodes. A disadvantage of the CA approach is the interaction between channels due to simultaneous stimulation, which may lead to distortion of the speech spectrum and degradation of speech intelligibility.

1.3.2 Feature-extraction strategy

The feature-extraction strategy used in modern CI devices can be separated into coarse features and fine features. Fan-Gang et al. (2008) state that for the general model the slowly varying envelope contributes to speech intelligibility, while the rapidly varying fine structure contributes to mainly auditory object formation. Nevertheless, the majority of current CI devices use coarse features and discard fine structures.

1.3.2.1 Coarse features

Spectral-envelope information is used for the early speech coding strategies (e.g., F0/F2, F0/F1/F2 and multi-peaks (MPEAK)). They are designed by using formant information to convey some information about the speech signals to the electrodes. However, this strategy has some disadvantages (McDermott, 1998); for instance, the

problems of estimation formant frequencies can be challenging, especially in noisy environments, there may be a loss of temporal resolution for the rapidly varying spectral features (e.g. low stimulation rate used), and there can be inappropriate processing of speech-like sounds.

In the early 1990s temporal-envelope information (Fan-Gang et al., 2008) was developed for speech coding strategies such as the continuous interleaved sampling (CIS), the spectral peak (SPEAK), and the advanced combinational encoder (ACE) strategies. The temporal-envelope information could provide better performance for speech intelligibility than the spectral-envelope information (Loizou, 1999; Fan-Gang et al., 2008). Although the temporal-envelope information uses implicit identification of speech features, the selection of frequency bands with the largest amplitudes usually represents the first three formants (McDermott, 1998).

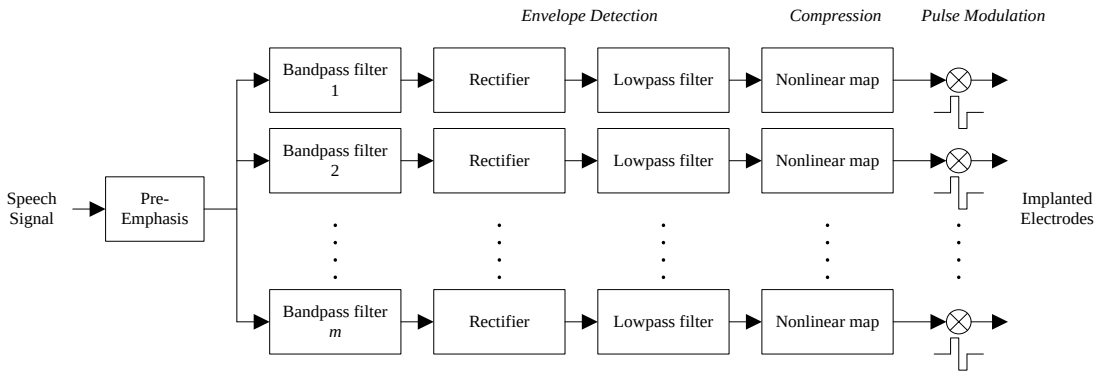


Figure 1.3 Block diagram of the CIS strategy. Adapted from Loizou (2006).

The Nucleus device as shown in Figure 1.2 supports the CIS, SPEAK, and ACE strategies. The CIS strategy as shown in Figure 1.3 (Loizou, 1999; Fan-Gang et al., 2008) is based on a fixed-channel strategy, and is implemented by all major manufacturers. The speech signal is pre-emphasised and then divided into a number of frequency bands. The envelopes of the outputs in each frequency band are extracted by rectification and lowpass filters. The extracted envelopes are compressed using a nonlinear map (e.g. a logarithmic map) to fit within the electrical dynamic range. Finally, the compressed amplitudes are used to modulate the stimulating pulse and transmitted to the implanted electrodes. The CIS strategy can reduce channel

interactions by stimulating channels asynchronously. In other words, only one electrode is stimulated at a time.

The SPEAK and ACE strategies are similar to the CIS strategy, but both are based on an *n*-of-*m* strategy. The *n*-of-*m* strategy was first introduced by Wilson and colleagues in 1988 (Fan-Gang et al., 2008). This strategy selects *n* envelope channels with the largest amplitudes from *m* channels (related to electrodes) for stimulation in each cycle (where $n<m$). Generally, the number of channels *m* represents the spectral resolution, whereas the channel stimulation rate represents the temporal resolution (Nogueira et al., 2005). The difference between the SPEAK and ACE strategies is the channel stimulation rate. The ACE strategy's stimulation rate is generally higher than SPEAK's to preserve more temporal details. The SPEAK strategy's channel stimulation rate is fixed at 250 pps, while the ACE strategy's channel stimulation rates vary between 250 and 2400 pps (Loizou, 2006).

The basic idea of the *n*-of-*m* strategy is to increase the temporal resolution, and to reduce the redundant information. The dominant channels can be updated more frequently by removing the less significant channels. This concentrates on the most important information conveyed to the auditory system (Nogueira et al., 2005; Buechner et al., 2009) and may reduce the overall SNR level (Wilson and Dorman, 2008b). This would also presumably reduce channel interaction further while still allowing for high resolution. Furthermore, the power consumption for CI stimulation can be decreased and this may lead to an increased battery life for the CI devices (Hu and Loizou, 2008). The ACE and SPEAK strategies have demonstrated either a significant improvement in speech recognition (Dorman et al., 2002; Skinner et al., 2002; Buechner et al., 2009), or at least they have been the CI users' preference over conventional CIS strategies (Kiefer et al., 2001; Skinner et al., 2002).

The ACE strategy is considered to be the default strategy for the Nucleus-24 processor (Cochlear, 2002) and is used by approximately 60% of CI users worldwide (Qazi et al., 2012). The overall CI stimulation rate is usually limited to 14,400 pps. The clinical channel stimulation rates range from 900 to 2,400 pps, and the number of channels varies between 8 and 12 (Hu and Loizou, 2008; Gopalakrishna et al., 2010b; Kokkinakis et al., 2011). The CI stimulation rate depends on the number of selected channels and the channel stimulation rate (i.e. CI stimulation rate $\geq$ the number of

selected channels × the channel stimulation rate). The fewer channels selected, the higher the maxima channel stimulation rate and vice versa. For example, no more than 8 maxima channels can be selected for a channel stimulation rate of 1,800 pps.

1.3.2.2 Fine features

Recently the development of CI systems has focused on fine structure processing, which is classified into spectral and temporal fine structures. The spectral fine structure needs to use more independent electrodes. It is difficult to increase the number of electrodes, and thus several techniques have been investigated to increase the number of functional channels by using virtual channels. The strategy of virtual channels was introduced by Wilson et al. in 1992 (Loizou, 2006; Wilson and Dorman, 2008b).

The principle of virtual channels is that the current of adjacent electrodes resulting in an electric field can produce the number of discriminable sites, approximately 2 to 9 sites between two adjacent electrodes depending on the channel separation. The difference between discriminable sites can be generated by using different ratios of the currents between two adjacent electrodes (e.g. 75/25, 50/50, and 25/75). The number of channels of information is increased by these intermediate sites beyond the number of physical electrodes. Some evidence (de Melo et al., 2012) has shown that speech perception performance for CI users can be improved in noisy environments with the use of the virtual channels.

Representation of temporal fine structure (Fan-Gang et al., 2008) is proposed in many ways for new strategies, such as increasing the electric stimulation carrier rate, extracting frequency modulation from the temporal fine structure to frequency modulate the carrier rate, and using multiple carriers for fine frequency structure. However, these strategies have been not demonstrated to provide benefits to CI users (Fan-Gang et al., 2008; Moon and Hong, 2014). Representation of temporal fine structure using new strategies may improve CI hearing in the future.

1.4 Noise reduction in cochlear implants

In quiet environments, most CI users can achieve high performance with speech recognition regardless of CI devices they use, because almost all CI devices perform

well in quiet listening conditions (Kokkinakis et al., 2012). Their speech recognition performance has improved steadily in quiet environments over a number of years (Fan-Gang et al., 2008). However, in noisy environments, many CI users complain of severe degradation in speech understanding. Recently CI research effort has increasingly focused on state-of-the art noise reduction strategies to achieve higher speech intelligibility in noisy environments.

Terms such as noise reduction, noise suppression, and speech enhancement have been used to describe methods that improve speech intelligibility and speech quality in noisy environments (Kokkinakis et al., 2012). Noise reduction algorithms were originally developed for NH listeners over many decades. Most of them are based on a single microphone and can be classified into four main categories (Loizou, 2007), namely spectral subtraction, Wiener filtering, statistical-model based methods, and subspace algorithms.

Spectral subtraction relates to the subtraction of noise spectrum estimates from the noisy speech spectrum. Wiener filtering works by providing a linear estimate of the clean speech spectrum, and is optimal in the mean-square sense. In addition, it can be performed in both the time and frequency domains, and it can be implemented either iteratively or non-iteratively. Statistical-model-based algorithms use an estimator in various statistical models and optimization criteria. Finally, subspace algorithms are based on the linear algebra theory that noisy speech can be decomposed into vector subspaces comprising a clean signal and a noise signal.

The development of noise reduction algorithms has made little progress in improving speech intelligibility, but much progress in improving speech quality in noisy environments (Loizou and Gibak, 2011). The comparative speech intelligibility of some algorithms encompassing the four categories of single-microphone noise reduction algorithms were investigated for NH listeners (Hu and Loizou, 2007; Li et al., 2011). Almost all noise reduction algorithms yielded little benefit or did not improve speech intelligibility in American English (Hu and Loizou, 2007) or other languages (i.e. Chinese and Japanese) (Li et al., 2011). Wiener filtering produced a significant improvement in speech intelligibility when compared with others, but in only car and white noise conditions (Hu and Loizou, 2007; Li et al., 2011).

However, some single-microphone noise reduction algorithms are successful in improving speech intelligibility for CI users. Most noise reduction algorithms deal with situations at 0 to 15 dB SNR level, which CI users can benefit (Fu et al., 1998). Many noise reduction methods for CI processors are based on a single microphone signal, while others exploit more than one microphone signal. Some good reviews of the literature can be found in Loizou (2006), Li (2009), and Kokkinakis et al. (2012).

1.4.1 Multi-microphone noise reduction strategies

Most of the multi-microphone noise reduction strategies are based on adaptive beamforming (ABF) algorithms and are implemented as pre-processing algorithms for CI processors. The ABF refers to signal processing that uses the spatial differences between at least two microphones to adaptively attenuate or preserve signals from particular directions (Vanhoesel and Clark, 1995).

Experimental results for an ABF algorithm with two microphones, one behind each ear, with four Nucleus CI users (Vanhoesel and Clark, 1995) indicated that there was a large improvement in speech intelligibility in conditions where reverberation is moderate but only one source is predominantly interfering with speech. An ABF with a two-microphone array in a single behind-the-ear hearing aid (Wouters and Vanden Berghe, 2001) provided significant improvements in word recognition in both speech-weighted noise and babble noise, corresponding to an average SNR improvement of approximately 10 dB among this group of four CI users.

The performance of the two-microphone adaptive beamformer BEAM (Spriet et al., 2007) with five Nucleus CI users was evaluated at different SNR levels and with two types of noise, speech-weighted noise and babble noise. This approach yielded an average SNR improvement of 5–16 dB for sentence recognition. The nonlinear spectral subtraction proposed by Kallel et al. (2012) and the multi-band spectral subtraction proposed by Kamath and Loizou (2002) were implemented as a pre-processing algorithm for CI processors (Kallel et al., 2012). Both algorithms were evaluated by three bilateral CI users and fifty NH listeners at different SNR levels of babble noise. These approaches provided an average improvement in the percentage correct word scores of 4–9% for bilateral CI users and 7–13 % for NH listeners.

Although such algorithms for multi-microphone noise reduction can bring benefits in speech intelligibility, these benefits are limited to situations where the speech and noise signals are spatially separated and may degrade in reverberant environments (Wouters and Vanden Berghe, 2001). Implants with two or more microphones are ergonomically difficult, and CI users may not like to wear headphones or a neck loop (Loizou, 2006). Most CI users would find this a cosmetically unappealing prosthesis. In addition, they are computationally complex, and it is difficult to optimise the particular algorithms to individual CI users (Li, 2009). Single-microphone noise reduction strategies, which can work under nondirectional conditions, are therefore more user-friendly and desirable.

1.4.2 Single-microphone noise reduction strategies

Single-microphone noise reduction strategies used in the latest CI processors (i.e. temporal-envelope information) can be divided into two main categories (Loizou, 2006; Kokkinakis et al., 2012). The first of these is the pre-processing noise reduction strategy, where the noisy speech is processed with a speech enhancement algorithm and then the enhanced speech is fed into the CI speech coding strategies. This approach is similar to the speech enhancement that is used in most modern communication devices (e.g. mobile phones). Another category is envelope-based noise reduction strategies. This approach is combined to form one part of the speech coding strategy to attenuate directly on noisy envelopes.

1.4.2.1 Pre-processing noise reduction strategies

A few single-microphone noise reduction algorithms have been proposed as a pre-processing approach for CI processors. Yang and Fu (2005) evaluated the spectral subtraction algorithms proposed by Marzinzik and Kollmeier (2002) with seven CI users wearing different CI devices. The results showed an average improvement in sentence recognition scores of 21% at various SNR levels (i.e. 0, 3, 6, and 9 dB) of speech-shaped noise, and this did not vary significantly with SNR levels. The performance of the nonlinear spectral subtraction proposed by Lockwood and Boudy (1992) was used in a study by Verschuur et al. (2006). Results indicated that there were

large improvements in sentence recognition scores of 5−10% at different SNR levels (i.e. 5 and 10 dB) of speech-shaped noise in a group of seventeen Nucleus CI users.

The study by Loizou et al. (2005) demonstrated that the subspace algorithm proposed by Hu and Loizou (2002) provided significant benefits to fourteen Clarion CI users in sentence recognition scores, with an average improvement of 44%, in conditions of 5 dB SNR speech-shaped noise. However, it is unclear whether such intelligibility benefits would be preserved if these algorithms were evaluated in nonstationary noise environments (e.g. babble noise).

The pre-processing approach has a few main disadvantages (Kokkinakis et al., 2012). First, speech enhancement algorithms can be implemented as pre-processing algorithms in several ways. They require the speech signal in the time domain to be transformed into any domain to reduce the noise, and then be reconstructed into the enhanced speech in time domain before transmitting them to the CI processor. Some algorithms have shown improvements in speech perception, but they are highly computationally complex (e.g. using the subspace algorithm). Therefore, they may be suitable for implementation on computers but not on wearable CI processors (Dawson et al., 2011). Second, speech enhancement algorithms sometimes provide unwanted distortion, which degrades speech perception. Third, there is no simple approach to optimise the operation of algorithms to individual CI users.

1.4.2.2 Envelope-based noise reduction strategies

The simple way to overcome the drawbacks of the pre-processing strategy is to directly apply attenuation to the envelopes after the step of envelope detection in speech coding strategies, as in Figure 1.3. Envelope-based noise reduction algorithms were integrated into the stage of the speech coding strategies as envelope-weighting or envelope-selection (Kokkinakis et al., 2012).

1.4.2.2.1 Envelope-weighting

A number of algorithms were proposed for envelope-weighting to attenuate noisy envelopes according to the estimated SNR in each channel. A sigmoidal-shaped gain function (Hu et al., 2007) was introduced as a simple algorithm to perform a weighting (values in the range of 0 to 1) in noisy envelopes of each channel. The envelope amplitudes in channels with high SNR levels were multiplied by a weight close to one,

whereas the envelope amplitudes in channels with low SNR levels were multiplied by a weight close to zero. This approach fits into the general category algorithms of noise reduction that enhance speech by spectral modification (e.g. spectral subtraction and Wiener filtering). Results showed large improvements in sentence recognition of 10−25% at different SNR levels (i.e. 5 and 10 dB) of babble noise in a group of five Clarion CI users, compared to the improvement of 7% for pre-processing algorithms reported in Yang and Fu (2005).

Principal components analysis (PCA) and independent component analysis (ICA) with soft thresholding (Li, 2008) can be used to reduce noise and signal redundancy. This approach significantly improved word recognition for ten Nucleus CI users at different SNR levels (i.e. 5, 10, and 15 dB) of babble noise and modulated noise. In Dawson et al. (2011)'s study, two gain functions were proposed and tested with thirteen Nucleus CI users. The first gain function used a combination of a *posteriori* SNR estimate (Mcaulay and Malpass, 1980) and a sigmoidal-shaped gain function (Hu et al., 2007). The second gain function used a combination of a *priori* SNR estimate (Mcaulay and Malpass, 1980) and a modified Wiener gain function (Loizou, 2007). This approach provided the greatest improvement in sentence recognition for speech-weighted noise; 1.77 dB for the first gain function and 2.14 dB for the second gain function. A sparse non-negative matrix factorisation (Hu et al., 2013) with five NH listeners provided an improvement sentence recognition in terms of speech intelligibility and speech quality at 0 and 5 dB, but not at 10 dB in babble noise.

1.4.2.2.2 Envelope-selection

A few techniques were introduced as criteria for envelope-selection in each channel to transmit useful information to electrodes. In the *n*-of-*m* strategy such as the ACE or SPEAK strategy, the channel-selection criterion with the largest amplitudes works well in quiet conditions, but it can be problematic in noisy conditions when the noise may completely dominate the speech.

Another channel-selection criterion was based on a psychoacoustic model (Nogueira et al., 2005), which was adopted in audio compression standards (MP3). This method was referred to as the psychoacoustic advanced combination encoder (PACE) strategy. The idea of this method was that amplitudes falling below a masking threshold would not be audible and so could be discarded. The PACE strategy was evaluated and

compared to the ACE strategy in sentence recognition in speech-shaped noise at 15 dB SNR in eight Nucleus CI users. This provided an average improvement over the ACE strategy of 17% for 4-of-20 channels, but no significant difference for 8-of-20 channels.

The SNR channel-selection criterion (Hu and Loizou, 2008) was proposed under the assumption that the true SNR values in each channel are known a *priori*. The idea of this criterion was that each channel was selected only when its SNR was more than or equal to 0 dB (speech-dominated channels), whereas each channel was discarded when its SNR was less than 0 dB (noise-dominated channels). Results revealed that this strategy could restore speech intelligibility to the level obtained in quiet conditions for six Clarion CI users. Sentence recognition was not dependent on different types of noise (babble noise and speech-shaped noise) and different SNR levels (0–10 dB).

The SNR channel-selection criterion can be also considered as envelope-weighting and can be implemented simply by multiplying the noisy envelopes with a binary gain function. The speech-dominated channels were assigned to a value of 1 and the noise-dominated channels were assigned to a value of 0. However, it cannot be implemented in real-world applications because of the fact that the SNR needs to be estimated from the noisy envelopes.

Overall, the ideal algorithms for noise reduction should be easy to implement and integrate into commercially available CI devices. The integration of envelope-based noise reduction algorithms into speech coding strategies has some advantages (Hu et al., 2007), including the lack of algorithmic delay related to the pre-processing approach, the low computational complexity and the ease of integration into existing speech coding strategies.

1.5 Performance evaluation

Most of the speech enhancement algorithms are usually evaluated in terms of speech intelligibility and speech quality. Speech intelligibility is related to the number of words that are identified correctly by the listeners, while speech quality is related to how natural speech sounds and the individual preferences of listeners (Ephraim and Cohen, 2004; Loizou, 2007; Li et al., 2011). In fact, improving speech intelligibility does not

correlate with enhancing speech quality (Ephraim and Cohen, 2004). Some cases of enhancing speech quality may lead to a decrease in intelligibility. This is due to the distortion of enhanced speech resulting from excessive noise reduction. In CI research, speech intelligibility is the most important criterion and is considered for evaluating performance improvements in CI processors.

Speech intelligibility measures used to assess the CI processors can be classified into subjective and objective intelligibility measures. The subjective intelligibility measures are regularly quantified in terms of the percentage of words that listeners can correctly identify. The percentage is often measured by using fixed SNR levels. There are two subjective tests. The first test uses speech stimuli transmitted directly to CI users. The second test uses vocoder simulation to simulate the speech processing of a CI processor, which is referred to as vocoded speech. The vocoded speech is then presented to NH listeners.

The listening test with CI users has numerous factors influencing CI processor performance. Two types of factors, namely CI user-related factors (e.g. duration of deafness, duration of CI use, age at implantation, electrode placement and insertion depth) and CI processor-related factors (e.g. number of channels, the stimulation rate and frequency-to-electrode allocation), may affect the speech perception of CI users. It is difficult to interpret the impact of each factor on speech intelligibility because of interaction between these factors (Fu et al., 1998; Loizou, 1998; Chen and Loizou, 2011).

Vocoder simulation with NH listeners has been widely used to evaluate the effects of different factors on speech intelligibility, avoid the confounding factors specified by individual CI users, and manipulate interesting parameters relating to the CI design. The vocoder simulations are not expected to predict the absolute levels of performance of individual CI users, but rather they can indicate trends of performance observed in CI users when a specific parameter of the speech coding strategy or a property of the acoustic signal is varied (Chen and Loizou, 2011).

Although, subjective listening tests are very important and necessary for the evaluation of intelligibility measures, these measures are very expensive and time consuming. In addition, this measure cannot be used for tuning parameters during the development of some stages of new algorithms (e.g. noise reduction algorithms and

speech coding strategies) or the comparison of different CI processors. Therefore, the objective intelligibility measures play an important role. They allow repeatable assessment at different stages of the algorithm development process as well as performance comparison among algorithms. These also help to provide guidance on how to improve the speech intelligibility of algorithms involved in the implementation of CI systems, and how to develop new algorithms involved in the implementation of CI systems in the proper direction.

Finally, for the objective intelligibility measures to be valid, they need to correlate well with subjective intelligibility measures. Some good literature reviews relating to objective intelligibility measures can be found in Rhebergen and Versfeld (2005), Jianfen et al. (2009), Christiansen et al. (2010), Ma and Loizou (2011) and Taal et al. (2011). The literature reviews of some of these objective intelligibility measures are briefly described in the next section.

1.5.1 Objective intelligibility measures

The prediction of speech intelligibility was first introduced by French and Steinberg (1947) who proposed the concept of the articulation index (AI). The AI was further developed to produce a new measure called the speech intelligibility index (SII) (ANSI, 1997). The SII was corrected in terms of hearing sensitivity loss and speech level as well as the upward and downward spread of masking (Christiansen et al., 2010). The SII is calculated from the SNR between the long-term speech spectrum and the long-term noise spectrum in each frequency band. Then, the auditory threshold is adjusted and the weighted SNR is summed across frequencies to produce the SII value. The SII value is a number between zero and unity. An SII of zero implies that no speech information is available and an SII of unity implies that all speech information is available to the listener. However, the SII works well for stationary noise, but not for nonstationary noise due to its calculation based on the long-term spectra of the speech and noise signals.

Another well-known intelligibility measure was the speech transmission index (STI) (Steeneken and Houtgast, 1980). The STI was first introduced to evaluate the quality of speech-transmission channels. However, the STI measure was also able to

successfully predict the intelligibility of reverberation, room acoustics, and additive noise (Steeneken and Houtgast, 1982; Houtgast and Steeneken, 1985). The method of STI calculation is similar to the SII method, which is based on the SNR in each frequency band. However, the SNR in each frequency band is related to the modulation depth and forms a weighted sum across frequencies to give the STI value. The STI value represents intelligibility between zero and unity. The meaning of the STI value is the same as the SII value.

Several attempts have been made to further develop the measures to predict speech intelligibility in various noisy environments and with different distortions produced by signal processing systems (e.g. hearing prostheses and noise-suppressed algorithms). Most intelligibility measures were based on the STI or the AI. The AI-based measures use the spectral-envelope information and include a short-term AI-based measure (AI-ST) (Jianfen et al., 2009), the coherence-based speech intelligibility index (CSII) (Kates and Arehart, 2005), the three-level CSII measures ($CSII_{high}$, $CSII_{mid}$, and $CSII_{low}$) (Kates and Arehart, 2005), and the I3 (Kates and Arehart, 2005). The STI-based measures use the temporal-envelope information and include the normalised covariance metric (NCM) (Goldsworthy and Greenberg, 2004; Jianfen et al., 2009; Chen, 2011).

The AI-ST is computed using short-term (30 msec) segments. In addition, the difference between the AI-ST and the SII is that the AI-ST does not use the auditory threshold and it does not account for the upward spread of masking. It was found to predict the speech intelligibility modestly in nonstationary noise (Jianfen et al., 2009). The CSII and I3 were introduced to predict intelligibility in additive noise, peak-clipping, and centre-clipping distortion in hearing aids (Kates and Arehart, 2005). Unlike the SNR computed for the SII, the SNR of the CSII and I3 are computed from clean speech and distorted (or processed) speech. The NCM differs from the STI with respect to the change in modulation depth. The STI uses the modulation transfer function (MTF), whereas the NCM uses the covariance between the clean speech and the processed speech (Jianfen et al., 2009). The NCM was found to yield high correlation for nonvocoded speech with noise-suppressed algorithms (Jianfen et al., 2009) and vocoded speech (Chen and Loizou, 2011).

Jianfen et al. (2009) evaluated the performance of the NCM with noise-suppressed speech (nonvocoded speech). The noise-suppressed speech was processed by some

algorithms encompassing the four categories of single-microphone noise reduction algorithms, similar to the study by Hu and Loizou (2007). The correlation of the NCM with intelligibility scores obtained by forty NH listeners was found to be quite high (r=0.89). Chen and Loizou (2011) studied the utility of the NCM for vocoded speech with and without noise reduction. The two noise reduction algorithms used as a pre-processing approach were the Wiener filtering proposed by Scalart and Vieira (1996) and the minimum mean-square error log-spectral amplitude (MMSE LSA) algorithm proposed by Ephraim and Malah (1985). Results indicated that the NCM can be used for processed speech and performed the best when compared with others. In addition, the NCM performed very well for vocoded speech degraded by room reverberation (Santos et al., 2012).

The perceptual evaluation of speech quality (PESQ) measure (ITU-T, 2000) is an existing objective measure, originally designed to evaluate speech quality. The PESQ was also used to predict the intelligibility of vocoded speech, performing well and producing high correlations with subjective listening tests in stationary and nonstationary noise (Chen and Loizou, 2010). However, the PESQ uses the vocoded speech as its input for predicting, rather than the temporal-envelope information, whereas the NCM calculation uses the temporal-envelope information with 20 channels, which is more similar to the CI processing strategy.

In vocoder simulation, the vocoded speech can be degraded by many levels of speech coding strategies (e.g. additive noise, reverberation, filtering and clipping), single-microphone noise reduction algorithms (Loizou, 2007), speech separation techniques like ideal time frequency segregation (ITFS), and so on. It is unclear whether conventional measures as previously described would be good for predicting their intelligibility. These conventional measures may be less suitable for techniques where noisy speech is processed by different types of time frequency-weightings (e.g. single-microphone noise reduction algorithms) (Taal et al., 2011). Hence, a speech intelligibility index should be able to predict the intelligibility of the vocoded speech reliably (Chen and Loizou, 2011; Taal et al., 2011).

The short-time objective intelligibility (STOI) measure (Taal et al., 2011) was proposed as a function of a time frequency-dependent intermediate measure. The STOI is similar to the NCM in that both measures are based on a correlation coefficient

between the temporal envelope of the clean and degraded speech in each frequency band. Unlike the NCM, however, which defines a correlation coefficient for the entire signal at once, the STOI determines a correlation coefficient for the short-time segments. Both represent intelligibility between zero and unity.

The STOI provided better performance compared to the reference objective measures. Five reference objective measures were the Dau Auditory model (DAU), a coherence-based speech intelligibility index (CSII), frequency-weighted segmental SNR (fwsSNR), a normalised subband envelope correlation (NSEC), and NCM. Only NCM showed a similar performance in the single-microphone noise reduction listening test (i.e. a minimum mean square error-short time spectral amplitude (MMSE-STSA) (Ephraim and Malah, 1984) and an improved version of MMSE-STSA (Erkelens et al., 2007)). In addition, the STOI was used to predict speech intelligibility of a noise reduction algorithm (i.e. the sparse coding shrinkage) (Sang, 2012). The obtained results from NH listeners were consistent with the trend of prediction of the STOI.

Therefore the NCM and STOI are chosen as preliminary measures, to guide the development of noise reduction algorithms in the wavelet packet-based speech coding strategy. The general principle of objective intelligibility measures for CI processors (Chen and Loizou, 2011) is shown in Figure 1.4. Objective measures calculate the relationship between vocoded clean speech and vocoded noisy speech (or vocoded noisy speech with noise reduction).

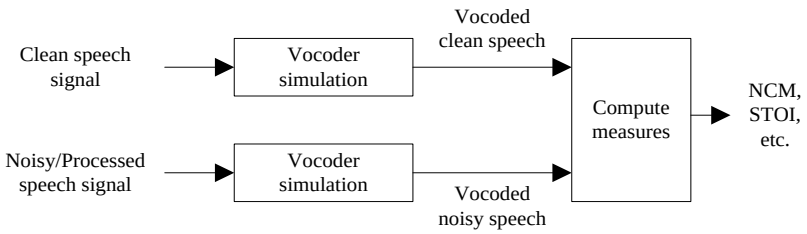


Figure 1.4 Computation of objective intelligibility measures of vocoded speech. Adapted from Chen and Loizou (2011).

1.5.1.1 The normalised covariance metric (NCM)

The NCM (Jianfen et al., 2009; Chen, 2011) is derived from the speech transmission index (STI) (Steeneken and Houtgast, 1980). The NCM uses the covariance of the envelope between the vocoded clean and vocoded noisy speech (with/without noise reduction). The NCM is calculated as follows. The vocoded signals are first decomposed into 20 bands across the signal bandwidth (125–8000 Hz in this study) using Butterworth filters. The envelope of each frequency band is computed using the Hilbert transform. The SNR is computed with a normalised correlation coefficient of envelopes between the vocoded clean and vocoded noisy speech (with/without noise reduction) in each frequency band. The values are limited to the range of [-15, 15] dB and mapped in the range of [0, 1]. These values are averaged across all frequency bands to produce the NCM value. More detailed information for how to compute the NCM is given in Appendix B.1.

1.5.1.2 The short-time objective intelligibility measure (STOI)

The STOI (Taal et al., 2011) is based on a correlation coefficient between the temporal envelopes of vocoded clean and vocoded noisy speech in the short-time region. First, the vocoded clean and vocoded noisy speech (with or without noise reduction) are processed in each frame with a length of 25.6 ms performed by Hann-windows with a 50% overlap. Next, the windowed signals are decomposed into 15 one-third octave bands. Then, the short-time temporal envelopes of the clean and noisy speech are normalised to compensate for global level differences and clipped to make sure that the sensitivity of the model is close to one time frequency-unit. Next, the short-time temporal envelopes of both are compared by means of a correlation coefficient. The short-time intermediate intelligibility measure across frequency bands is averaged to a rating value. More details for computing STOI are explained in Appendix B.2.

1.6 Conclusion

Enhancing speech intelligibility in noisy speech has been attempted over the last decade, but little progress has been made in designing algorithms due to their drawbacks and limitations as described above. The development of effective algorithms (i.e. speech

coding strategies and noise reduction algorithms) with low complexity is considered to be the main drawback. The objective of this study is to investigate whether a wavelet packet-based speech coding strategy with envelope-based noise reduction algorithms can improve speech intelligibility in noisy speech for CI users. The intelligibility performance of these algorithms is evaluated using objective measures and subjective tests with NH listeners for different noise types and SNR levels.

1.7 Outline of Thesis

With the motivation of improving speech intelligibility in noisy speech for CI processors, wavelet analysis is exploited to develop a novel speech coding strategy. The design, analysis, and evaluation of the speech coding strategy are discussed and organised as follows:

Chapter 2 gives a comprehensive review of wavelet analysis. Discrete wavelet transforms (DWTs) are described in terms of filter banks. DWTs can be classified into real-valued and complex wavelets. The real-value wavelets including standard DWT, stationary wavelet transform (SWT) and wavelet packet transform (WPT), are explained in terms of their structure of decomposition and reconstruction. The benefits of wavelets and their applications are given.

Chapter 3 gives some details associated with CI design, such as the concept of the basilar membrane model and auditory filter banks. These lead to the design of the structure of a Bark scale wavelet packet used in the speech coding strategy. This strategy is compared to the structure of a standard ACE strategy.

Chapter 4 presents the principle of noise reduction techniques (i.e. analysis, suppression and synthesis). The noise reduction algorithms, namely time-frequency spectral subtraction (TFSS) and time-adaptive wavelet thresholding (TAWT), are described and compared to ideal binary masking (IdBM) as a baseline for denoising performance. These algorithms are integrated into the speech coding strategy as an envelope-based noise reduction strategy to directly attenuate the envelope of noisy speech. An evaluation with objective intelligibility measures (i.e. the normalised

covariance metric (NCM) and short-time objective intelligibility (STOI)) is used to predict the trend of performance, before a listening test with normal-hearing listeners.

Chapter 5 presents the study with normal-hearing listeners, to show the contribution of improving speech intelligibility in noise reduction algorithms in a wavelet packet-based speech coding strategy. The performance evaluation with NH listeners can be divided into two parts: the effects of parametric variation in wavelet packet filter banks on speech intelligibility (i.e. filter spacing, types of mother wavelet and frame lengths) to find the optimal parameters and the comparison of noise reduction algorithms in cases of different types of noise and different SNR levels. The sentence scores obtained from normal-hearing listeners and the predicted values of NCM and STOI are assessed for validity.

In Chapter 6, a general discussion is presented of the limitations of wavelet packets, the objective intelligibility measures, vocoder simulation, and performance evaluation for developing speech coding strategies with noise reduction algorithms. Finally, some limitations of this study are discussed, and directions for future research are given.

Chapter 7 presents the conclusion and contributions of the research.

Chapter 2: Wavelet Analysis

2.1 Introduction

Transformations are useful tools to enable the exploration of signal characteristics. In the analysis stage, a signal is transformed into another domain by techniques such as discrete Fourier transform (DFT), discrete cosine transform (DCT), wavelet transform (WT), and so on. The oldest and best-known method is the Fourier transform (FT) that transforms any signal from the time domain to the frequency domain. However, the FT is not always the best tool to analyse real signals. It is not appropriate for analysing nonstationary signals and it is not able to reveal inherent information in nonstationary signals (Peng and Chu, 2004).

This problem has been partly resolved by using the short-time Fourier transform (STFT) based on time-frequency analysis. For many years the STFT has been the most popular method for analysing nonstationary signals like speech. However, the shortcoming of STFT is that it uses the same window for analysing the different frequency bands, which provides constant resolution at all frequencies. This property does not reflect the structure of speech. Wavelets are characterized by having a time resolution which increases with high frequency. Consequently, wavelets provide a natural candidate with which to compute features for speech processing in a CI system.

This chapter gives a brief overview of wavelet evolution, the main wavelet theory and its application to speech processing. The FT and the STFT are explained and compared in order to better understand the basic concept of wavelet analysis. Important concepts related to wavelet analysis, including dilation, translation, multiresolution analysis and filter banks, are considered.

2.2 Fourier transform (FT)

The FT is a well-known mathematical tool and a helpful method of representing signals from the time domain to the frequency domain. The FT of any signal $x(t)$ is given by

$$X_{FT}(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \tag{2.1}$$

The FT contains basis functions that are sinusoidal waves. A signal is decomposed into sine waves of different frequencies. The FT has a good ability to extract information efficiently. However, the limitation of FT is that it cannot offer both time and frequency localisation of a signal at the same time; it only provides easily accessible information about the global frequency content. This is not a serious shortcoming for stationary signals that do not change over time. However, it does hinder its direct application for nonstationary signals that change over time, such as speech signals.

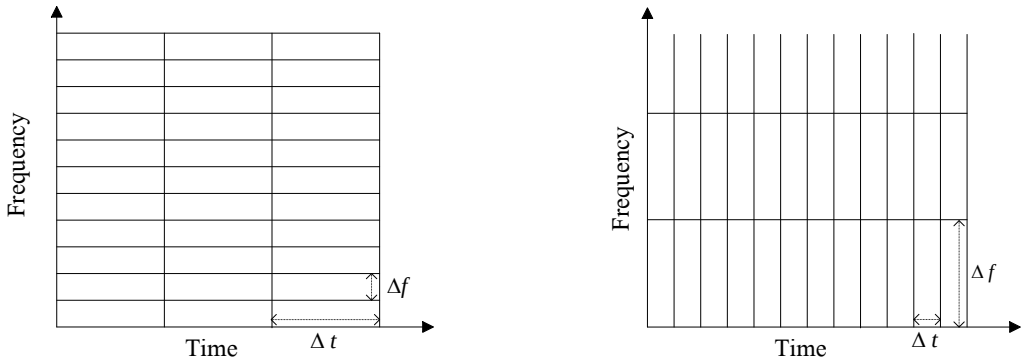
The short-time Fourier transform (STFT) was introduced to overcome this problem by using a fixed-length window $w(t)$ shifted to be centred at time τ. This window is translated along the time axis, analysing the frequency content of the signal in the windowed time interval. Mathematically the STFT can be defined as:

$$X_{STFT}(\tau, f) = \int_{-\infty}^{\infty} x(t) w(t-\tau) e^{-j2\pi ft} dt \tag{2.2}$$

Even though the STFT has demonstrated utility in numerous applications, it has disadvantages. Due to its use of a single window length for analysing the whole signal, the time-frequency resolution of signal analysis is the same at all locations in a time-frequency plane. The accuracy of the information obtained from the STFT is limited by the size and shape of the window. Many naturally-occurring signals contain long-lasting frequency components, but high frequency components may require shorter time windows. The STFT offers fixed time and frequency resolution so is not well suited to analysis of such signals.

Figure 2.1 shows the time-frequency plane of the STFT with Heisenberg boxes. Heisenberg's uncertainty principle suggests that in modelling a signal one cannot be arbitrarily precise in both time and frequency simultaneously. This principle states that the product of time resolution Δt and frequency resolution Δf is constant. That means that the boxes in the time-frequency plane have the same area. The STFT uses a fixed

window length, and thus Δt and Δf are constant all the whole plane. Figure 2.1 (a) illustrates that a longer window provides better frequency resolution but poorer time resolution. Figure 2.1 (b) shows that a shorter window provides better time resolution but poorer frequency resolution. It is impossible to obtain good time resolution and good frequency resolution using STFT. The WT gives a better trade-off between time and frequency resolutions than the fixed window length used in the STFT. The details of WT are described in next section.



(a) STFT with longer window (b) STFT with shorter window

Figure 2.1 The time-frequency plane of STFT.
Adapted from Vetterli and Herley (1992).

2.3 Wavelet transform (WT)

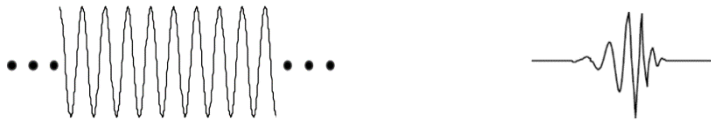
2.3.1 Wavelets

A wavelet is a waveform with a set oscillatory structure that is nonzero for a limited duration, with additional mathematical properties (Fugal, 2009; Mallat, 2009). Within the constraints of the required mathematical properties, wavelets have different shapes and sizes. The difference between sinusoidal waves and wavelets is shown in Figure 2.2.

A wavelet transform (WT) is performed using a wavelet basis function. A signal is decomposed by using translated and dilated versions of the wavelet basis function to produce a correlation of signals and localise energy concentration in the time-frequency

domain. This function is informally called a “mother wavelet”, which can be thought of as a bandpass filter (Vetterli and Herley, 1992). The mother wavelet can be specified by a set of numbers referred to as the coefficients of the wavelet filter. There are many different types of mother wavelets such as the Mexican hat wavelet, Daubechies, Symlet and Coiflet.

The mother wavelets can be selected according to the characteristics of signals and the requirement of each application (Fugal, 2009). For example, the Mexican hat wavelet is employed in vision analysis, because its characteristics are similar to the computation performed by the retina. The Morlet wavelet is used in atmospheric indices (e.g. cyclical change in air pressure and in storm tracks). The Haar wavelet is well suited to edge detection. The Daubechies and Symlet wavelet is often used in speech and image processing.



(a) Sinusoidal function (b) Wavelet function

Figure 2.2 Characteristics of sinusoidal and wavelet functions.

Wavelet theory was introduced in 1984 by Morlet, who formalised the continuous wavelet transform (CWT) (Peng and Chu, 2004). In the next year Mayer constructed orthonormal wavelets with very good time and frequency localisation properties. Mayer and Mallat developed the concept of multiresolution analysis (MRA) which is useful for constructing other orthonormal wavelets and for computing the wavelet decomposition of signals from their finest approximation resolution using a recursive filtering algorithm. In 1988 Daubechies constructed a set of orthonormal wavelet basis functions with compact support that have become the foundation of many wavelet applications. Wavelet development from continuous to discrete signal analysis, developed by Daubechies and Mallat, is widely accepted and credited.

WTs are broadly classified into continuous wavelet transforms (CWTs) and discrete wavelet transforms (DWTs). DWTs, as shown in Figure 2.3, can be classified into two main types: real-valued DWTs and complex-valued DWTs. The real-valued DWTs use real-valued filter coefficients and give real-valued wavelet coefficients. In contrast, the complex-valued DWTs also use real-valued filter coefficients but give complex-valued wavelet coefficients. The class of real-valued DWTs can be divided into three basic forms: the standard DWT, stationary wavelet transforms (SWTs), and wavelet packet transforms (WPTs). The ℂWTs can be divided into two classes: dual-tree DWT-based ℂWTs and projection-based ℂWTs.

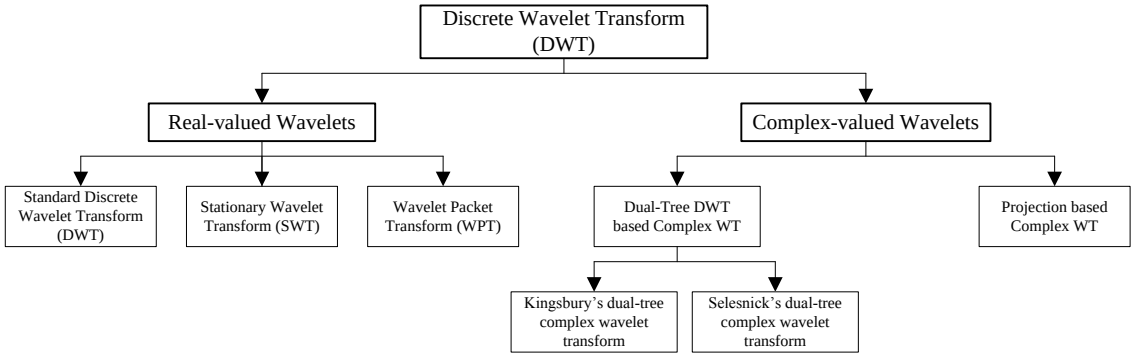


Figure 2.3 The classification of discrete wavelet transforms.

2.3.2 Continuous wavelet transforms (CWTs)

Let $\psi_{a,b}(t)$ be a wavelet basis function (Daubechies, 1992), which is generated in dilated and translated versions:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{2.3}$$

where the real numbers $a(a>0)$ and b denote the dilation and the translation respectively. The factor $1/\sqrt{a}$ is introduced to guarantee energy preservation. The CWT of any signal $x(t)$ is defined as:

$$C(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi \left(\frac{t-b}{a} \right) dt \tag{2.4}$$

where $C(a,b)$ are known as the wavelet coefficients.

The wavelet function $\psi_{a,b}(t)$ is stretched and contracted by changing the dilation parameter a, which covers different frequency ranges. With larger a the wavelet function becomes stretched and corresponds to low frequency components. The wavelet function with smaller a becomes contracted and represents high frequency components. Variation in the dilation parameter a also changes the window length. The wavelet is shifted over the signal by changing the translation parameter b.

WT provides a flexible time-frequency window. The frequency resolution of the WT is good at low frequencies while the time resolution becomes good at high frequencies. This approach is reasonable in practical applications when a signal has low frequency components of long duration and high frequency components of short duration. Figure 2.4 presents the time-frequency resolution of WT with Heisenberg boxes. It is clear that WT uses longer time windows at lower frequencies and shorter time windows at higher frequencies.

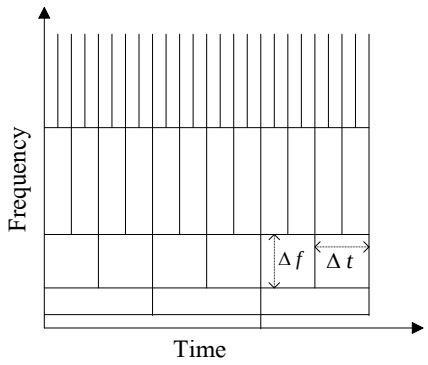


Figure 2.4 Time-frequency plane of WT. Adapted from Vetterli and Herley (1992).

2.3.3 Discrete wavelet transforms (DWTs)

The CWT is infinitely redundant due to the continuous values of dilation a and translation b. The CWT of these parameters also means the transforms are not suitable

for implementation in digital form. The transformation can be discretised by selecting a suitable set values of a and b at which to evaluate the CWT. The general sampling strategy adopted is defined by $a = a_0^j$ and $b = ka_0^j b_0$ where $a_0 > 1$ and $b_0 > 0$ are fixed, and $j, k \in \mathbb{Z}$ (Daubechies, 1992). The different values of j correspond to the different widths of the wavelets. A discrete set of wavelet basis functions is generated, so that Equation (2.3) becomes:

$$\psi_{j,k}(t) = a_0^{-j/2}\psi(a_0^{-j}t - kb_0) \tag{2.5}$$

It can be shown that for critical sampling $a_0 = 2$ and $b_0 = 1$, so that $a = 2^j, b = k2^j$ to produce the minimal basis. In order to preserve all information about the decomposed function, the sampling cannot be coarser than this critical sampling.

The dilation $a=2^j$ is by a power of 2, sometimes called dyadic. Thus the dyadic parameter of the wavelet basis function is given as:

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k) \tag{2.6}$$

This is actually an octave band filter and it can be interpreted as a form of constant-Q filtering, where Q represents the quality factor of the filter and is defined as the centre frequency f_c divided by its bandwidth.

2.3.3.1 Implementation of DWT

The implementation of DWT can be viewed as either multiresolution analysis or a filter bank as follows.

2.3.3.1.1 Multiresolution analysis

The wavelets can be constructed from the concept of multiresolution analysis (MRA) which was introduced by Mallat and Meyer (Daubechies, 1992; Vetterli and Herley, 1992; Meyer, 1993; Mallat, 2009). A multiresolution approximation is a sequence of closed subspaces $V_j, j \in \mathbb{Z}$ of $L^2(\mathbb{R})$ having the following properties, which form a hierarchy:

$$\cdots V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \cdots$$

$$V_j \subset V_{j-1} \tag{2.7}$$

This is a causality property that verifies that a signal approximation at a given resolution contains all the necessary information to compute a signal approximation at coarser resolutions.

The nested spaces have an intersection $\cap_j V_j = \{0\}$ which implies that the details of a signal approximation will be lost when the resolution reduces to 0. A union $\cup_j V_j = L^2(\mathbb{R})$ that is dense in $L^2(\mathbb{R})$ imposes that the signal approximation converges to the original signal. The hierarchy (2.7) is constructed such that *V*-spaces are self-similar:

$$f(2t) \in V_j \quad \Leftrightarrow \quad f(t) \in V_{j+1} \tag{2.8}$$

That means the dilation in space V_j by 2 enlarges the details by 2 (Vetterli and Herley, 1992). This guarantees that it determines an approximation at a coarser resolution.

There exists a scaling function $\phi(t)$ that derives an approximation in space V_j of signals in space V_{j-1}. The set of functions $\phi_{j,k}(t) = 2^{-j/2}\phi(2^{-j}t - k)$ is an orthonormal basis for the space V_j. In particular, if $\phi(t) \in V_1$ and $\phi(2t) \in V_0$, since $V_1 \subset V_0$ the scaling function $\phi(t)$ can be represented as:

$$\phi(t) = \sqrt{2} \sum_{k=-\infty}^{\infty} h(k)\phi(2t - k) \tag{2.9}$$

The wavelet function $\psi(t)$ is an orthonormal basis of the different space W_j. Let W_j be the orthogonal complement of V_j in V_{j-1}:

$$V_{j-1} = V_j \oplus W_j \tag{2.10}$$

V_{j-1} is equivalent to V_j plus some added detail according to W_j. In other words, a space V_{j-1} of a multiresolution approximation is decomposed into a coarser approximation

space V_j plus a detail space W_j. For $j < J$, the iteration of Equation (2.10) can be written as:

$$V_j = \ldots V_{j+3} \oplus W_{j+3} \oplus W_{j+2} \oplus W_{j+1} \oplus W_j$$

$$V_j = V_J \oplus \bigoplus_{k=0}^{J-j-1} W_{J-k} \tag{2.11}$$

Since the wavelet function $\psi(t) \in W_1 \subset V_0$, the wavelet function $\psi(t)$ can be generated from the scaling function $\phi(t)$. This is introduced as:

$$\psi(t) = \sqrt{2} \sum_{k=-\infty}^{\infty} g(k)\phi(2t-k) \tag{2.12}$$

The $h(k)$ and $g(k)$ are associated with coefficients of the lowpass filter (scaling filter) and the highpass filter (wavelet filter), respectively. The DWT is derived from the concept of MRA based on Equations (2.9) and (2.12). Hence any signal $x(t)$ can be represented in terms of wavelet and scaling functions as:

$$x(t) = \sum_{k=-\infty}^{\infty} c_0(k)\phi_{0,k}(t) + \sum_{k=-\infty}^{\infty} \sum_{j=0}^{J-1} d_j(k)\psi_{j,k}(t) \tag{2.13}$$

with the approximation coefficients $c_j(k)$ and the detail coefficients $d_j(k)$ at level $j = 0,1,2,\ldots,J-1$, and where J is the number of levels.

The pair of filters of wavelet decomposition, $h(k)$ and $g(k)$ are related to each other and are known as a quadrature mirror filter (QMF) pair with $g(k) = \pm(-1)^k h(L-k-1)$, $\sum h(k) = \sqrt{2}$ and $\sum g(k) = 0$, where N is the number of filter coefficients. The pair of filters for the perfect reconstruction, $\tilde{h}(k)$ and $\tilde{g}(k)$ are related to the filters of wavelet decomposition by $\tilde{g}(k) = g(L-k-1)$ and $\tilde{h}(k) = h(L-k-1)$ (Appendix C). There exists a trade-off between the filter length L and computation time. Higher filter lengths are smoother and are better able to distinguish between the different frequencies, but they require more computation time.

2.3.3.1.2 Filter bank

The DWT of any signal in $L^2(\mathbb{R})$ can be implemented by two-channel filter banks which are filtering signals with a lowpass filter $h(k)$ and a highpass filter $g(k)$ (Mallat, 2009). The filtered signals are downsampled by 2 to provide approximation coefficients $c_j(k)$ and detail coefficients $d_j(k)$ for the lowpass and highpass filters respectively. At the next level *j*, the approximation coefficients are decomposed. Such a wavelet decomposition is a recursive algorithm and provides successively coarser resolution coefficients given as:

$$c_{j+1}(k) = \sum_{n=-\infty}^{\infty} h(n-2k)c_j(n) \tag{2.14}$$

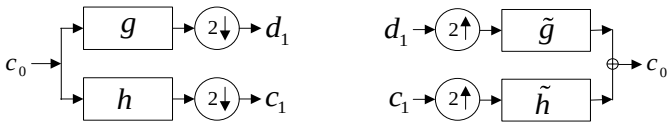
$$d_{j+1}(k) = \sum_{n=-\infty}^{\infty} g(n-2k)c_j(n) \tag{2.15}$$

Wavelet reconstruction processes by upsampling and filtering. The reconstructed signal is the sum of the approximation coefficients and the detail coefficients at a coarser resolution. This is given as:

$$c_j(k) = \sum_{n=-\infty}^{\infty} \tilde{g}(k-2n)c_{j+1}(n) + \sum_{n=-\infty}^{\infty} \tilde{h}(k-2n)d_{j+1}(n) \tag{2.16}$$

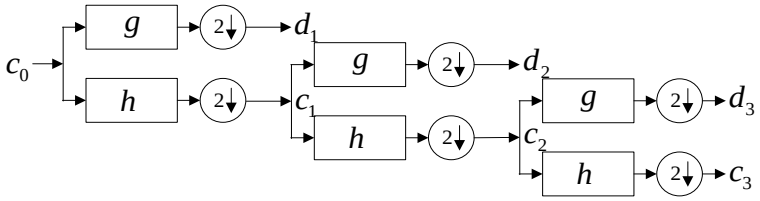
The decomposition and reconstruction of DWT can be considered as a tree-structured filter bank, as shown in Figure 2.5 and Figure 2.6, when $\{c_0(k)\}, k \in \mathbb{Z}$ denotes the input signal of wavelet decomposition and the output signal of the wavelet reconstruction. The symbols $\downarrow 2$ and $\uparrow 2$ in circles indicate the operation of downsampling by 2 and upsampling by 2, respectively. Downsampling (or decimation) by 2 means discarding all the odd or even samples of wavelet coefficients, whereas upsampling by 2 means adding zeros between the samples of wavelet coefficients.

The difference in implementation between CWT and DWT are that CWT employs all possible integer factors of dilation and translation (e.g. 2, 3, 4, and 5), while the dilation of DWT uses powers of 2. Another difference is that CWT uses only one wavelet filter while DWT uses four filters for decomposition and reconstruction.

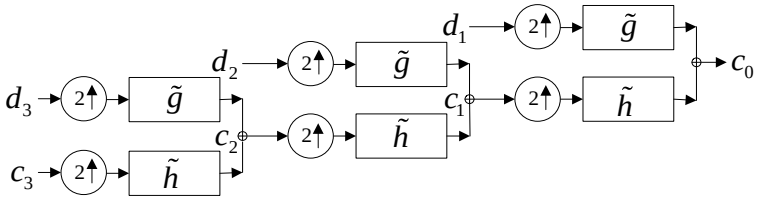


(a) (b)

Figure 2.5 DWT for the first level (J=1) decomposition (a) and reconstruction (b).



(a)



(b)

Figure 2.6 DWT for three-level (J=3) decomposition (a) and reconstruction (b).

2.3.3.2 Limitation of discrete wavelet transform

This DWT can be referred to as a standard DWT. The standard DWT suffers from some fundamental problems (Kingsbury, 2001; Selesnick et al., 2005) specifically: shift variance, and oscillation. Shift variance is the property whereby a small shift in a signal can lead to relatively large unpredictable changes of wavelet coefficients around a singularity, which is a large wavelet coefficient (Selesnick et al., 2005) and provides the most information about the signal (Peng and Chu, 2004). It can result in significant variation in the energy distribution between wavelet coefficients at different scales (Kingsbury, 2001). Generally, singularity extraction with standard DWT-based

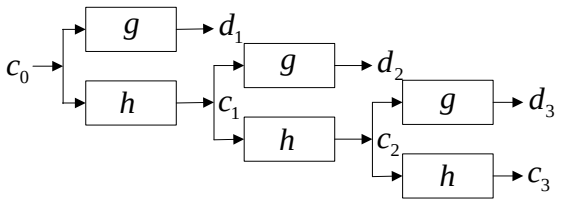
processing yields large wavelet coefficients. However, wavelet functions are bandpass filters, and these wavelet coefficients may oscillate around singularities which are overlapped to provide small or zero wavelet coefficients (Selesnick et al., 2005).

2.4 Extension of discrete wavelet transforms

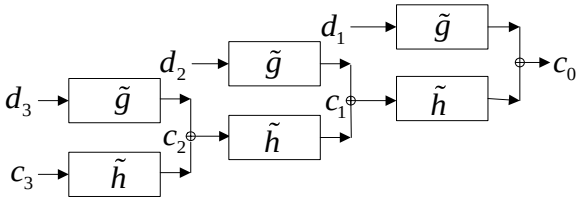
The standard DWT may not be good enough for some applications. The standard DWT can be extended to the stationary wavelet transform (SWT) and the wavelet packet transform (WPT) by changing some of procedures associated with decomposition and reconstruction in the standard DWT.

2.4.1 Stationary wavelet transforms (SWTs)

The SWT is sometimes referred to as the undecimated DWT (UDWT), redundant DWT (RDWT), or shift-invariant DWT (SIDWT) along with other terms. The SWT can be implemented by removing the up/downsampling operation in the standard DWT and inserting zeros between filter coefficients in the pair of filters. An example of the decomposition and reconstruction of SWT is shown in Figure 2.7. The frequency allocation for SWT is the same as that for DWT. The approximation coefficients $c_j(k)$ and the detail coefficients $d_j(k)$ have the same size as the input signal $c_0(k)$ at level $j = 0,1,2,...,J-1$. Hence SWT has redundancy, but not to the same degree as CWT.



(a)



(b)

Figure 2.7 SWT for three-level (J=3) decomposition (a) and reconstruction (b).

2.4.2 Wavelet packet transforms (WPTs)

In 1992, Coifman, Meyer and Wickerhauser (1992) introduced the WPT, which is a further generalisation of the standard DWT. The WPT decomposes a signal into approximation coefficients and detail coefficients and then decomposes recursively on both to give a binary tree structure. Therefore the WPT provides a much richer frequency subband of possibilities in signal analysis, which cannot be obtained by using standard DWT or SWT.

The filter bank algorithm of WPT decomposition (Mallat, 2009) can be represented by:

$$w_{j+1,2n}(k) = \sum_{p=-\infty}^{\infty} g(p-2k) w_{j,n}(p) \tag{2.17}$$

$$w_{j+1,2n+1}(k) = \sum_{p=-\infty}^{\infty} h(p-2k) w_{j,n}(p) \tag{2.18}$$

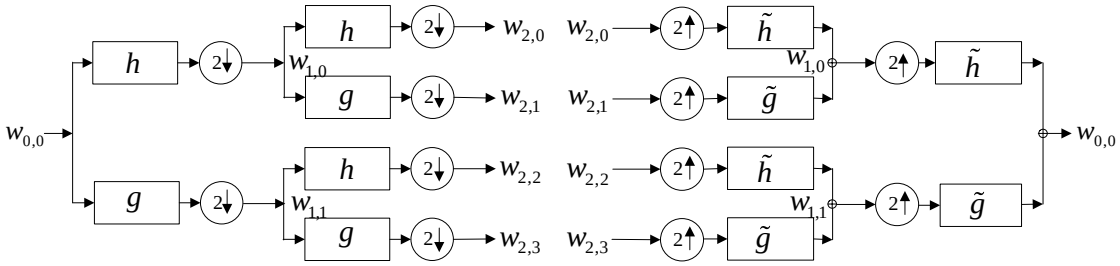
Each internal node $w_{j,n}$ in the binary tree is decomposed into child nodes $w_{j+1,2n}$ and $w_{j+1,2n+1}$. The WPT reconstruction can be expressed as:

$$w_{j,n}(k) = \sum_{p=-\infty}^{\infty} \tilde{g}(k-2p) w_{j+1,2n}(p) + \sum_{p=-\infty}^{\infty} \tilde{h}(k-2p) w_{j+1,2n+1}(p) \qquad (2.19)$$

where $w_{j,n}(k)$ are wavelet coefficients which are defined by the k^{th} index of the n^{th} subband (node) at level j in the structure of the binary tree. Hereafter, $w_{j,n}(k)$ is used as the sequence of all wavelet coefficients (i.e. approximation coefficients $c_j(k)$ and detail coefficients $d_j(k)$), which are derived by the standard DWT at level j; k is the coefficient index.

2.4.2.1 The tree-structured filter bank of WPT

The decomposition structure of WPT can be either a full binary tree (Figure 2.8) or an admissible tree (Figure 2.9). A full binary tree decomposes any signal into 2^j nodes. An admissible tree (Mallat, 2009) is a binary tree where any node has either zero or two child nodes. In other words, an admissible tree has independence to stop or continue the decomposition at any node.



(a) (b)

Figure 2.8 WPT with a full binary tree for two-level (J=2) decomposition (a) and reconstruction (b).

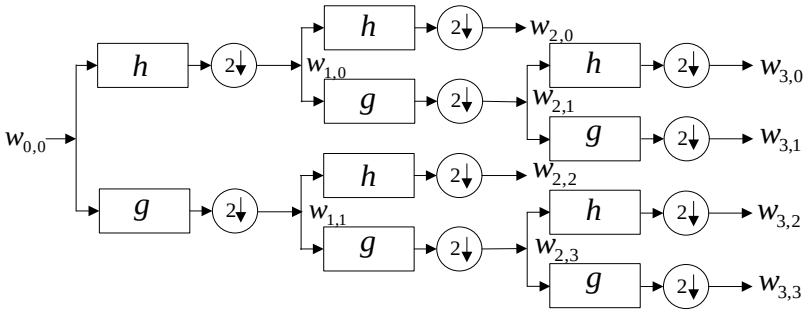


Figure 2.9 Example of WPT with an admissible tree
for three-level (J=3) decomposition.

2.4.2.2 Frequency ordering

Let E be sets of terminal nodes (j,n) and $E \subset \left\{(j,n): 0 \geq j < J, 0 \leq n < 2^j\right\}$ of a wavelet decomposition tree (Cohen, 2001). A terminal node (j,n) is associated with a subband whose bandwidth and centre frequency are given by:

$$\Delta f_{j,n} = 2^j \cdot f_s / 2 \tag{2.20}$$

$$f_{c(j,n)} = \left[GC^{-1}(n) + 0.5\right] 2^j \cdot f_s / 2 \tag{2.21}$$

where GC^{-1} is the inverse Grey code permutation of n, and f_s is the sampling rate of the signal. The lower and upper frequency of each subband is $[n, n+1] \times 2^j \cdot f_s / 2$.

2.5 Complex wavelet transforms

Complex wavelet transforms (ℂWTs) were introduced to overcome some of the limitations of the real-valued standard DWT. ℂWTs can be widely divided into two classes: dual-tree DWT-based ℂWT and projection-based ℂWT.

The well-known form of ℂWT is dual-tree DWT-based ℂWT, otherwise known as Kingsbury's dual-tree ℂWT (Kingsbury, 2001) and Selesnick's dual-tree ℂWT

(Selesnick et al., 2005). Dual-tree ℂWT employs two real-valued standard DWTs, where the first and second standard DWTs are the real and imaginary parts of the wavelet coefficients. The two real-valued standard DWTs use two different sets of filters. Both parts are operated in parallel to decompose and reconstruct the signal. The implementation of dual-tree ℂWT is illustrated in Figure 2.10. Projection-based ℂWT was introduced by Fernandes et al. (2003). This transform represents the conversion of a real signal to a complex form, followed by a DWT of the complex mapping.

The ℂWT provides advantages of reduced shift variance and improved directionality in two and higher dimensions. This has been most frequently applied in image processing and is suitable for several applications such as classification, feature extraction, motion estimation, coding, and watermarking. Further details of ℂWT and its applications are given in Kingsbury (2001), Fernandes et al. (2003) and Selesnick et al. (2005).

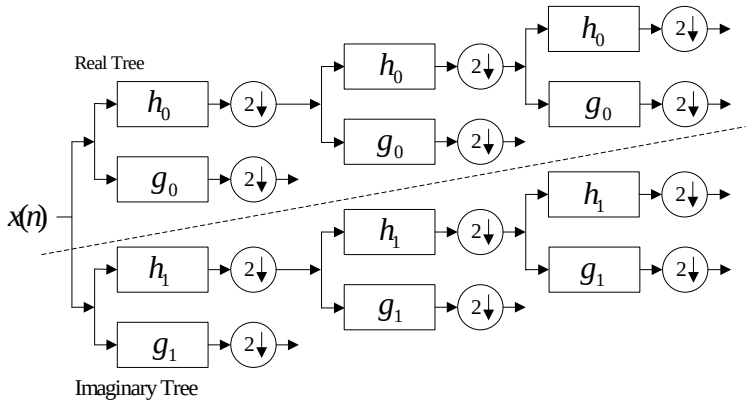


Figure 2.10 The decomposition of the dual-tree ℂWT.

2.6 Wavelets and their applications

In the late 1980s, WT has been successfully utilised and applied to various research fields. Their applications cover research areas as diverse as acoustics, speech and audio processing, image processing, telecommunications, medicine and biology, physics and seismology. The application of wavelets can be found in a vast amount of available

literature (Kingsbury, 2001; Peng and Chu, 2004; Selesnick et al., 2005; Fugal, 2009; Mallat, 2009). All wavelet-based applications may be grouped into a few main types associated with the wavelet's properties (Peng and Chu, 2004), namely time-frequency analysis, feature extraction, singularity detection, signal denoising, and data compression. Brief descriptions of some specific applications, including time-frequency analysis, signal denoising and data compression, are given here.

2.6.1 Time-frequency analysis

Time-frequency signal analysis is a powerful tool for the analysis and processing of nonstationary signals. Signals are characterised in a time-frequency plane and potentially reveal a picture of the signal's components in the temporal localisation. The WT has good utility in terms of time-frequency analysis as explained in Section 2.3. The WT yields high frequency resolution and low time resolution at low frequency. In contrast, it yields low frequency resolution and high time resolution at high frequency. Such time-frequency analysis using WT provides excellent time-frequency localised features of information simultaneously (Selesnick et al., 2005).

Several researchers have employed WT for analysing speech signals. Kadambe and Boudreauxbartels (1992) proposed a pitch detector based on the standard DWT which was suitable for both low-pitched and high-pitched speakers and was robust to noise. Tan et al. (1994) found that the SWT can locate the spectral changes of the speech signal accurately. This can be easily identified the speech into voice, plosives, fricative and silence. Voice activity detection (VAD), which is based on DWT (Stegmann and Schroder, 1997) and WPT (Chen et al., 2007), utilised the flexibility of WT in terms of time-frequency resolution to compute robust parameters for VAD decisions in noisy environments.

2.6.2 Signal denoising and data compression

The wavelet basis function has a property of compact support that provides good energy concentration information. Therefore, a singularity of the signal is large wavelet coefficients while others have small wavelet coefficients. This reflects the ability for separation between useful signal and noise. Many real-world signals are represented in

wavelet domain by a few large coefficients which are the key to sparsity (Donoho and Johnstone, 1994; Selesnick et al., 2005). The sparsity of wavelet coefficients allows near-optimal signal processing based on simple thresholding i.e. keeping the large wavelet coefficients and killing the small ones without significant errors in representing the characteristics of the signal (Selesnick et al., 2005). This is key for signal denoising and data compression.

The wavelet-based denoising approach was successfully developed by Donoho and Johnstone (1994). This method is simply performed in three steps, which are: signal decomposition, modification of wavelet coefficients with wavelet thresholding, and signal reconstruction. The wavelet coefficients can be denoised by setting all wavelet coefficients below a threshold value to zero. This can nearly optimally reduce noise while preserving the important information of the original signal. This approach has been developed and modified into several versions in the past few decades to develop techniques appropriate to different applications.

Wavelet thresholding has been widely applied in the area of speech enhancement. A variety of methods have been considered, including classical wavelet thresholding (i.e. soft thresholding and hard thresholding) (Pinter, 1996; Bahoura and Rouat, 2001; Chang, 2002; Chen and Wang, 2004; Bahoura and Rouat, 2006), modified hard thresholding based on standard DWT (Sheikhzadeh, 2001), and WPT (Ghanbari and Karami-Mollaei, 2006). Moreover, wavelet shrinkage has been effectively combined with other algorithms for speech enhancement to increase noise reduction performance in noisy speech, such as spectral subtraction (Shao and Chang, 2007) and multitaper spectra estimation (Hu and Loizou, 2004).

Various techniques of enhancing speech have been developed in the wavelet domain such as Wiener filtering (Cohen, 2001), blind adaptive filter (Veselinovic and Graupe, 2003), minimum mean square error-short time spectral amplitude (MMSE-STSA) (Tasmaz and Ercelebi, 2008), Kalman filtering (Shao and Chang, 2006), hidden Markov models (HMMs) (Shao and Chang, 2011) and blind source separation (Ashino et al., 2010; Litvin and Cohen, 2011).

Data compression becomes an economic factor for either storage or transmission of data. The idea of data compression is to use fewer bits to represent the same information at some given representation (lossless compression), or to use fewer bits to

represent the given data approximately (lossy compression). Actually, the principles of data compression are similar to signal denoising. The small wavelet coefficients can be set to zero. The greater the number of zeros, the lower the number of bits in the encoding stage of data compression. Wavelet-based data compression can often obtain a high compression ratio and maintain the singularities of signals in areas such as audio compression (Sinha and Tewfik, 1993; Srinivasan and Jamieson, 1998; Reyes et al., 2003) and speech compression (Agbinya, 1996; Carnero and Drygajlo, 1999)

2.7 Discussion and conclusion

Wavelet analysis has established a remarkable reputation as a powerful tool for signal analysis, signal denoising and data compression. The strength of WT compared to FT is in time-frequency analysis and compact support. In time-frequency analysis, the dilation and translation of WT can lead to signal analysis with variable length windows for analysing different frequency components. This allows practical and efficient representation for many types of signals (e.g. nonstationary signals), but it may not be suitable for FT. The compact support of wavelets influences the sparsity of wavelet coefficients, which is useful and important for the performance of signal denoising and data compression.

WT can be classified into continuous wavelet transforms (CWTs) and discrete wavelet transforms (DWTs). DWTs can be divided into real-valued DWTs and complex-valued DWTs. Real-valued DWTs are more appropriate for real-time applications than CWTs and complex-valued DWTs due to their lower redundancy. The class of real-valued DWTs can be divided into three general forms: standard DWT, stationary wavelet transform (SWT), and wavelet packet transform (WPT).

The filtering process for standard DWT is a recursive process with decomposing only on low frequency components and downsampling by 2. The filtering process of SWT is similar to standard DWT but the downsampling step is removed. The filtering process of WPT is similar to standard DWT, but WPT is iterated on both the low and high frequency components of the signal. The different filtering process of real-valued DWTs results in a difference in computational complexity. The computational complexities of standard DWT, SWT, and WPT are $O(n)$, $O(n^2)$ and $O(n\log_2 n)$

operations, respectively (Shukla, 2003; Mallat, 2009), where n is the length of data samples and O is a symbol used in complexity theory. Both SWT and WPT are generally higher redundancy and computational complexity than standard DWT.

WPT and FFT have the same computational complexity, which requires $O(n\log_2 n)$ operations. However, their computational complexities may different depending on computational algorithms and implementations (Mallat, 2009). The computational complexity of WPT relates to the length of filter coefficients, the decomposition levels, and the decomposition structures (i.e. a full binary tree and an admissible tree). These can result in less computational cost than FFT and bandpass filters (BPFs) (i.e. finite impulse response (FIR) filters and infinite impulse response (IIR) filters) in the same application (e.g. CI applications) (Gopalakrishna et al., 2010b).

The WPT decomposes recursively on both low (approximation coefficients) and high (detail coefficients) frequency components of the signal, but not for the standard DWT or others. Consequently, the WPT offers more frequency bands for signal analysis than the other DWTs. The WPT provides flexibility in selecting the number of frequency bands, and setting centre frequencies and bandwidths. Therefore, the WPT is more suitable for CI processors than other DWTs. In addition, WPT has more benefit than BPFs and FFT; these are summarised in Table 2.1. A comprehensive overview is provided in the next chapter (Section 3.4).

Table 2.1 Comparison of advantages and disadvantages among different filter banks in CI applications.

Criteria	**Filter banks**		
	BPFs	**FFT**	**WPT**
Signal analysis	Time domain	Frequency domain	Time-frequency domain
Temporal and spectral resolution	Good temporal resolution, but poor frequency resolution	Poor temporal resolution, but good frequency resolution	Good temporal resolution, and good frequency resolution
Configuration design of filter banks	Difficult	Simple	Simple
Computational complexity	High	Medium	Low
Loss of temporal information	High	High	Low
Loss of spectral information	Low	High	Low

Chapter 3: Wavelet packet-based speech coding strategy for cochlear implants

3.1 Introduction

In recent decades, developments in bio-signal processing have led to a trend of mimicking real bio-systems. The human auditory system has remarkable capabilities to detect, separate, and recognise speech, music and other environmental sounds (Yang et al., 1992). The functional principle of human auditory perception is incorporated in the design and implementation of human-machine communication systems, especially hearing prosthesis for analysis, synthesis, and transmission. The adoption of such auditory processing techniques has usually led to substantial improvements in the performance of these systems (Yang et al., 1992).

Many applications mimicking human auditory models can be applied to speech analysis, speech synthesis, speech coding, speech recognition, speech enhancement, room acoustics, and algorithms for the objective evaluation of speech intelligibility and quality. In hearing prosthesis, wavelet transforms (WTs) have been considered for employment in prosthesis devices as a compensation algorithm for hearing-impaired people, including multiband dynamic range compression (Drake et al., 1993), nonlinear automatic gain control in hearing aids (Li et al., 2000), noise reduction in hearing aids (Li et al., 2001) and speech processing in cochlear implants (Yao and Zhang, 2002; Gopalakrishna et al., 2010b).

Ideally, a cochlear implant (CI) would be able to imitate and replace the auditory functions of the inner ear (Zeng, 2004). An understanding of cochlear function will provide insights into many aspects of the auditory processing of speech signals. This understanding will motivate the development of novel approaches for speech processing in the auditory system in order to improve the performance of CIs. The main purpose of this chapter is to describe the important points and some details related to the design of CIs based on wavelet packet transform (WPT). Some criteria associated with CI design are considered, including filter banks, frequency scales, the structures of wavelet decomposition trees, and types of mother wavelet.

3.2 Basilar membrane model

Pioneering research by Georg von Békésy in the 1950s (Loizou, 1998) showed that the 35-mm basilar membrane (BM) running along the cochlea in the inner ear is responsible for separating received frequencies into different spatial locations along its length. A sinusoidal stimulation takes the pattern of a traveling wave that propagates from base to apex along the BM. The amplitude of the wave reaches a maximum at a particular position before slowing down and decaying rapidly. Consequently, different positions on the BM correspond to specific frequencies according to their maximum amplitudes (Figure 3.1). With the largest amplitude of displacement, high frequencies are characterised at the base while low frequencies are at the apex. A frequency that gives a maximum response at particular position on the BM is known the characteristic frequency (CF) for that position.

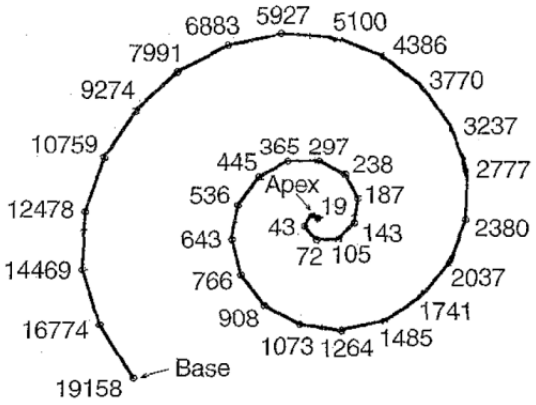


Figure 3.1 Diagram of the basilar membrane showing the base and the apex. The positions of maximum displacement in response to sinusoids of different frequencies (in Hz) are indicated (Loizou, 1998).

When complex natural sounds are decomposed into different frequency components, they produce maximum displacement at different positions along the BM. These positions on the BM can be modelled as a filter bank of a large number of overlapping bandpass filters, commonly approximately 10,000 filters. Each bandpass filter with its bandwidth has a certain centre frequency according to the characteristic

frequency. These filters are known as the auditory filter, and their bandwidth is called the critical bandwidth (CB). The critical bandwidth theory claims that the same bandwidth plays an important role in terms of harmonic discrimination, masking effects, and other psychoacoustic phenomena (e.g. perception of loudness, pitch, and timbre) (Harma et al., 2000; Chen and Wang, 2004).

The critical bandwidth and the shapes of auditory filters on the whole range of audible frequencies have been directly measured in experiments using many different methods. Experimental measures have included the absolute threshold of complex sounds, the masking of a band of noise by two tones, sensitivity to phase differences, and loudness. Most of the methods for estimating the auditory filter shape are based on assumptions about the power spectrum model of masking, such as psychophysical tuning curves, the rippled-noise method and the notched-noise method (Moore, 2008). Further details of measurements are given in Glasberg and Moore (1990) and Moore (2008).

In auditory processing, it has been found that the performance of WT filters is equivalent to the performance of auditory filters by analysing properties of the BM model (Yang et al., 1992; Yao and Zhang, 2002). The BM is sensitive to higher frequencies at the base (analysing short transients with lower frequency resolution), and it is sensitive to low frequencies at the apex (analysing long transients with higher frequency resolution). Hence a WT decomposition with similar characteristics to those of cochlear filters may be effective for speech and auditory processing.

3.3 Auditory filter banks

Auditory filter banks are bandpass filters designed to mimic the frequency resolution of human auditory perception (Smith and Abel, 1999). An ideal bandpass filter is used to separate signals by accepting signals within a desired frequency band, and to provide potentially useful spectral transforms of speech signals. The auditory filter can be considered as a weighting function, which is used in the spectrum of acoustic signals to determine the effective magnitude of the output of the filter (Glasberg and Moore, 1990). The output of the filter bank of analysis signals affects the information transmitted to the auditory nerves in the brain. The critical bandwidth of the auditory

filter is an important thing, and it can be determined by a wide variety of models based on experimental techniques.

3.3.1 Cochlear mapping

The relation between centre frequency and position on the BM has been modelled by Greenwood (1990), resulting in the cochlear frequency-position function. This function closely agrees with Békésy's cochlear coordinates. Greenwood's function is given by:

$$f_c = 165.4(10^{0.06x} - 1) \tag{3.1}$$

and the first derivative of Equation (3.1) is:

$$\frac{df_c}{dx} = 22.9 \times \frac{f_c + 165.4}{165.4} \tag{3.2}$$

where x is the location on the cochlea (in millimetres), f_c is the centre frequency (in Hertz) corresponding to that location and df / dx is the bandwidth related to a 1-mm range on the cochlea (Harma et al., 2000).

3.3.2 Auditory frequency scales

In physics, frequency is normally expressed in units of Hertz (Hz). In speech and hearing research, various frequency scales have been proposed in other units. The frequency scales and their critical bandwidths are usually based on a model of auditory filters, and they may be derived in many different ways (Harma et al., 2000), such as frequency-position maps of the cochlea, critical-band measures, or pitch scaling experiments. The frequency scales can be in the form of linear or nonlinear scales such as one-third octave, Bark, ERB, Mel and so on. Even though all the scales differ somewhat in terms of their numerical values (Zwicker and Terhardt, 1980), most of the frequency scales tend to be linear functions of frequency in the low-frequency region (0−1 kHz), and logarithmic functions in the mid- (1−5 kHz) and high-(5−8 kHz) frequency regions (Miller, 1989).

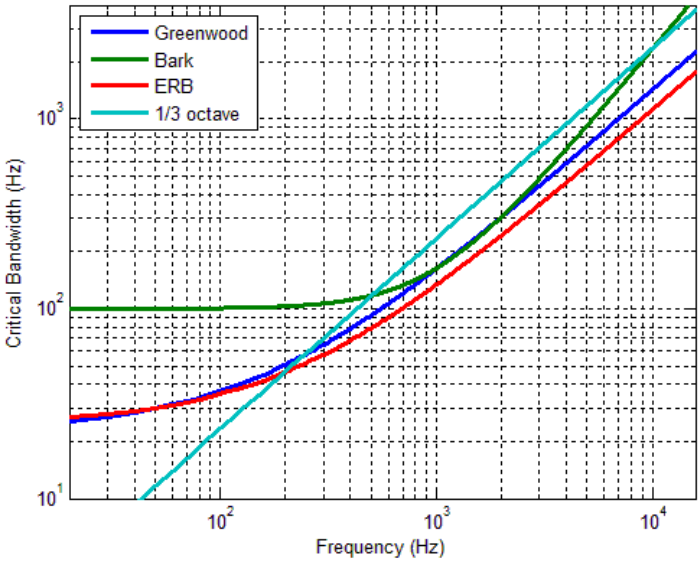


Figure 3.2 Comparison of different frequency scales.

3.3.2.1 Comparison of different frequency scales

One-third octave filter banks are usually used in the audio industry as a convenient idealisation of auditory filters. The Bark scale is derived from measurements of the characteristic frequencies of the human auditory system (Zwicker and Terhardt, 1980). It is an approximate linear scale for frequencies below 500 Hz and an approximate logarithmic scale for higher frequencies. The Mel (melody) scale has been employed based on a subjective measure of pitch magnitude. The Mel scale is parallel to the Bark scale, with a Bark unit corresponding to 100 Mels. A newer frequency scale is the ERB (equivalent rectangular bandwidth) scale (Glasberg and Moore, 1990) which is found by using the notched-noise method. The ERB scale is conceptually similar to the Bark scale.

The different frequency scales, including ERB, Bark, and one-third octave, may be compared. The critical bandwidths may be plotted on a log-log scale as in Figure 3.2 (Harma et al., 2000). The ERB is very close to the bandwidth of Greenwood's function. The critical bandwidths of ERB are narrower than those of the Bark scales, especially at frequencies below 500 Hz. For frequencies below 500 Hz, the ERB scale is neither linear, like the Bark scale, nor logarithmic, but something in between (Hermes and

Vangestel, 1991). The critical bandwidths of the Bark scale are wider at low and high frequencies, but the correspondence is excellent at the central range of hearing from 700 Hz to 4 kHz (Harma et al., 2000).

3.3.2.2 Frequency-to-place map on electrode array

The electrode array directly interfaces between the electrical output of the speech processor and the auditory neural tissue. The amount of signal energy in each frequency band of speech coding strategy should be directed to the correct position in the electrically stimulated cochlea to achieve a high level of speech recognition (Stakhovskaya et al., 2007). The use of filter banks requires the specification of the critical bandwidth in each frequency band that relates to a particular electrode. There are two important limitations for specified critical bandwidths (Fourakis et al., 2004).

The first is that electrode insertion cannot be accurately aligned with the tonotopic organisation of the cochlea, which is that the apical part of the cochlea encodes low frequencies, while the basal part encodes high frequencies. This is because of the individual insertion depth of the electrode array and the total length of the electrode array, which are dependent on the type of implant (Baumann and Nobbe, 2006; Fan-Gang et al., 2008). Moreover, each manufacturer uses a different electrode array in terms of both the number of electrodes and the electrode spacing (Fan-Gang et al., 2008) (Appendix D.1). Consequently, the intended pitches for perception differ from those that are actually perceived. The speech signal is therefore less intelligible. In addition, the speech sounds unnatural and “high-pitched” or “Donald Duck-like” (Loizou, 1998). The assigned centre frequencies of electrodes should correspond as closely as possible to the positionally determined frequencies along the cochlea.

The second limitation is that there is currently no provision for the programming audiologist to specify the frequency ranges of critical bandwidth values to electrodes when creating speech processor programmes (referred to as MAPs) for individual CI users. The critical bandwidth for the Nucleus processor is commonly specified through frequency tables (Cochlear, 2002) as part of the programming software for the creation of MAPs for individual CI users. The existing frequency tables do offer some flexibility in the number of filters (8 to 22) and the electrodes that can be allocated to different frequency ranges of the incoming speech signals.

The first limitation is of great importance. Many researchers have studied the pattern of electrical stimulation delivered to CI electrodes in relation to the depth and angle of electrode insertion, but this issue is beyond the scope of this thesis. The latter limitation is considered in simulating speech processing in CIs. Although Greenwood's function has been used to estimate the centre frequency of an electrode array (Baumann and Nobbe, 2006; Stakhovskaya et al., 2007), none of the frequency-to-electrode allocations are actually matched to the Greenwood function.

There are possible limitations of the application of the Greenwood function to CIs (Stakhovskaya et al., 2007). The first limitation is that the Greenwood function may provide accurate estimates of the frequency-to-electrode allocations only if the position of spike initiations in the electrical excitation of the spiral ganglion cell is close to the organ of Corti. Another important limitation is that accurate estimates of the position of the electrodes in the cochlea require knowledge of the total length of the organ of Corti, which cannot be determined in most temporal bone and imaging studies. Frequency estimation using the average length of the organ of Corti may be inaccurate due to substantial individual variability. Therefore, most commonly the speech frequency range is divided up between the available electrodes, regardless of the depth of insertion.

Different manufacturers have different approaches to frequency-to-electrode allocation. Some speech coding strategies use a logarithmic scale, while other CI processors use both linear and logarithmic scales (Loizou, 2006). The existing filter bandwidths of the Nucleus processor, as specified by the manufacturers, do not explicitly define a certain filter bank approach (Nogueira et al., 2005). The filter bandwidths are linearly spaced below 1 kHz, and logarithmically spaced above 1 kHz. The recommended frequency tables for the Nucleus processor (Cochlear, 2002), especially 128-point FFT, are almost the same as the Bark scale (Nogueira et al., 2006) as shown in Figure 3.4 (a).

3.4 Wavelet packet filter banks

All CI speech coding strategies are based on a filter bank approach, which is the first stage of speech processing. They use a filter bank which decomposes the speech signals

into multiple frequency channels to determine the characteristics of auditory filters and provide spectral and temporal information. There are different implementations of filter banks, including finite impulse response (FIR) filters, infinite impulse response (IIR) filters (Buechner et al., 2009), short-time Fourier transform (STFT) (Cochlear, 2002), and wavelet transforms (WT).

The use of bandpass filters (BPFs), namely FIR and IIR filters, is based on signal analysis in the time domain. BPFs provide good temporal resolution, but limited frequency resolution by the number of channels (Gopalakrishna et al., 2010b). In addition, their filter configurations make it difficult to design critical bands (Nie *et al*., 1998). Although BPFs provide good temporal resolution, the temporal information is limited by lowpass filters at the envelope detection stage. In CI processors with noise reduction, the noise is usually mixed with the speech signal across the entire frequency band. This may be difficult to achieve by means of BPF techniques (Yao and Zhang, 2002). In contrast, the signal in the time domain is transformed into other domains (e.g. STFT and WT), and transformed signals can be easily discriminated between speech and noise signals. This is more useful for denoising techniques.

The use of STFT is based on signal analysis in the frequency domain. STFT provides good frequency resolution, but limited temporal resolution by the update frame rate (Gopalakrishna et al., 2010b). Therefore, very high stimulation rates can be obtained by increasing the overlap between analysed frames, and this may not necessarily provide new information. In other words, there is a lack of temporal information improvement with high stimulation rates (Loizou, 2006). Moreover, the temporal resolution of filter banks implemented by the speech processor and the temporal resolution determined by its stimulation rates may be misaligned (Nogueira et al., 2006). These result in limitations in speech perception. Nevertheless, STFT is more efficient in terms of speed than BPFs.

The use of WT is based on signal analysis in the time-frequency domain. The WT approach is introduced to address the limitation of STFT implementation in terms of temporal resolution (Nogueira et al., 2006; Gopalakrishna et al., 2010b). The temporal resolution should sufficiently represent the temporal features of speech information, and higher temporal resolution can lead to better speech perception (Buechner et al., 2009). Time-frequency analysis of WT is similar to human auditory perception (Yao and

Zhang, 2002; Derbel et al., 2008) and can be adapted to the time-frequency features of CI systems (Nogueira et al., 2006).

In principle, the overall CI stimulation rate (e.g. the *n*-of-*m* strategy) is constrained by the channel stimulation rate (temporal resolution) and the number of selected channels. In order to design an optimum configuration for the CI stimulation rate, the balance between temporal and frequency resolution may be alleviated by using wavelet-based strategies (Yao and Zhang, 2002).

The implementation of a wavelet-based speech coding strategy proposed by Gopalakrishna et al. (2010b) provides a lower amount of spectral leakage, allows for high stimulation rates and achieves lower computational complexity compared to other commonly used strategies in CIs. The spectral leakage is a good measure to indicate how much the energy of one frequency band is leaked into adjacent frequency bands. Hence, the lower spectral leakage leads to better performance of CI processors in terms of good frequency specificity and less distortion of information. Gopalakrishna et al. (2010b) have shown that the WPT-based strategy yielded lower spectral leakage than that obtained with a STFT-based strategy, but it was almost the same as the BPF-based strategy.

High stimulation rates can provide better fine temporal representation of speech information than low stimulation rates. This strategy can provide a high stimulation rate, which is equal to the sampling rate of the input signal. This can lead to better speech recognition performance, especially in some CI devices with sufficiently wide electrode spacings (Loizou, 2006). Due to the increased channel interaction concomitant with a high stimulation rate, wider electrode spacings provide smaller amounts of channel interaction at the same high stimulation rate. Therefore, most of the benefits of high stimulation rates were reported by Med-El CI users (the widest electrode spacing) but not with Nucleus CI users (the smallest electrode spacing).

Gopalakrishna et al. (2010b) have shown that the WT approach has lower computational complexity than BPF and STFT implementations. Different structures of WPT have different computational complexity. WPT with an admissible tree has lower computational complexity than WPT with a full binary tree. This is because WPT with an admissible tree can be designed directly for electrodes. The low computational complexity can reduce memory requirements, save in execution time and minimize the

power requirement. This offers important advantages of minimised device requirements, real-time implementation and prolonged battery life for CI processors.

Overall, the WT offers some important advantages over both BPFs and STFT, such as a simple design in terms of filter configuration, good spectral and temporal resolution, low computational cost, and appropriate properties for speech coding and denoising as discussed in Chapter 2.

The evolution of wavelet-based strategies for CI system has been developed over the last fifteen years. These can be classified into two groups: continuous wavelet transforms (CWTs) (Yao and Zhang, 2002; Cheikhrouhou et al., 2004; Guan et al., 2005; Derbel et al., 2008) and discrete wavelet transforms (DWTs) (Nie et al., 1998; Nogueira et al., 2006; Paglialonga et al., 2006; Paglialonga et al., 2008; Gopalakrishna et al., 2010b).

A commonly-used mother wavelet in CWTs is the complex Morlet wavelet, because it is easy to select centre frequencies and the bandwidth such that they match the Bark scale or ERB scale (Cheikhrouhou et al., 2004; Derbel et al., 2008). The bionic wavelet transform (BWT) is derived from the Morlet wavelet and has also been used to develop adaptive wavelet strategies (Yao and Zhang, 2002; Derbel et al., 2008). There is a difference between the CWT and the BWT. The window size of the CWT varies with the analysing frequency, but all windows at a certain scale along the time-axis are fixed. The window size of the BWT can be adjusted in the same scale. The BWT achieves a better trade-off between time and frequency resolution and preserves more of the energy of the signal than the CWT (Yao and Zhang, 2002). However, the CWT and BWT both produce redundancies and have a high computational cost.

DWTs are implemented in a filter bank decomposition approach. The standard DWT-based speech coding strategies (Nie et al., 1998) provide a fast and efficient algorithm. Their results are consistent with IIR bandpass filters in terms of the characteristics of the waveforms in each band. Their speech recognition performance is also similar to that of the ACE and CIS strategies (Paglialonga et al., 2006; Paglialonga et al., 2008). However, the filtering process of the standard DWT is iterated only on low frequency components. This provides a limited number of channels and limited frequency ranges in each channel, making them inappropriate to apply to CI devices.

A further generalisation of the standard DWT is the wavelet packet transforms (WPTs). The filtering process of the WPT is iterated in both the low and high frequency components. The filter banks of the WPT can be varied over the frequency ranges and the decomposition structure can be simply adjusted for the approximation of critical bands. An appropriate structure of WPT can closely mimic the critical band according to a perceptual auditory model.

Many WPT structures are designed based on the Bark scale, the Mel scale and the ERB scale. A commonly-used frequency scale for the critical band is the Bark scale, which has been widely used by speech researchers. Bark scale wavelet packets are used in many applications such as wavelet packet-based CIs (Nogueira et al., 2006; Gopalakrishna et al., 2010b), speech enhancement (Carnero and Drygajlo, 1999; Cohen, 2001; Chen and Wang, 2004; Shao and Chang, 2007), source separation (Litvin and Cohen, 2011), and speech compression (Carnero and Drygajlo, 1999).

A small body of research has studied CI speech coding strategies based on wavelet packet filter banks (Behrenbruch and Lithgow, 1998; Nogueira et al., 2006; Gopalakrishna et al., 2010b). Recently, wavelet packet-based strategies have been successfully produced for real-time implementations (Gopalakrishna et al., 2010b). Moreover, it provides better speech recognition performance than a commercial ACE strategy for CI users at 15 dB SNR (Nogueira et al., 2006). Wavelet packet-based strategies are expected to be used in future generations of CIs (Gopalakrishna et al., 2010b).

Designing wavelet packet filter banks relates to the selection of the perceptual auditory model, the structure of wavelet packet-decomposition trees and mother wavelets. Further details are given in the next section.

3.4.1 The Bark frequency scale

The concept of the Bark frequency scale assumes that the width of critical bands of human hearing is one Bark. That means that the distance of the bandwidth from the lower band edge to the upper band edge is 1 Bark. The representation of energy over the Bark scale closely corresponds to the obtained information processing in the ear.

Theoretically, the Bark scale ranges from 1 to 24 Barks; the bandwidth of Bark scales is only defined up to 15.5 kHz for the highest sampling rate at 31 kHz. The frequency range of human auditory processing ranges from 20 to 20000 Hz and covers a total of 25 Barks (Carnero and Drygajlo, 1999; Smith and Abel, 1999).

The relation of frequency and critical band rate z can be approximately expressed by:

$$z(f) = 13\tan^{-1}(7.6\times10^{-4} f) + 3.5\tan^{-1}(1.33\times10^{-4} f)^2 \quad \text{[Bark]} \quad (3.3)$$

The critical bandwidth (CB) can be found by:

$$CB(f) = 25 + 75(1+1.4\times10^{-6} f^2)^{0.69} \quad \text{[Hz]} \quad (3.4)$$

where f is the frequency in Hertz (Hz). In the CI system, the underlying sampling rate is selected to be 16 kHz which produces a bandwidth of 8 kHz. The WPT decomposes the frequency range [0 8] kHz into 22 subbands as listed in Table 3.1.

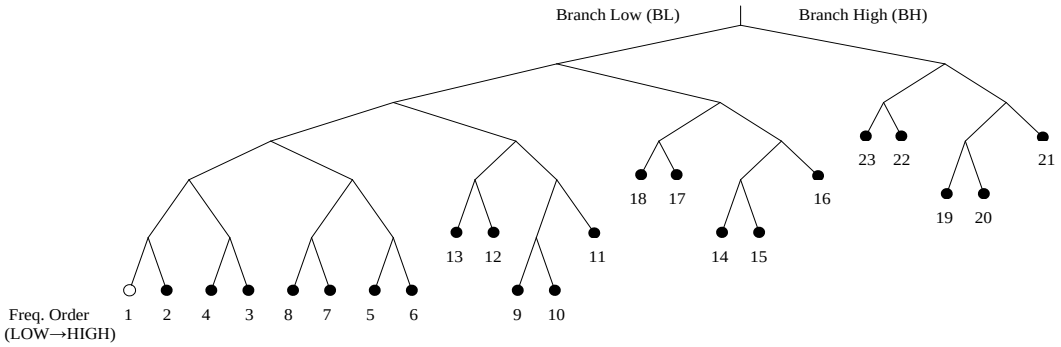
3.4.2 Structure of Bark scale wavelet packet

Two different structures of wavelet packet-decomposition tree are shown in Figure 3.3 (Gopalakrishna et al., 2010b); these will be used in the experiment. Structures with both an admissible tree (Figure 3.3 (a)) and a full binary tree (Figure 3.3 (b)) are generated from a six-level decomposition of the WPT. A given node connects the left and right branches to its child nodes. The left and right branches denote lowpass and highpass filter, respectively. Therefore, the left and right child nodes also correspond to a lower and a higher frequency component, respectively. Consequently, the frequency order of each node changes its position in the wavelet packet tree as shown in Figure 3.3. The frequency ordering is explained in Section 2.4.2.2.

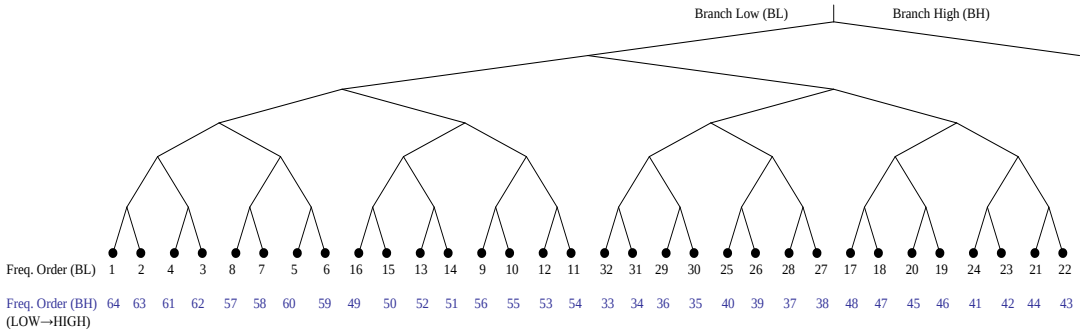
The 23-band WPT is designed to directly approximate the critical bands of the human auditory system using the 22 channels available to the Nucleus-24 processor. The 23 subbands are selected from the entire set of wavelet packet bands. Consequently the lowest frequency band, shown as a white node in Figure 3.3 (a), is not used, because

this frequency band plays no significant role in speech perception (Nogueira et al., 2006).

The six-level decomposition of WPT with a full binary tree consists of 64 nodes ($2^6 = 64$). The 64-band WPT with a frequency spacing of 125 Hz is grouped together to obtain 22 channels with different frequency bands. The 64-band WPT is treated like the 64 FFT bins by the Nucleus-24 processor. The filter spacing of the 64-band WPT is allocated by using a linear spacing in the low frequencies (≤1 kHz) and logarithmic spacing thereafter (>1 kHz) (Cochlear, 2002). Therefore, the bandwidth (Δf) and centre frequencies (f_c) of the 23-band and 64-band WPT for 22 channels at a 16 kHz sampling rate are slightly different, as shown in Table 3.1 (Cochlear, 2002; Shao and Chang, 2007; Gopalakrishna et al., 2010b) where f_l and f_u are the lower and upper frequencies, respectively. Figure 3.4 compares the difference between centre frequency and bandwidth of wavelet packet tree and those of Bark scales.

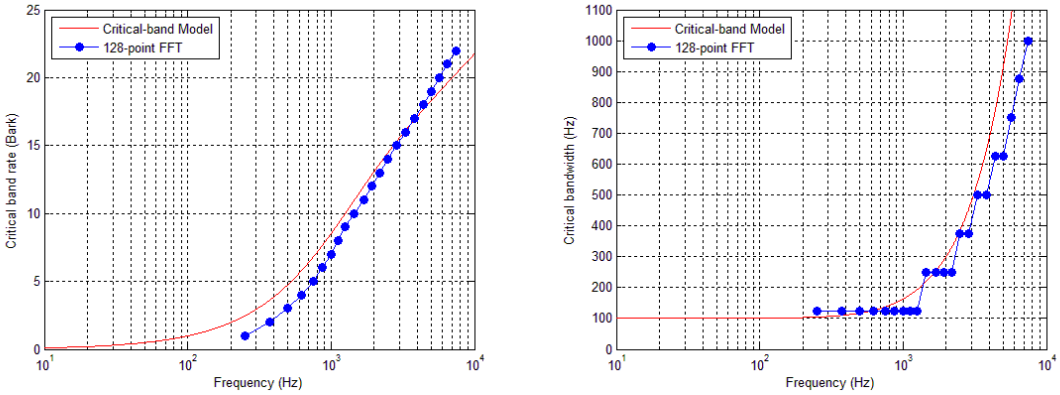


(a) 23-band WPT

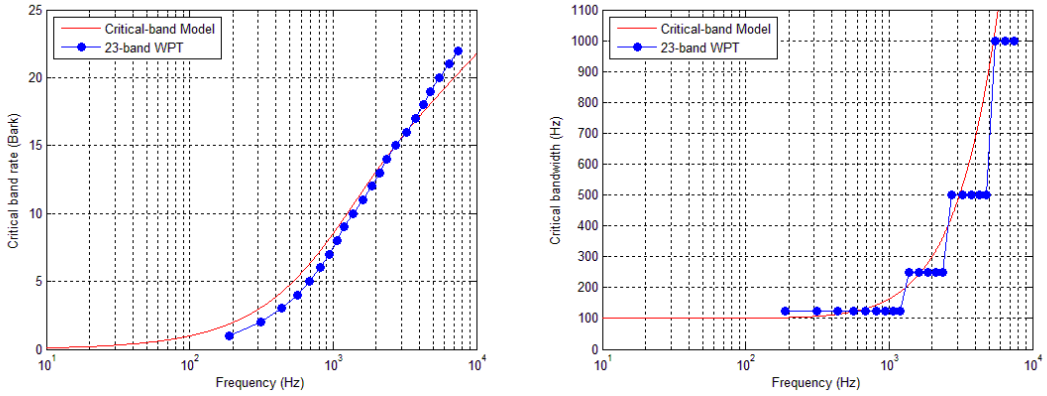


(b) 64-band WPT

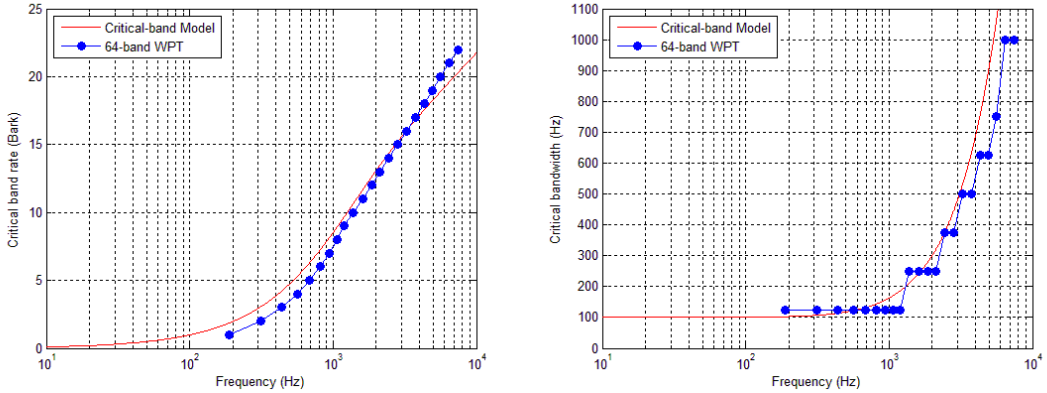
Figure 3.3 Two structures of wavelet packet-decomposition tree.



(a) 128-point FFT



(b) 23-band WPT



(c) 64-band WPT

Figure 3.4 Comparison of WPT and 128-point FFT with Bark scale: centre frequencies (left) and bandwidths (right).

Table 3.1 Frequency band and centre frequency in each channel at 16 kHz sampling rate

Electrode channel number	Bark Scale			128-point FFT		
	$[f_l \quad f_u]$	f_c	Δf	$[f_l \quad f_u]$	f_c	Δf
22	0-100	50	100	188-313	250.0	125
21	100-200	150	100	313-438	375.0	125
20	200-300	250	100	438-563	500.0	125
19	300-400	350	100	563-688	625.0	125
18	400-510	450	110	688-813	750.0	125
17	510-630	570	120	813-938	875.0	125
16	630-770	700	140	938-1063	1000.0	125
15	770-920	840	150	1063-1188	1125.0	125
14	920-1080	1000	160	1188-1313	1250.0	125
13	1080-1270	1170	190	1313-1563	1437.5	250
12	1270-1480	1370	210	1563-1813	1687.5	250
11	1480-1720	1600	240	1813-2063	1937.5	250
10	1720-2000	1850	280	2063-2313	2187.5	250
9	2000-2320	2150	320	2313-2688	2500.0	375
8	2320-2700	2500	380	2688-3063	2875.0	375
7	2700-3150	2900	450	3063-3563	3312.5	500
6	3150-3700	3400	550	3563-4063	3812.5	500
5	3700-4400	4000	700	4063-4688	4375.0	625
4	4400-5300	4800	900	4688-5313	5000.0	625
3	5300-6400	5800	1100	5313-6063	5687.5	750
2	6400-7700	7000	1300	6063-6938	6500.0	875
1	7700-9500	8500	1800	6938-7938	7437.5	1000

Electrode channel number	23-band WPT			64-band WPT		
	$[f_l \quad f_u]$	f_c	Δf	$[f_l \quad f_u]$	f_c	Δf
22	125-250	187.5	125	125-250	187.5	125
21	250-375	312.5	125	250-375	312.5.	125
20	375-500	437.5	125	375-500	437.5	125
19	500-625	562.5	125	500-625	562.5	125
18	625-750	687.5	125	625-750	687.5	125
17	750-875	812.5	125	750-875	812.5	125
16	875-1000	937.5	125	875-1000	937.5	125
15	1000-1125	1062.5	125	1000-1125	1062.5	125
14	1125-1250	1187.5	125	1125-1250	1187.5	125
13	1250-1500	1375.0	250	1250-1500	1375.0	250
12	1500-1750	1625.0	250	1500-1750	1625.0	250
11	1750-2000	1875.0	250	1750-2000	1875.0	250
10	2000-2250	2125.0	250	2000-2250	2125.0	250
9	2250-2500	2375.0	250	2250-2625	2437.5	375
8	2500-3000	2750.0	500	2625-3000	2812.5	375
7	3000-3500	3250.0	500	3000-3500	3250.0	500
6	3500-4000	3750.0	500	3500-4000	3750.0	500
5	4000-4500	4250.0	500	4000-4625	4312.5	625
4	4500-5000	4750.0	500	4625-5250	4937.5	625
3	5000-6000	5500.0	1000	5250-6000	5625.0	750
2	6000-7000	6500.0	1000	6000-7000	6500.0	1000
1	7000-8000	7500.0	1000	7000-8000	7500.0	1000

3.4.3 Mother wavelet

The selection of mother wavelets or wavelet filters is essential for CI processors. Some of the most well-known mother wavelets are Haar, Daubechies, Coiflets, Symlets, Meyer and Biorthogonal wavelets (Appendix C). The different mother wavelets are used in the CI processor based on wavelet packet filter banks such as Haar (Nogueira et al., 2006), Daubechies (Nogueira et al., 2006; Gopalakrishna et al., 2010b), Symlets (Gopalakrishna et al., 2010b), and mixed mother wavelets (Daubechies and Symlets) (Nogueira et al., 2006).

The complicated computations and the aliasing of the speech coding strategy also depend directly on the filter length of the mother wavelet. The longer the filter length, the more complex the computation and the longer the processing time (Nogueira et al., 2006; Gopalakrishna et al., 2010b). The Haar wavelet is the simplest method of implementation, but it may be limited in terms of filter lengths. This leads to a worse frequency resolution and aliasing in each level of wavelet decomposition (Nogueira et al., 2006). Daubechies and Symlets with various filter lengths have similar results (Gopalakrishna et al., 2010b). However, the most reasonable strategy for selecting optimal mother wavelets may be chosen by a comparison of the analysis results among these mother wavelets (Sang et al., 2009). For this thesis, a Symlet with order 8 (filter length of 16) yielded the best information envelope and electrodogram compared to other wavelet filters.

3.5 Speech coding strategy

The stages of wavelet packet-based speech coding strategies are similar to those in the ACE strategy. The ACE strategy is a FFT-based speech coding strategy and an *n*-of-*m* channel selection strategy (Nogueira et al., 2005; Loizou, 2006). A signal is decomposed into *m* channels and only the *n* most important channels are selected. A set of processing parameter values used by an individual CI user are collected by MAP (Fourakis et al., 2004), such as the centre frequencies of the channels and corresponding bandwidths, the number of channels selected, the channel stimulation rate, the implant stimulation rate, the current threshold level, and the current comfort level (Appendix

D.2). The details of ACE and wavelet packet-based speech coding strategies are as follows.

3.5.1 Advanced Combination Encoder (ACE) strategy

The analysis stages of the ACE strategy in a Nucleus-24 processor are as follows. The speech signal is captured by a microphone at a sampling rate of 16 kHz, and it is first pre-emphasised by a filter that amplifies high-frequency components in particular. The emphasised signal is windowed using a hanning window (8 ms, N=128 samples). The overlapping window adapts to the channel stimulation rate – for example, a 75% overlap for a channel stimulation rate of 500 pps, and a 90% overlap for a channel stimulation rate of 1200 pps. After that, the FFT is used to decompose the windowed signal into frequency bands. The 128-point FFT provides 128 spectral coefficients (128 bins). Due to the symmetry property of FFT, the first 64 bins are used and the second 64 bins are discarded without loss of information.

The 64 FFT bins with linear spacing are rearranged to mimic the critical bands of the auditory system by summing the powers of adjacent bins to provide 22 channels with different frequency ranges. The frequency range in each channel is based on a critical band, and is defined by the frequency table of the Cochlear Corporation (Cochlear, 2002). The apical one-third of the channels are allocated with a linear spacing to frequencies up to 1 kHz, and the basal two-thirds of the channels are allocated with logarithm spacing to frequencies above 1 kHz.

The power of the envelope in each channel is calculated as a weighted sum of the FFT bin powers. The envelope channels with the largest amplitude are selected for stimulation. In clinical practice, 8 to 12 maximum envelopes (Hu and Loizou, 2008; Gopalakrishna et al., 2010b) are selected and compressed to fit with the individual CI user's dynamic range between threshold and comfortable loudness levels. Finally, the compressed amplitudes are used to modulate the stimulating pulse which is delivered to the implanted electrode. In each frame of the speech signals, n electrodes are stimulated sequentially and one cycle of stimulation is completed (Nogueira et al., 2005). The number of pulses/second (pps) thus determines the rate of stimulation on a single

channel, also known as the channel stimulation rate. Further details of the ACE strategy are provided in Appendix D.3.

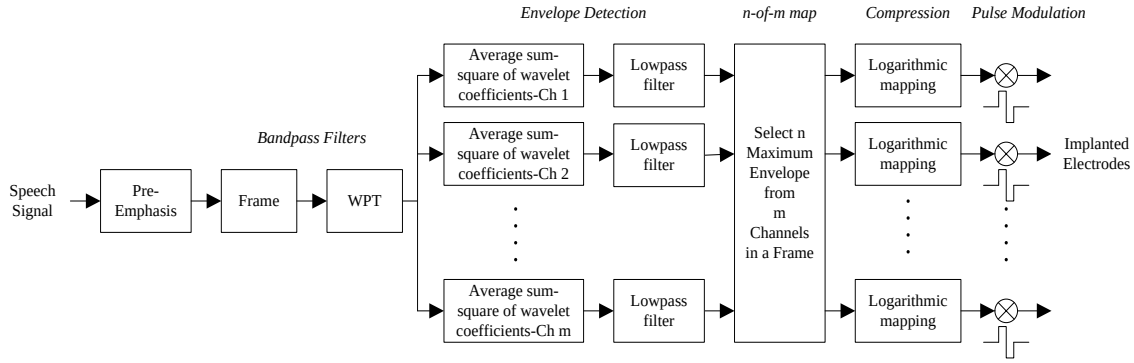
3.5.2 Wavelet packet-based speech coding strategy

In CI processors, the stages of a wavelet packet-based speech coding strategy are similar to those of the ACE strategy. A block diagram of the analysis stage in a CI processor is shown in Figure 3.5 (a). The speech signal is recorded by a microphone at a 16 kHz sampling rate and is initially pre-emphasised by a first-order Butterworth filter that amplifies high-frequency components between approximately 1.5 kHz and 5 kHz. The pre-emphasis signal provides the frequency response associated with the HS8 microphones in Nucleus processors.

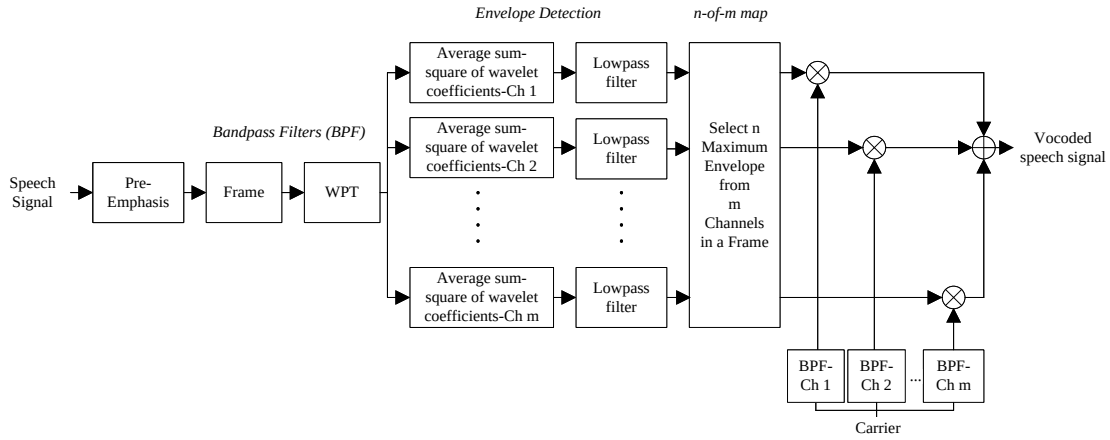
After pre-emphasis the signal is processed frame by frame using a sliding window of 128 samples (8 ms) with an overlap of 75% and a channel stimulation rate of 500 pps. The window overlapping technique is the same as in ACE. It is adapted to the channel stimulation rates in the CI user's MAP. The higher the channel stimulation rate, the greater the overlapping of windows and the temporal information. The signal in each frame is then decomposed into different frequency bands using the WPT. The spectral coefficients of the WPT in each band differ from those of FFT in the ACE strategy in each band. WPT consists of a number of wavelet coefficients, whereas FFT contains only spectral coefficients. The number of wavelet coefficients in each band depends on the decomposition levels.

The power in each band is computed using the average sum-square of the wavelet coefficients. In the 64-band WPT, the 64 frequency bands are computed by summing the power of consecutive frequency bands with frequency ranges used in the Nucleus-24 processor to generate 22 channels (Cochlear, 2002). The power per band is weighted following the ACE strategy. The envelopes are smoothed with a low-pass filter. The 12 maximum envelopes (12-of-22 channels) are selected and compressed to fit within the electrical dynamic ranges defined by the CI user's threshold and comfort levels. Finally, the compressed amplitudes are used to modulate the stimulating pulses and sent to the implanted electrodes.

In a vocoder simulation or CI hearing simulation (Figure 3.5 (b)), acoustic models can be thought of as producing vocoded signals. Vocoded signals are used to test NH listeners. The noise-band vocoder is most commonly used, and provides the most natural sound. The 12 maximum envelopes (12-of-22 channels) are selected and then used to modulate white noise, which is filtered by the bandpass filter in the same channel as the WPT. A vocoded speech signal is synthesised by summing the modulated signals of each channel.



(a) Analysis stages in the speech coding strategy.



(b) Vocoder simulation.

Figure 3.5 Wavelet packet-based speech coding strategies. Adapted from (Gopalakrishna et al., 2010a) and Mourad Ghrissi (2012).

3.6 Conclusion

WPT has some advantages for CI processors. The filter banks of WPT provide flexibility in specified frequency ranges. The decomposition structure can be simply adjusted in relation to auditory-inspired frequency components to match a perceptual auditory scale such as the Bark scale. The property of PWT has a trade-off between time and frequency representation which produces a good match of signals and localises energy concentration with few large coefficients. In addition, WPT is more efficient in terms of speed than bandpass filters and STFT. Such advantages can lead to appropriate designs and the effective development of speech coding strategies in CI system.

Chapter 4: Noise reduction in wavelet packet-based speech coding strategy

4.1 Introduction

Since humans live in a natural environment where noise is everywhere and unavoidable, ambient noise is generally merged into speech signals. This background noise causes a speech degradation, which can lead to overall unintelligibility and decreases the performance of speech coding, speech recognition and communication applications considerably (Chen et al., 2006). Therefore, techniques for efficient noise reduction in realistic listening environment are required, especially for hearing impaired (HI) listeners.

The speech-reception threshold (SRT) for sentences (50% correct) in noisy environments can be explained with signal-to-noise ratio model (Festen, 1987; Festen and Plomp, 1990). The SRT of NH listeners is reached approximately -5 dB SNR. The SRT of HI listeners is reached approximately up to 10 dB SNR, which depends on hearing loss (Festen, 1987). Noise reduction algorithms would be beneficial to HI listeners at higher SNR levels. Some noise reduction algorithms may work well for HI listener, but not work for NH listeners. Generally, HI listeners require perfectly noise reduction algorithms to match their individual hearing capabilities, increase their comfort level when listening, and improve their speech intelligibility (Ephraim and Cohen, 2004).

Noise reduction in speech processing is a complicated problem for a number of reasons (Chen et al., 2006). First, the nature and complex characteristics of speech and noise signals vary over time and may change from one application to another. It is very difficult and complicated to develop an adaptable algorithm that will work in different environments. Another reason is that the purpose of noise reduction depends on the specific context and application. Some applications need to increase intelligibility and quality or improve overall speech perception, while others aim to improve the accuracy of automatic speech recognition systems or simply to decrease the listener's fatigue. It is not easy to satisfy all purposes at the same time.

Generally, there are three stages in noise reduction techniques for speech processing (Loizou et al., 2007); these are analysis, suppression and synthesis (Figure 4.1). The analysis stage is when the speech signal is transformed into another domain. This relies on the capability of discriminating between speech and noise. The larger the difference between speech and noise signals, the more reduction there may be in the noise signals. The suppression stage is the main stage of most algorithms. The transformed signal is modified or weighted by multiplying with a gain function (suppression function) to control noise reduction across a wide range of SNR levels. Finally, at the synthesis stage the modified signal is transformed back to the time domain.

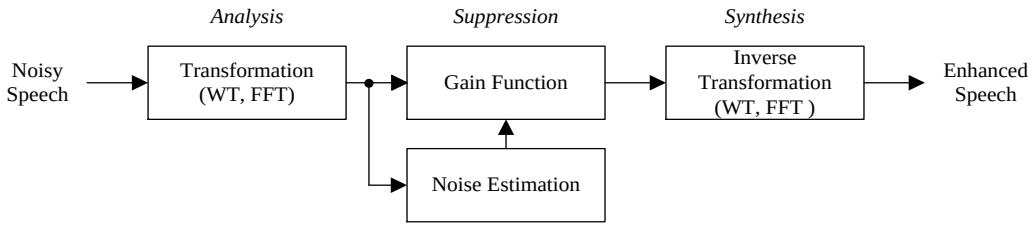


Figure 4.1 Three main stages in noise reduction techniques.

The parameters of optimal gain functions involve noise estimation. This noise estimation should continuously adapt to access information for noise spectrums in different noisy environments (Martin, 2001; Loizou et al., 2007). Where there is no prior information about noise sources, adaptive techniques using the statistical properties of speech and noise are usually used to accurately track noise. The noise level should not be under- or overestimated. An accurate noise estimate can effectively denoise and highly enhance speech. In contrast, overestimated noise may lead to the removal of speech information, further distortions in enhanced speech and reductions in speech intelligibility. Meanwhile underestimated noise may lead to greater amounts of residual noise. Therefore, the optimal gain function should be a trade-off between the amount of noise reduction, speech distortion, and the level of residual noise (Virag, 1999).

When the Bark scale wavelet packets, which reflect the human auditory system, are combined with an appropriate gain function, this may lead to the improvement of speech intelligibility and quality (Cohen, 2001; Chen and Wang, 2004). Two noise reduction algorithms, namely time-frequency spectral subtraction (TFSS) and time-adaptive wavelet thresholding (TAWT), are applied in wavelet packet-based speech coding strategies. Both algorithms are compared with ideal binary masking (IdBM) as a baseline for denoising performance.

The IdBM is used for noise reduction where information about clean speech and noise is known. The TFSS and TAWT algorithms are applied in this study since both have some main advantages. These approaches are simple in their implementation, which only requires an estimation of the noise spectrum. They offer high flexibility in the variation of parameters to compromise between noise reduction and speech distortion. Additionally, they do not require the explicit voice activity detection (VAD). For these reasons, both are suitable for the real-time implementation of CI systems in diverse environments.

This chapter is organised as follows. The concept of combined noise reduction and speech coding in a wavelet packet-based speech coding strategy is presented. The noise reduction algorithms selected for use in this study are presented next. The section on performance measurements presents various measures - both visual inspection (e.g. waveform and electrodogram) and objective speech intelligibility measures- that are used for evaluating the algorithms before they are tested with NH listeners. A summary is given in the last section.

4.2 Combined noise reduction and speech coding strategy

The speech coding strategies in CI processors are almost the same as noise reduction algorithms, in which the tasks is to decompose the signal into uncorrelated components and then process these components separately. Thus, both algorithms can be combined into one system, using a common processing structure to decrease the computational load and the complexity of the system.

This concept has been widely applied in data compression (Sinha and Tewfik, 1993; Srinivasan and Jamieson, 1998; Carnero and Drygajlo, 1999) and some noise reduction algorithms in CIs (Hu et al., 2007; Hu and Loizou, 2008; Li, 2008; Kokkinakis et al., 2011; Hu et al., 2013). To reduce the effect of noisy backgrounds, noise reduction algorithms are integrated into wavelet packet-based speech coding strategies to reduce noise directly in noisy envelopes (Figure 4.2).

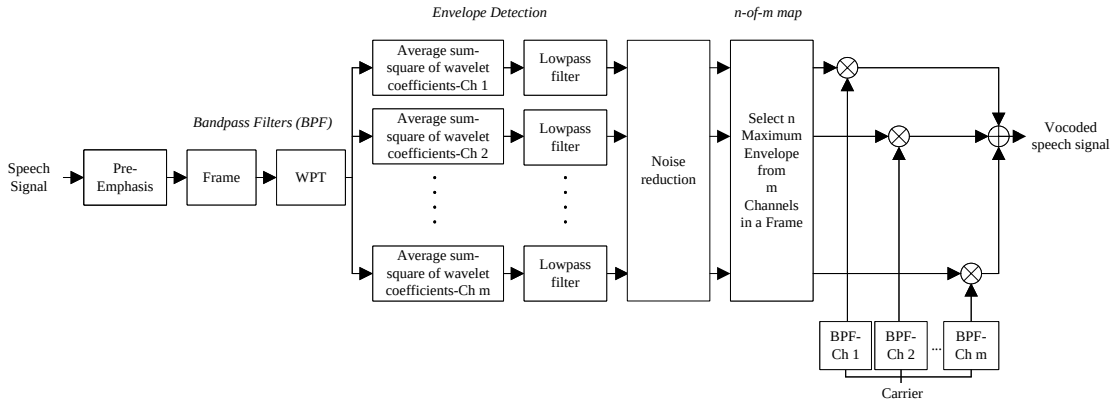


Figure 4.2 Block diagram of vocoder simulation for noise reduction in a wavelet packet-based speech coding strategy.

4.3 Noise reduction algorithms

4.3.1 Wavelet packet energy

Assume that noisy speech $y(n)$ is composed of clean speech $x(n)$ and the additive noise $d(n)$. Then:

$$y(n) = x(n) + d(n) \tag{4.1}$$

Taking the WPT of both sides gives:

$$Y_{j,n}(k) = X_{j,n}(k) + D_{j,n}(k) \tag{4.2}$$

where $Y_{j,n}(k)$, $X_{j,n}(k)$ and $D_{j,n}(k)$ are wavelet coefficients of the n^{th} subband (node) at level j for noisy speech, clean speech and noise, respectively. k is the coefficient index in each subband.

The noisy signal is divided into frames of length M =128 samples with 96 overlapping samples (75%). Each frame is calculated using WPT. The number of wavelet coefficients in each subband depending on the decomposition level *j*, is $K_j = 128/2^j$. In a single frame, the energy of each subband can be calculated using the average sum-square of the wavelet coefficients, thus:

$$E_Y(i,n) = \frac{1}{K_j}\sum_k \left|Y_{j,n}(k)\right|^2 \tag{4.3}$$

where $E_Y(i,n)$ is the energy of the i^{th} frame and the n^{th} subband, $Y_{j,n}(k)$ is the wavelet coefficient of the noisy signal in the n^{th} subband and level j, and k is the coefficient index ($k = 0,2,\ldots,K_j - 1$). The energy of the clean speech and the noise signal can be computed as noisy speech in the wavelet domain.

In the 23-band WPT, the energy of the first subband is discarded to provide 22 channels because it plays no role in speech perception. In the 64-band WPT, it is computed by summing the energy of consecutive subbands with frequency ranges, as in Table 3.1, to generate 22 channels. Then, the envelope amplitudes in each channel are smoothed using a lowpass filter as shown in Figure 4.2. This stage provides the time-frequency (T-F) envelope amplitude matrix, which represents the number of frames and channels. From Equation (4.2), the T-F envelope amplitude matrix at the i^{th} frame and n^{th} channel (subband) can be defined as:

$$Y(i,n) = X(i,n) + D(i,n) \tag{4.4}$$

where $Y(i,n)$, $X(i,n)$ and $D(i,n)$ are the T-F envelope amplitudes matrix for the noisy speech, clean speech and noise, respectively and $n = 0,1,2,\ldots,N-1$ channels ($N = 22$). The noise reduction algorithms are processed in the T-F envelope amplitude matrix. The differences and similarities between noise reduction techniques will be described and discussed in the next sections.

4.3.2 Ideal binary mask (IdBM)

Channel selection using the maximum amplitude criterion can become problematic for noisy environments (Dorman et al., 1997; Hu and Loizou, 2008; Kokkinakis et al., 2011). When noise dominates, the channels selected can be noise, because those channels have the maximum amplitudes. The ideal binary mask (IdBM) is employed to compensate for this shortcoming. In fact, the IdBM has been introduced as a goal of computational auditory scene analysis (CASA), which attempts to computationally extract sound mixtures into individual streams corresponding to different sound sources (Wang, 2005). The IdBM is applied to the criterion for selecting envelope channels, which is based on the true signal-to-noise ratio (*SNR*) to improve speech intelligibility in noisy environments (Hu and Loizou, 2008; Kokkinakis et al., 2011).

The IdBM is defined as a binary T-F mask, which is equivalent to a binary gain function. This approach is called ideal because its construction requires prior knowledge of the clean speech and noise information before both are mixed. The binary gain function takes the value of 1 when the *SNR* in the corresponding T-F envelope amplitude matrix exceeds a threshold value, and the value of 0 otherwise (Wang, 2005; Hu and Loizou, 2008). The T-F envelope matrix of enhanced speech is obtained as follows:

$$\hat{X}(i,n) = IdBM(i,n) \cdot Y(i,n) \tag{4.5}$$

$$IdBM(i,n) = \begin{cases} 1 & , SNR(i,n) \geq 0 \\ 0 & , SNR(i,n) < 0 \end{cases} \tag{4.6}$$

$$SNR(i,n) = 10\log_{10}(X^2(i,n) / D^2(i,n)) \tag{4.7}$$

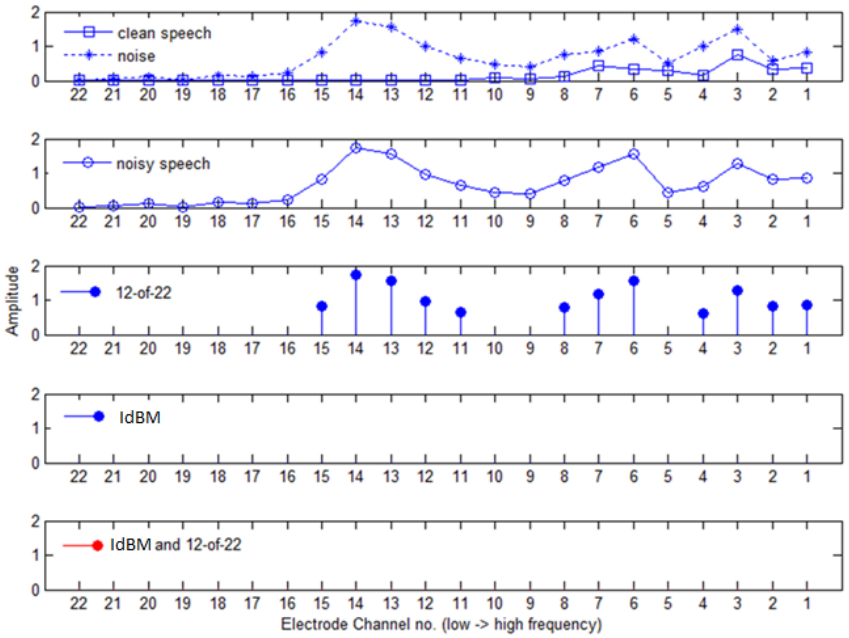
where $\hat{X}(i,n)$, $Y(i,n)$, $X(i,n)$ and $D(i,n)$ are the T-F envelope matrices at the i^{th} frame and n^{th} channel for the enhanced speech, noisy speech, clean speech and noise signal, respectively.

The threshold of *SNR* was 0 dB for this study. The threshold value of 0 dB has been found to work well and produce optimality in studies employing the IdBM (Wang, 2005). This threshold was reasonable because the purpose of IdBM-based channel selection was to retain the speech-dominated channels and to remove the noise-

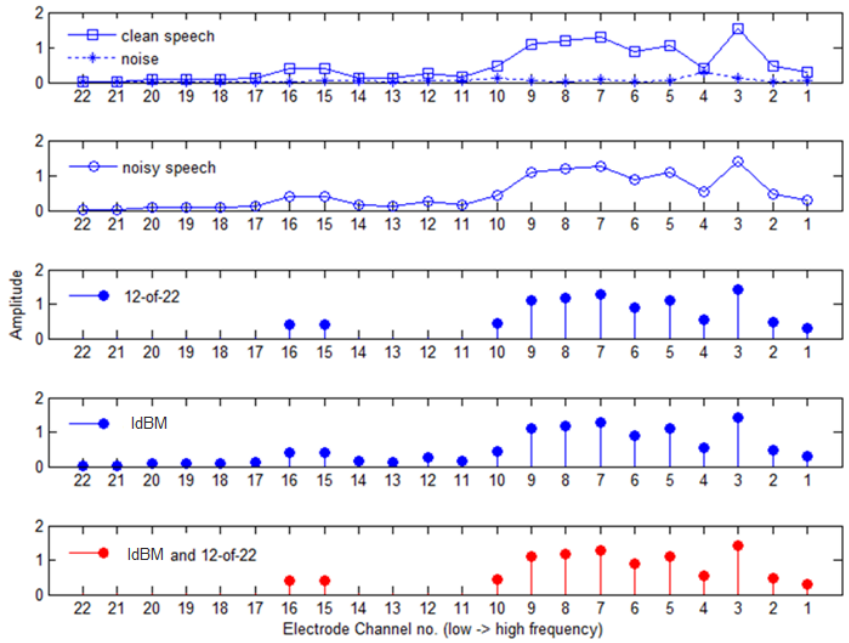
dominated channels (Hu and Loizou, 2008), as shown in Figure 4.3. The speech-dominated channels (i.e. $SNR \geq 0$) contain important information about clean speech, whereas the noise-dominated channels (i.e. $SNR < 0$) contain little information about clean speech because speech signals are severely masked by noise and the speech components of the mixture are almost inaudible.

The number of channels selected corresponding to *SNR* can vary from 0 (i.e. no channels are selected) to 22 (i.e. all channels are selected). For noise-dominated channels as shown in Figure 4.3 (a), the IdBM will not select any channels, while the *n*-of-*m* strategy will select 12 channels with the largest amplitudes. For speech-dominated channels as shown in Figure 4.3 (b), the IdBM will select all channels while the *n*-of-*m* strategy will only select the 12 channels with the largest amplitudes. This can be a disadvantage of the IdBM strategy when the speech-dominated channels number more than 12. This is unnecessary for speech intelligibility (Dorman et al., 2002), especially in quiet or high SNR conditions. In this study, a combination of IdBM-based channel selection and the *n*-of-*m* strategy are used, as shown in Figure 4.3.

From Equation (4.5), the T-F envelope amplitudes of noisy speech with $SNR \geq 0$ dB are retained while the envelope amplitudes with $SNR < 0$ dB are removed to reduce noise in the CI processors. Figure 4.4 illustrates the noise reduction with IdBM. The clean speech is shown in Figure 4.4 (a). The babble noise at 5 dB SNR and the noisy speech is shown in Figure 4.4 (b) and (c), respectively. The IdBM is shown in Figure 4.4 (d). The result of IdBM provides enhanced speech, which is shown in Figure 4.4 (e). The enhanced speech is much closer to the clean speech. Informal listening to the enhanced speech results in clear intelligibility, like the clean speech.

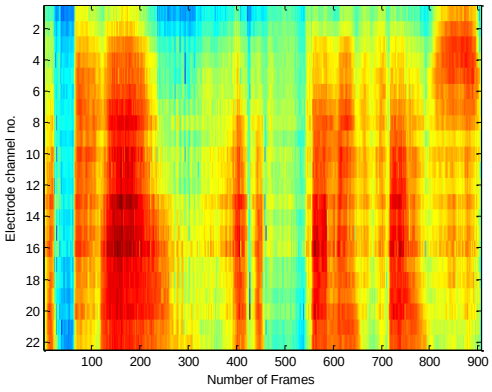


(a)

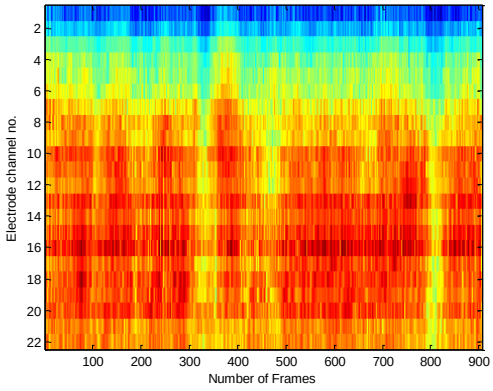


(b)

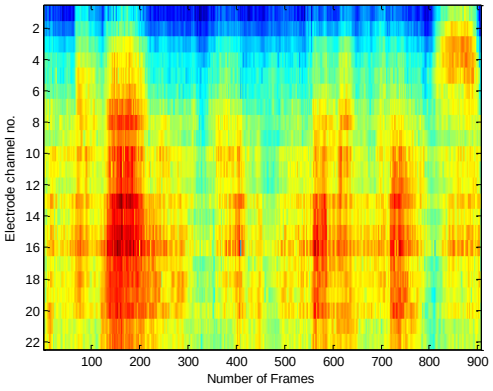
Figure 4.3 Example illustrating the channel selection in a frame using the *n*-of-*m* strategy, IdBM, and a combination of IdBM and the *n*-of-*m* strategy. The first panel shows the amplitudes of the clean speech and noise signal. The second panel shows the amplitudes of the noisy speech. The bottom three panels show the amplitudes selected by the *n*-of-*m* strategy, IdBM, and the combination of IdBM and the *n*-of-*m* strategy, respectively. (a) The noise dominates the clean speech. (b) The clean speech dominates the noise.



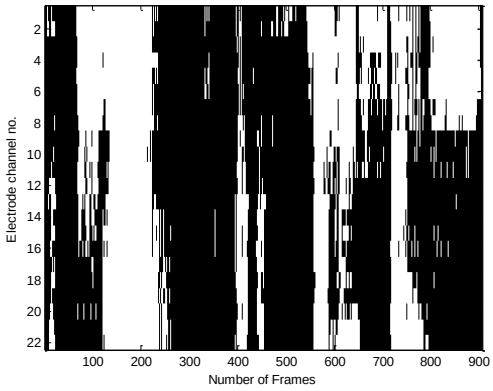
(a)



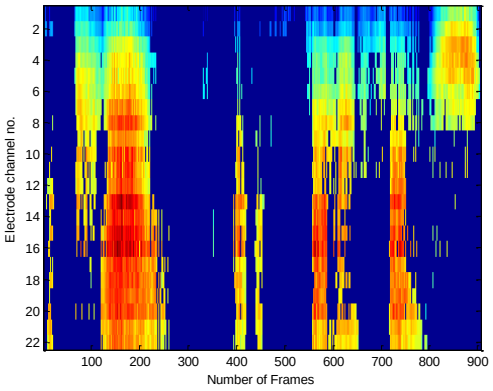
(b)



(c)



(d)



(e)

Figure 4.4 Example illustrating the concept of IdBM for the BKB sentence "*The clown had a funny face*". (a) Clean speech. (b) Babble noise at 5 dB SNR. (c) Noisy speech. (d) The IdBM, where white pixels indicate 1 and black pixels 0. (e) Enhanced speech using IdBM.

4.3.3 Time-frequency spectral subtraction (TFSS)

The spectral subtraction proposed by Boll (1979) is one of the earliest and most well-known techniques for speech enhancement. This technique is based on a simple implementation where enhanced speech is obtained by subtracting the noise estimation from the noisy speech. Numerous studies have proposed different implementations and configurations of spectral subtraction to find the optimized spectral subtraction for their applications. The objective here is to apply time-frequency spectral subtraction (TFSS) in wavelet packet-based speech coding strategies.

4.3.3.1 Power spectral subtraction and error analysis

From Equation (4.4), the estimated power spectrum of the enhanced speech in the wavelet domain using power spectral subtraction (Berouti et al., 1979; Virag, 1999) can be expressed as follows:

$$\hat{X}^2(i,n) = \max\left(Y^2(i,n) - \hat{D}^2(i,n), 0\right) \tag{4.8}$$

where $\hat{X}^2(i,n)$ and $\hat{D}^2(i,n)$ represent the estimated power spectrum of the enhanced speech and noise, respectively. $Y^2(i,n)$ is the power spectrum of the noisy speech. The $\max(\cdot)$ operator is used to guarantee that $\hat{X}^2(i,n)$ always has a positive value.

Generally, the noise level $D^2(i,n)$ is unknown, but it can be estimated from nonspeech frames. The estimated power spectrum of the enhanced speech $\hat{X}^2(i,n)$ may be negative values as a result of spectral subtraction because the noise estimation $\hat{D}^2(i,n)$ may be inaccurate due to the random variation of the noise spectrum. These negative values are set to zero. This process produces tones at random times and frequencies which result in an artefact called musical noise (Berouti et al., 1979).

Musical noise can be reduced by improving the estimation of noise. Several techniques have been proposed to reduce this effect: magnitude averaging (Boll, 1979), over-subtracting the estimation of the noise spectrum and spectral-flooring the estimation of the negative values (Berouti et al., 1979), a minimum mean-square error (MMSE) estimation of short-time spectral amplitude (Ephraim and Malah, 1984), an adaptation of subtraction parameters related to the masking properties of human

perception (Virag, 1999), and the estimation of cross terms associated with the phase differences between the noisy/clean speech and noise (Lu and Loizou, 2008).

The power spectral subtraction in Equation (4.8) can be written in terms of the gain function $G(i,n)$ as:

$$\hat{X}^2(i,n) = G^2(i,n) \cdot Y^2(i,n) \tag{4.9}$$

$$G(i,n) = \frac{\hat{X}(i,n)}{Y(i,n)} = \sqrt{1 - \frac{\hat{D}^2(i,n)}{Y^2(i,n)}} = \sqrt{1 - \frac{1}{\gamma(i,n)}} \tag{4.10}$$

where $G(i,n)$ always takes positive values in the range of $0 \le G(i,n) \le 1$, and γ is a *posteriori* SNR ($\gamma(i,n) \triangleq Y^2(i,n) / \hat{D}^2(i,n)$). The gain function $G(i,n)$ is used to modify the amplitude of the noisy speech between the speech and noise regions. Regions containing only speech signals are unmodified (i.e. $G(i,n) = 1$), whereas regions containing only noise are removed (i.e. $G(i,n) = 0$). Regions containing both speech signals and noise are modified to reduce the noise according to the *posteriori* SNR γ.

Substituting (4.4) into (4.8), the power spectral subtraction can be rewritten as:

$$\hat{X}^2(i,n) = X^2(i,n) + D^2(i,n) - \hat{D}^2(i,n) + 2\,\mathrm{Re}(X(i,n)D^*(i,n))$$

$$\hat{X}^2(i,n) = Y^2(i,n) - \hat{D}^2(i,n) + 2\,\mathrm{Re}(X(i,n)D^*(i,n)) \tag{4.11}$$

The power spectrum estimate of the enhanced speech $\hat{X}^2(i,n)$ includes error terms of noise variation (i.e. $D^2(i,n) - \hat{D}^2(i,n)$) and cross terms of clean speech and noise (i.e. $2\,\mathrm{Re}(X(i,n)D^*(i,n))$) (Shao and Chang, 2007). The cross terms are commonly set to zero because the clean speech and noise are assumed to be uncorrelated. This assumption leads to an inaccurate subtraction rule (Lu and Loizou, 2008). Some researchers have attempted to assess the effect of neglecting the cross term (Evans et al., 2006) and compensate for the cross term in spectral subtraction (Lu and Loizou, 2008).

4.3.3.2 Cross term to perceptual time-frequency spectral subtraction

The cross term estimate proposed by Lu and Loizou (2008) is applied in this study. This cross term can be represented by a geometric perspective on spectral subtraction, which

provides the difference between the phases of noisy/clean speech and noise. The gain function can be created using the relationship between the difference of phases and trigonometric principles (Appendix E.1). This gain function, dependent on the estimation of *priori* SNR $\hat{\xi}$ and *posteriori* SNR $\hat{\gamma}$ parameters, can be expressed as follows:

$$G\left(\hat{\xi},\hat{\gamma}\right)=\sqrt{\left(1-\frac{(\hat{\gamma}+1-\hat{\xi})^2}{4\hat{\gamma}}\right)\Big/\left(1-\frac{(\hat{\gamma}-1-\hat{\xi})^2}{4\hat{\xi}}\right)} \tag{4.12}$$

The parameters $\hat{\xi}$ and $\hat{\gamma}$ in the gain function $G\left(\hat{\xi},\hat{\gamma}\right)$ are estimated according to:

$$\hat{\gamma}(i,n)=\beta\cdot\hat{\gamma}(i-1,n)+(1-\beta)\cdot\min\left(\hat{\gamma}_I(i,n),20\right) \tag{4.13}$$

$$\hat{\xi}(i,n)=\alpha\cdot\hat{\xi}_I(i-1,n)+(1-\alpha)\cdot\left(\sqrt{\hat{\gamma}(i,n)}-1\right)^2 \tag{4.14}$$

$$\hat{\xi}_I(i,n)\triangleq\frac{\hat{X}^2(i,n)}{\hat{D}^2(i,n)}\text{ and }\hat{\gamma}_I(i,n)\triangleq\frac{Y^2(i,n)}{\hat{D}^2(i,n)} \tag{4.15}$$

where the subscript I indicates the instantaneous values. β and α are weighting factors, which were set to $\beta=0.60$ and $\alpha=0.98$. Both factors control the trade-off between the noise reduction and the speech distortion. Both values were selected based on informal listening tests and predicting the objective speech intelligibility (i.e. NCM and STOI) between the vocoded clean speech and the vocoded noisy speech with TFSS. The $\min(\cdot)$ operator was used to give a maximum of 13 dB ($=10\log_{10}(20)$) and to avoid over-attenuation of the signals (Lu and Loizou, 2008).

This gain function is employed in the time-frequency spectral subtraction (TFSS) according to the following steps. Initially, an estimate of the noise power spectrum $\hat{D}^2(i,n)$ is averaged from the first five frames. Then $\hat{D}^2(i,n)$ is updated by a noise estimation algorithm (Martin, 2001), which is obtained by the minimum tracking method, since the power spectrum of the noisy speech regularly decays to the noise power level. This method tracks minimum values of a smoothed power spectrum for the noisy speech and multiplies by a constant to compensate for the bias noise estimate. This method has been found to work well for nonstationary environments.

Finally, the T-F envelope amplitude matrix of the enhanced speech is computed by a multiplication of the gain function $G\left(\hat{\xi},\hat{\gamma}\right)$ with the T-F envelope amplitude matrix of the noisy envelopes:

$$\hat{X}(i,n)=G\left(\hat{\xi},\hat{\gamma}\right)\cdot Y(i,n) \tag{4.16}$$

The *posteriori* SNR $\hat{\gamma}(i,n)$ in Equation (4.13) is weighted to reduce rapid fluctuations and also to limit the over-suppression of the signal for large values of $\hat{\gamma}(i,n)$. The weighting factor β of $\hat{\gamma}(i,n)$ can improve the estimate of the enhanced speech. The *priori* SNR $\hat{\xi}(i,n)$ in Equation (4.14) is weighted to control the average of spectral information positioned on past and present frames. This is similar to the decision-directed approach (Ephraim and Malah, 1984), which updates amplitude estimates using information from the past frames.

This gain function has two main advantages (Lu and Loizou, 2008). First, it is not derived using any statistical model about the statistical distributions of the speech and noise (e.g. Gaussian, Gamma, Laplacian or Raleigh distributions). In addition, the best statistical model is currently undetermined (Ephraim and Cohen, 2004). Second, the estimation of the parameters $\hat{\xi}$ and $\hat{\gamma}$ are instantaneous values, which are updated directly from estimates of the instantaneous noise as in Equation (4.15). This doesn't only use one average of noise estimate from the initial noise segment of the signal, as in Ephraim and Malah (1984). As a result, this gain function provides a more accurate estimate of enhanced speech and is well appropriate for real-time implementations in noisy environments.

4.3.4 Time-adaptive wavelet thresholding (TAWT)

4.3.4.1 Conventional wavelet thresholding

Donoho and Johnstone (1994) have proposed wavelet thresholding for noise reduction. This algorithm consists of three steps: forward wavelet transform of the noisy signals, thresholding the wavelet coefficients and inverse wavelet transform. Wavelet thresholding utilises statistical differences between the wavelet coefficients of speech

and noise signals. Nonsignificant coefficients have small absolute values, they are probably noise and they should be removed or attenuated. Significant coefficients have large absolute values, they are important components of signals and they should be retained. Therefore, the wavelet coefficients below a selected threshold are treated as nonsignificant information and set to zero, whereas the significant ones are kept.

The soft-thresholding gain function (Donoho and Johnstone, 1994; Donoho, 1995) was introduced in Equation (4.17) and its characteristic of signals is shown in Figure 4.5. The soft thresholding gain function T_S sets the absolute values of wavelet coefficients below the selected threshold λ to zero. The absolute values of wavelet coefficients above the selected threshold λ are replaced by shrinking the wavelet coefficients of the noisy speech $Y(i,n)$ with the selected threshold λ.

$$\hat{X}(i,n) = T_S(Y,\lambda) = \mathrm{sgn}(Y(i,n)) \max(|Y(i,n)| - \lambda, 0) \tag{4.17}$$

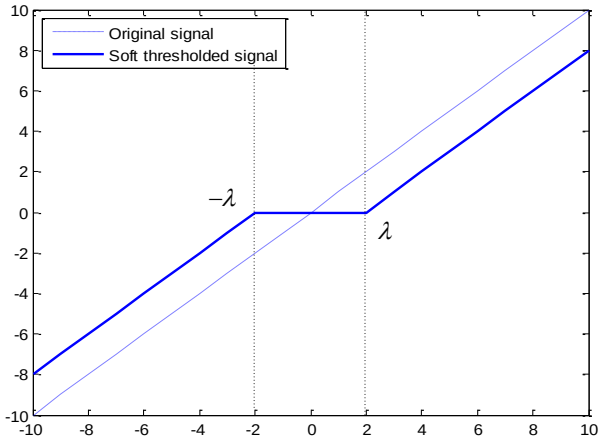


Figure 4.5 The soft-thresholding gain function.

However, applying a selected threshold λ to all wavelet coefficients can lead to over-thresholding of speech regions. This not only reduces additional noise but also removes some speech components such as unvoiced sounds. In order to solve the problem of the limit of a selected threshold in conventional wavelet thresholding, an adaptive threshold (Bahoura and Rouat, 2001; Chen and Wang, 2004) without any prior

knowledge of the noise level is applied in this study. The obtained results give better performance than those using an MMSE gain function (Ephraim and Malah, 1984).

4.3.4.2 Time-adaptive wavelet thresholding

The time-adaptive wavelet thresholding (TAWT) algorithm (Chen and Wang, 2004) is different from conventional wavelet thresholding (Donoho and Johnstone, 1994). This technique is based on the Teager energy operator (TEO) and the adaptation of the wavelet threshold.

The TEO was modelled by Teager (Teager and Teager, 1990) and was further investigated by Kaiser (Kaiser, 1993). The TEO is a simple nonlinear function and a very local property of the signal, dependent on the three adjacent samples of the signal with indexes $i-1$, i, and $i+1$. It is used to enhance the discriminability of speech and noise (Bahoura and Rouat, 2001; Chen and Wang, 2004). The TEO is a powerful tool that has been used in many speech applications (Bahoura and Rouat, 2001; Chen and Wang, 2004; Bahoura and Rouat, 2006; Dimitriadis et al., 2011).

The TAWT algorithm is computed in the following steps. The TEO coefficients $T(i,n)$ can be calculated from samples of three adjacent amplitude envelopes as:

$$T(i,n) = Y^2(i,n) - Y(i+1,n)Y(i-1,n) \tag{4.18}$$

where $Y(i,n)$ is the T-F envelope amplitude matrix of the noisy speech at the i^{th} frame and the n^{th} channel. The temporal masking $M(i,n)$ is constructed by smoothing the TEO coefficients, defined by:

$$M(i,n) = T(i,n) * h(i,n) \tag{4.19}$$

where * denotes the convolution operation and $h(i,n)$ is the lowpass filter.

The adaptive threshold values $\lambda(i,n)$ are constructed from the temporal masking $M(i,n)$. If $M(i,n)$ below the variance of $M(i,n)$ is set to zero, otherwise temporal masking $M(i,n)$ is normalised as follows:

$$M'(i,n) = \begin{cases} \left[\dfrac{M(i,n)}{\max(M(i,n))} \right], & M(i,n) > \mathrm{var}(M(i,n)) \\ 0 \quad , & otherwise \end{cases} \tag{4.20}$$

The parameter of $M'(i,n)$ is close to 1 for speech regions and close to 0 for noise regions. Therefore the adaptive threshold values $\lambda(i,n)$ can be expressed as:

$$\lambda(i,n) = \lambda_n(1 - M'(i,n)) \tag{4.21}$$

$$\lambda_n = \sigma_n \sqrt{2\log(N\log_2(N))} \quad \text{and} \quad \sigma_n = MAD_n / 0.6745 \tag{4.22}$$

where λ_n represents the channel-dependent threshold values (Bahoura and Rouat, 2001), N is the total frames, σ_n is the noise variances with the median of the absolute deviation (MAD_n) of all the wavelet coefficients $Y(i,n)$ at the n^{th} channel, and 0.6745 is a normalisation factor, which is approximated from fine-scale wavelet coefficients (Donoho, 1995). The enhanced speech $\hat{X}(i,n)$ is modified by the soft thresholding gain function as:

$$\hat{X}(i,n) = \mathrm{sgn}(Y(i,n)) \max(|Y(i,n) - \lambda(i,n)|, 0) \tag{4.23}$$

4.4 Objective speech intelligibility

In the previous chapter, a frame length of 128 samples (the default in the ACE strategy) and a sym8 (Symlet with order 8) were chosen for the wavelet packet-based speech coding strategy. The sym8 yielded good results based upon the information envelope and electrodogram when compared to other wavelet filters. For noise reduction techniques, various algorithmic parameters were chosen for the TFSS, but not for the TAWT. Suitable parameters β and α for the TFSS were chosen based on informal listening tests and the average values of predictions from the normalised covariance metric (NCM) and short-time objective intelligibility (STOI) in all conditions.

Vocoded speech, with and without noise reduction algorithms in situations of different of noise types and SNR levels, were evaluated using the NCM and STOI to

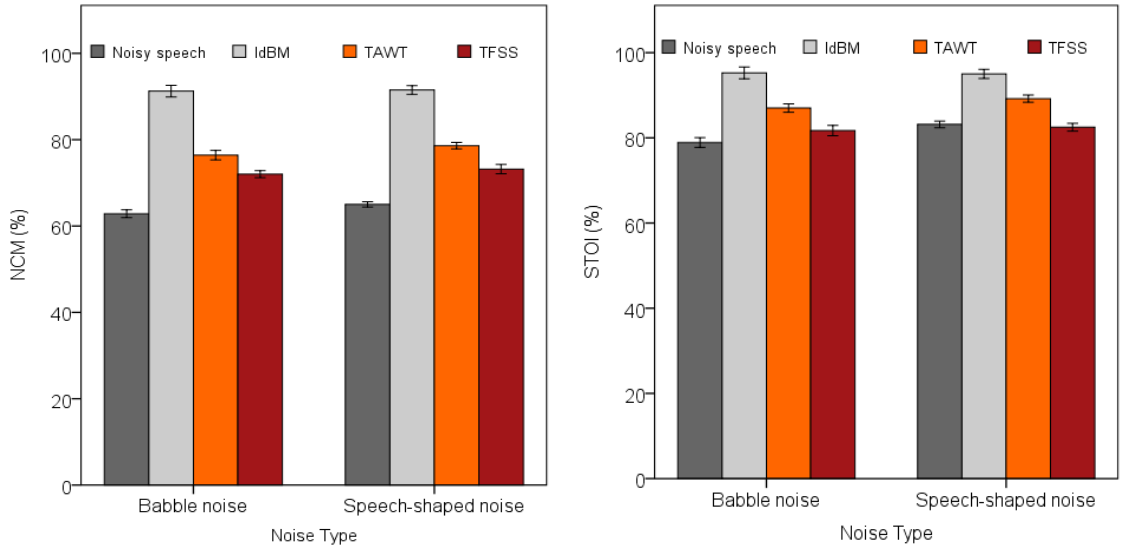
predict the direction of performance, before a listening test with NH listeners. Both the NCM and STOI were computed using vocoded clean speech (as a reference speech) and vocoded noisy speech with and without noise reduction algorithms.

The NCM and STOI values for each condition were obtained from the average of 336 Bamford-Kawal-Bench (BKB) sentences (details are provided in Section 5.2.1.1) per condition. In the conditions of different noise types, the sentences were corrupted by two types of noise, i.e. babble and speech-shaped noise at 5 dB SNR. There were a total of 16 conditions (4 algorithms × 2 noise × 2 wavelet packet structures). In conditions with different SNR levels, the sentences were corrupted by babble noise at 0, 5 and 10 dB SNR. There were a total of 18 conditions (3 algorithms × 3 SNR levels × 2 wavelet packet structures).

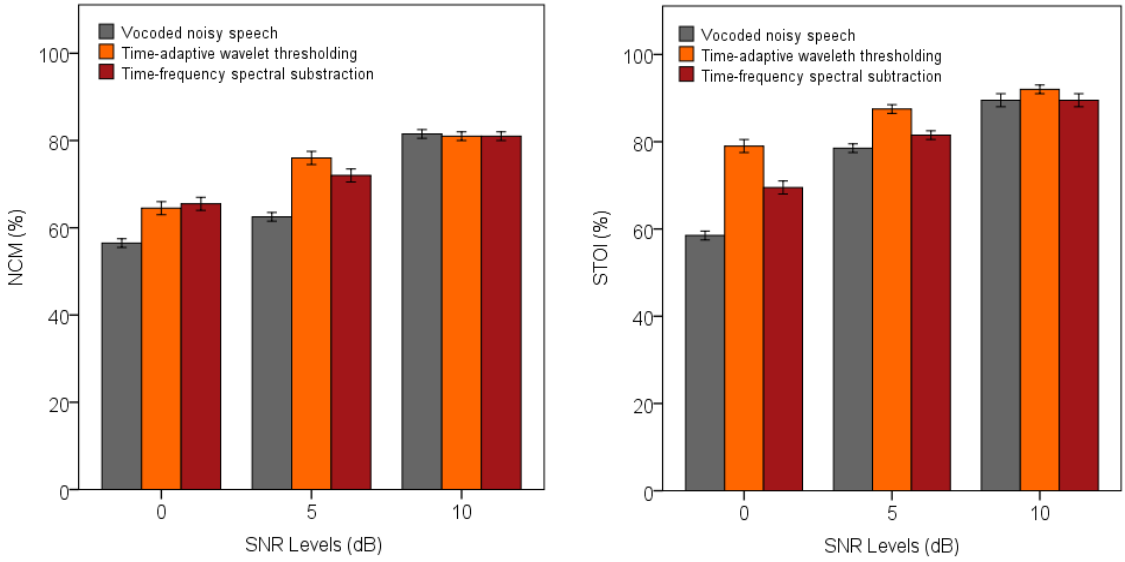
Figure 4.6 shows the comparative results of the NCM (left) and STOI (right) for processing with and without noise reduction algorithms, in terms of different noise types (Figure 4.6 (a)) at 5 dB SNR and different SNR levels (Figure 4.6 (b)) in babble noise. The results of the NCM and STOI have the same trend of performance in almost all conditions except at 0 dB SNR. The STOI provides considerably higher values of speech intelligibility than the NCM in all conditions.

A one-way analysis of variance (ANOVA) revealed a statistical significant (F [15, 31] =532.78, p<0.0005) in different noise types (Figure 4.6 (a)) and a statistical significant (F [17, 35] =77.44, p<0.0005) in different SNR levels (Figure 4.6 (b)) for processing on the intelligibility measures examined. Post-hoc tests (Bonferroni) were used to assess differences between values of the intelligibility measures obtained in the different conditions.

As can be seen from Figure 4.6 (a), the IdBM provided significantly better performance than others. The TAWT yielded significantly better performance than the TFSS in both noises at 5 dB SNR. In Figure 4.6 (b), the TAWT and TFSS provided significantly better performance than vocoded noisy speech at 0 and 5 dB SNR, but not at 10 dB SNR for babble noise. Both the TAWT and TFSS showed no significant difference in almost all SNR levels. The TAWT provided only significantly better performance than TFSS for the STOI at 0 and 5 dB SNR.



(a)



(b)

Figure 4.6 The comparison of performance between noise reduction algorithms (i.e. TFSS and TAWT) in terms of intelligibility measures (i.e. NCM and STOI). Each column denotes intelligibility measures in NCM and STOI. The rows represent different noise types (a) and different SNR levels (b). The error bars indicate ± 1 standard error.

4.5 Discussion

4.5.1 Differences between noise reduction algorithms

The noise reduction algorithms applied in the wavelet packet-based speech coding strategy (i.e. TFSS and TAWT) have been discussed and compared to IdBM. All the algorithms are relatively simple in their implementation, do not require the explicit voice activity detection, and have low computing complexity. This makes them more suitable for real-time implementation in CI systems.

In noise estimation, all algorithms are simply implemented by gain functions in noisy envelopes in each channel. These algorithms employ different techniques in their gain functions. The IdBM requires prior knowledge of speech and noise information in the present frame to find the exact *priori* SNR of the signal. The TFSS requires the noise estimation algorithm proposed by Martin (2001) and information from previous and present frames to estimate the *priori* SNR and the *posteriori* SNR of the gain function. This approach needs to adjust some parameters (e.g. weighting factors) to get the optimal performance for different types of noise and different SNR levels.

The TAWT uses the TEO to construct temporal masking in each channel. This temporal masking is applied to estimate adaptive threshold values for the soft-thresholding gain function. The TAWT does not require both an explicit estimation of noise level or any knowledge of the *priori* SNR and the *posteriori* SNR, using only a noise variance in each channel. In addition, the TAWT can reduce the data redundancy by setting the envelope amplitudes below the threshold λ to zero.

The algorithms may provide more error estimation for nonstationary noise or lower SNR levels than stationary noise or higher SNR levels. The algorithms may work better for speech-shaped noise than for babble noise. This is because the characteristics of babble noise vary rapidly. Then the noisy envelopes may be weighted by the gain function with inaccurate noise estimation, which leads to decreased intelligibility of performance (Fu and Nogaki, 2005). The noise reduction algorithms should be further developed for nonstationary noise. Figure 4.6 (a) shows that the TFSS and TAWT in speech-shaped noise give a slightly higher performance than those in babble noise when predicting performance with both NCM and STOI.

At lower SNR levels, speech signals are heavily masked by noise and noisy envelopes are multiplied by a weight close to zero. Little speech information in each channel is probably not intelligible and distracting information to CI users (Hu et al., 2007). In contrast, at sufficiently high SNR levels, noisy envelopes are multiplied by a weight close to one. CI users can understand most of the important information and ignore the noise. Figure 4.6 (b) shows that TFSS and TAWT at higher SNR levels (e.g. 10 dB SNR) are slightly better or almost the same as noisy speech. This is because noisy speech with and without noise reduction may allow discrimination between speech and noise, with the capabilities of the human auditory system (Verschuur et al., 2006).

Ideally, the noise reduction algorithms in CI systems should be able to automatically detect noise environment changes and select optimized parameters for the noise reduction algorithms. Currently, there are no noise estimation techniques that can track noise spectra accurately. However, Hu et al. (2007) suggested that noise estimation does not have to be very accurate to obtain an exact weight for multiplying to the noisy envelopes. It is enough if the noise estimation performs sufficiently well to discriminate high from low SNR envelopes.

A few studies in CI noise reduction suggested that in practice, noise reduction for CI users should provide more aggressive gain functions and show significant intelligibility performance improvement (Hu et al., 2007; Dawson et al., 2011). In contrast, noise reduction for NH listeners is designed to be less aggressive in maintaining listening quality. This is because of the perceptible difference between NH listeners and CI users.

In noise reduction strategies for CIs, rather than TFSS as the envelope-based strategy, various techniques for spectral subtraction are applied as pre-processing strategies and they are able to improve intelligibility performance for CI users (Yang and Fu, 2005; Verschuur et al., 2006). Some studies reported that algorithms of spectral subtraction carry low computational complexity (Verschuur et al., 2006). However, none of the spectral subtraction algorithms are applied as envelope-based strategies, which are expected to provide the same performance improvement as pre-processing strategies but require a lower computational load. Figure 4.6 indicates that the trend of

TFSS as the envelope-based strategy may improve performance for CI users both in different noise types and at different SNR levels.

Other techniques based on envelope-weighting with a gain function include sigmoidal-shaped gain function (Hu et al., 2007), PCA/ICA with soft thresholding (Li, 2008), sigmoidal-shaped gain function with a *posteriori* SNR estimate (Dawson et al., 2011), modified Wiener gain function (Dawson et al., 2011), and sparse non-negative factorisation (Hu et al., 2013). These algorithms are reported to be able to improve intelligibility performance for NH listeners and CI users.

Interestingly, TAWT is similar to PCA/ICA based on soft thresholding, which is a stage in envelope-based algorithms. For TAWT, the stage of soft thresholding is applied directly in the wavelet domain with sparseness properties, while the soft thresholding of PCA/ICA is applied in the ICA domain. Therefore, TAWT requires lower computational complexity than PCA/ICA. Furthermore, Figure 4.6 shows that TAWT trends to improve intelligibility for CI users.

4.5.2 Validity of objective intelligibility measures

The NCM and STOI are used to pre-evaluate the intelligibility performance of noise reduction algorithms. Both NCM and STOI were found that they work well with NH listeners for non-vocoded noisy speech with noise reduction algorithms (Jianfen et al., 2009; Sang, 2012). The NCM was confirmed to be good measure with NH listeners for vocoded noisy speech without noise reduction algorithms (Chen and Loizou, 2011).

When TAWT and TFSS were evaluated by using NCM and STOI, TAWT had better trend for intelligibility performance than TFSS both in different noise types and at different SNR levels. The trend of intelligibility performance of the obtained results was expected to be consistent with previous research, which examined CI systems with similar noise reduction methods evaluated by NH listeners and CI users. Nevertheless, the use of various objective measures might increase reliability for predicting the performance of noise reduction algorithms.

Finally, subjective intelligibility measures with listening tests are required to obtain reliability of intelligibility performance. Noise reduction algorithms were

assessed by subjective intelligibility measures with NH listeners in Chapter 5 and were compared with objective intelligibility measures to justify the correlation and reliability of objective intelligibility measures.

4.6 Conclusions

Noise reduction techniques based on wavelet packet transform, namely TAWT and TFSS, were implemented in the wavelet packet-based speech coding strategy. All the noise reduction techniques removed noise and retained important speech information. They were expected to benefit CI users in terms of speech intelligibility.

Vocoded noisy speech with and without noise reduction techniques were pre-evaluated by objective intelligibility measures including NCM and STOI. The comparative results indicated that IdBM might bring more benefit while the TAWT and TFSS might bring either less benefit or no benefit in terms of speech intelligibility. The TAWT provides the better trend of intelligibility performance than the TFSS.

Although the IdBM provides the best possible performance, it is impossible for application in the real world because its approach uses an ideal method in which information about speech and noise are known and noise estimation is accurate. The TFSS needs to adjust some parameters for the best performance in each condition. Consequently, the TAWT is the more suitable and realistic method for noise reduction than the TFSS in the real-world situation.

Chapter 5: Evaluation of wavelet packet-based strategies for normal-hearing listeners

5.1 Introduction

CI manufacturers provide several speech coding strategies in their CI systems (Loizou, 2006) – for instance, the Cochlear Corporation supports the ACE and CIS strategies in their Nucleus device (Cochlear, 2002). CI users benefit from increasing the number of speech coding strategies, as at least one of them might be more useful than the others. This also allows a large number of parameters to be configured in CI processors and the complexity of selecting the optimal subset of parameters associated with each strategy.

The parameters of CI processors specified in a CI user's MAP (Appendix D.2) can be varied to optimise speech recognition performance for individual CI users, such as channel stimulation rate, filter spacing and the number of channels selected (Dorman et al., 1997; Loizou et al., 2000; Fourakis et al., 2007; Kasturi and Loizou, 2007). It is an important issue to identify the optimal subset of parameters for fitting CI users. The optimal subset of parameters is a good starting point and saves time in selecting parameters during the fitting of new CI users (Loizou et al., 2000).

Designing the wavelet packet-based speech coding strategy with noise reduction algorithms, the parametric variation of the wavelet packet filter bank and the noise reduction algorithms may affect speech recognition performance. The parametric variation of the wavelet packet filter bank includes the wavelet packet structures, the types of mother wavelet and the frame lengths. These parameters were quite difficult to evaluate using objective intelligibility measures because their objective intelligibility values are almost the same. Thus the listening test is more suitable for evaluating these parameters. The parametric variation of noise reduction algorithms was evaluated using the objective intelligibility measures to guide the adjustment of parameters to improve speech intelligibility before the listening test with NH listeners.

In this chapter, the evaluation of the wavelet packet-based speech coding strategies for NH listeners can be divided into two parts: the effects of parametric

variation in wavelet packet filter banks on speech intelligibility (i.e. filter spacing, type of mother wavelet and frame lengths) and the comparison of noise reduction algorithms (i.e. TFSS and TAWT). The first part aims to further explore the optimal parameters of wavelet packet filter banks that may affect speech recognition in both quiet and noisy conditions. The second part aims to investigate and compare vocoded noisy speech, with and without noise reduction algorithms for different noise types and SNR levels.

5.2 Effect of wavelet packet filter banks on speech intelligibility

Three speech processing parameters were examined to study the effect of parametric variation of wavelet packet filter banks on speech intelligibility. Experiment 1 examined the effect of filter spacing. The filter spacing design relates to the wavelet packet filter banks and the number of channels allocated in the formant regions. The wavelet packet-based strategy was also compared to the commercial ACE strategy. Experiment 2 examined the effect of a perceptually optimised wavelet. This mother wavelet was based on an auditory model (Karmakar et al., 2011) and was compared to Symlet, which was applied in CI processors (Nogueira et al., 2006; Gopalakrishna et al., 2010b). Experiment 3 examined the effect of frame lengths. Different frame lengths may provide different speech recognition performance.

5.2.1 Experiment 1: Effect of filter spacing

An important stage in all CI processors is the decomposition of speech signals into frequency bands. Therefore, the signal bandwidth and filter spacing need to be considered to find the optimal frequency-to-electrode allocation (Kasturi and Loizou, 2007).

The signal bandwidth is constrained by the Nyquist theorem to provide half of the sampling frequency. A sampling frequency of 16 kHz is commonly used in the CI processor, and the bandwidth is between 0 Hz and 8 kHz. Bandwidths ranging from 6.7 to 9.9 kHz have no significant effects on consonant and vowel recognition (Loizou et al., 2000). A bandwidth of 4 kHz is very important for understanding speech (Loizou,

1998). This bandwidth contains the first three formants denoted as F1 (0.3−1 kHz), F2 (1−3 kHz) and F3 (>3 kHz) (Hillenbrand et al., 1995; Loizou, 2006), which are the frequency bands for most vowels. However, a small bandwidth (i.e. 0−4 kHz) may result in consonant confusions (e.g. f/s, p/t and t/k), especially a female speakers. A wide bandwidth (i.e. 0−8 kHz) can reduce the consonant confusion (Loizou, 1998; Loizou et al., 2000).

The filter spacing in each channel of CI processors requires specific frequency ranges. The optimal spacing of frequency bands to the number of electrodes (12−22) is becoming more important to find the best mapping of frequency-to-electrode allocation, and it might also have an important effect on perception outcomes for CI users. Filter banks with narrow frequency spacing provide considerable flexibility for setting centre frequencies and bandwidth in each channel of CI processors. Consequently, more channels can be easily allocated in the low-frequency region (Kasturi and Loizou, 2007; Mourad Ghrissi, 2012). The channel density in the low-frequency region plays a critical role in speech recognition (Fourakis et al., 2004; Loizou, 2006; Fourakis et al., 2007) and melody recognition (Kasturi and Loizou, 2007).

There are many ways of allocating the filter spacing in signal bandwidth such as logarithmic, Mel and Bark scales. These frequency scales relate to the assignment of the number of channels in the formant regions, which may influence intelligibility, at least on vowel recognition tasks (Loizou, 2006). A study (Loizou, 2006) found that Clarion CI users obtained a significant benefit in vowel recognition using the Bark scale over Mel and logarithmic scales, because the Bark scale had the highest number of channels in the F1/F2 region. A similar outcome was reported in Fourakis et al. (2004) and Fourakis et al. (2007), which indicate that there was performance improvement with the assignment of more channels to the F1/F2 regions for Nucleus CI users. In addition, Kasturi and Loizou (2007) found that a small difference in the number of channels in the low-frequency region produced a difference of 34 percentage points in melody recognition for NH listeners and Clarion CI users.

The aim of this experiment was to determine whether the number of channels of different wavelet packet filter banks based on Bark scale could affect to speech recognition.

5.2.1.1 Method

A. Subjects

Nine NH listeners participated in this experiment. All subjects were native speakers of British English (6 males, 3 females, from 18 to 34 years of age) and all had normal hearing thresholds (< 20 dB HL). They were staff and students at the University of Southampton and were paid for their participation. Testing was approved by the University of Southampton Experimentation Safety and Ethics Committee.

B. Stimuli

The BKB (Bamford-Kowal-Bench) sentences (Bench et al., 1979) were used. They are composed of 21 lists with each list consisting of 16 sentences (21 lists × 16 sentences = 336 sentences) and 50 key words (3−4 words per sentence). The sentences are composed of no more than seven syllables and their vocabulary reflects the natural language usage of younger and more hearing-impaired children. All the BKB sentences were recorded by a male speaker of standard British English at a 22 kHz sampling rate. They were resampled to 16 kHz for the experiment to simulate the speech processing in a CI system.

All sentences were separately processed offline using ACE and wavelet packet-based strategies. They were corrupted by babble and speech-shaped noises at 5 dB SNR. A level of 5 dB SNR is encountered in many everyday environments (e.g. class rooms, and work environments) (Wilson and Dorman, 2008a). There were a total of 15 conditions (5 filter banks × 3 noises), as listed in Table 5.1.

Table 5.1 All conditions in this study.

Filter banks	Quiet	Noise level at 5 dB SNR	
		Babble noise	Speech-shaped noise
128-point FFT	C1	C6	C11
23-band WPT	C2	C7	C12
32-band WPT	C3	C8	C13
64-band WPT	C4	C9	C14
128-band WPT	C5	C10	C15

For the ACE strategy, the 128-point FFT with a frequency spacing of 125 Hz defaulted in the Nucleus device was used to compare with the wavelet packet-based strategy. For the wavelet packet-based strategy, four wavelet filter banks with sym8, including 23-, 32-, 64- and 128-band WPT, were implemented. The 23-band WPT was generated from a six-level decomposition. The 32-, 64- and 128-band WPT were generated from five-, six- and seven-level decompositions, and their frequency spacing was 250, 125 and 62.5 Hz, respectively. For all wavelet filter banks except 23-band WPT, the frequency bands were calculated by summing the power of adjacent frequency bands to generate 22 channels.

The filter spacing of the ACE and wavelet packet-based strategy was allocated using the Bark scale (as in Sections 3.3 and 3.4). The frequency bands and centre frequencies of the wavelet packet filter banks were specified as in Table 3.1 (Section 3.4) for the 23- and 64-band WPT and as in Table 5.2 for the 32- and 128-band WPT. Figure 5.1 shows the centre frequencies of all filter banks. It can be seen that the centre frequency of the 32-band WPT has a different frequency map to the 23-, 64- and 128-band WPTs. This is because the 32-band WPT has the widest frequency spacing (i.e. 250 Hz), so it is difficult to allocate frequency ranges close to the Bark scale or to form the signals sent to electrodes. The number of channels in each formant region for different filter banks is shown in Table 5.3.

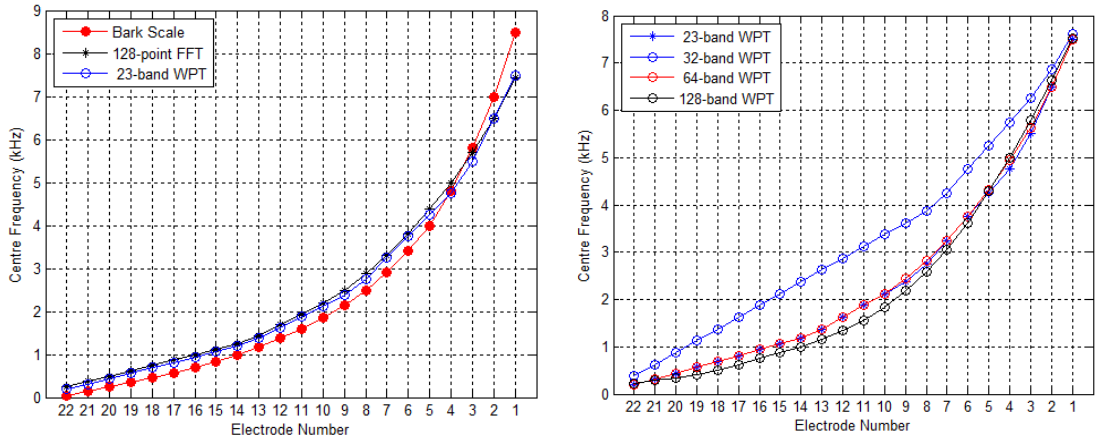


Figure 5.1 Centre frequencies of WPT and FFT filter banks. Comparison between 23-band WPT and 128-point FFT (left) and comparison among WPT filter banks (right)

Table 5.2 Frequency band and centre frequency in each channel of 32- and 128-band WPT at 16 kHz sampling rate.

Electrode channel number	32-band WPT			128-band WPT		
	$[f_l \quad f_u]$	f_c	Δf	$[f_l \quad f_u]$	f_c	Δf
22	250-500	375	250	187.5-250.0	218.75	62.5
21	500-750	625	250	250.0-312.5	281.25	62.5
20	750-1000	875	250	312.5-375.0	343.75	62.5
19	1000-1250	1125	250	375.0-437.5	406.25	62.5
18	1250-1500	1375	250	437.5-562.5	500.00	125.0
17	1500-1750	1625	250	562.5-687.5	625.00	125.0
16	1750-2000	1875	250	687.5-812.5	750.00	125.0
15	2000-2250	2125	250	812.5- 937.5	875.00	125.0
14	2250-2500	2375	250	937.5-1062.5	1000.00	125.0
13	2500-2750	2625	250	1062.5-1250.0	1156.25	187.5
12	2750-3000	2875	250	1250.0-1437.5	1343.75	187.5
11	3000-3250	3125	250	1437.5-1687.5	1562.50	250.0
10	3250-3500	3375	250	1687.5-2000.0	1843.75	312.5
9	3500-3750	3625	250	2000.0-2375.0	2187.50	375.0
8	3750-4000	3875	250	2375.0-2812.5	2593.75	437.5
7	4000-4500	4250	500	2812.5-3312.5	3062.50	500.0
6	4500-5000	4750	500	3312.5-3937.5	3625.00	625.0
5	5000-5500	5250	500	3937.5-4625.0	4281.25	687.5
4	5500-6000	5750	500	4625.0-5375.0	5000.00	750.0
3	6000-6500	6250	500	5375.0-6187.5	5781.25	812.5
2	6500-7250	6875	750	6187.5-7062.5	6625.00	875.0
1	7250-8000	7625	750	7062.5-8000.0	7531.25	937.5

Table 5.3 The number of channels in the F1/F2 region of all filter banks.

Formant region	Frequency range	Filter banks				
		128-point FFT	23-band WPT	32-band WPT	64-band WPT	128-band WPT
F1	0.3-1 kHz	6	7	3	7	8
F2	1-3 kHz	9	8	8	8	7
others	3-8 kHz	7	7	11	7	7

C. Procedure

The experiment was carried out in a sound-treated room. The subjects were asked to sign a consent form. Before the actual testing, a pure tone audiogram test was carried out to confirm that the subjects had normal hearing thresholds (≤ 20 dB HL, between 250 and 8000 Hz). The speech stimuli were presented using a Dell Latitude E4300

laptop, routed through a Creek Audio OBH-21SE headphone amplifier and presented unilaterally through a Sennheiser HDA280 circumaural headphone. Levels of speech stimuli in all experiments were presented at a comfortable conversional level (65 dB (A)).

Subjects were fully tested in a total of 15 conditions over two sessions on separate days, lasting approximately 1.25 hours each. They used their preferred ear (left or right) that was most comfortable for them to listen to the vocoded speech for the entire test. They were asked to write down the sentences that they heard. In the training session, they were asked to listen to one sentence list in both quiet and noisy conditions in a five-minute test in order to familiarise themselves with the vocoded speech and the testing procedures. This sentence list was not included in the actual testing.

In the testing session, two lists of BKB sentences (32 sentences) per condition were used to provide 100 keywords (100 percent). The sentences were scored in terms of the percentage of correct key words per condition, expressed as "percent correct." No list was repeated across the conditions in each session. The order of conditions and the list-to-condition mapping in each session was randomised across subjects. Subjects were given a five-minute break every 30 minutes during the test, or whenever they needed to take a rest.

D. Statistical analysis

The obtained scores were analysed using SPSS software version 21. A Shapiro-Wilk test (sample size < 50) was used to test the normality of the data distribution. For data with normal distribution ($p>0.05$), an analysis of variance (ANOVA) with repeated measure was used to investigate the difference between mean scores with different factors. Post-hoc tests (Bonferroni) were used to indicate differences between mean scores in the individual pair relationships in various conditions.

For data with non-normal distribution ($p<0.05$), a nonparametric Friedman's ANOVA was used to investigate the difference between mean scores with various factors. Post-hoc tests (Wilcoxon) were used to assess differences between mean scores in the individual pair relationships in various conditions.

5.2.1.2 Results

The boxplot and the mean percent correct scores for various filter banks in quiet and noisy conditions are shown in Figure 5.2. A Shapiro-Wilk test indicated that all the quiet conditions were not normally distributed, while all the noisy conditions were normally distributed.

In quiet conditions, a nonparametric Friedman's ANOVA with repeated measures showed a significant main effect of the filter banks (χ^2 [4, 9] =18.667, p=0.001). Post-hoc tests revealed a significant main effect of individual pairs of filter banks. The sentence score of the 32-band WPT was significantly lower than the others.

In noisy conditions, a two-way ANOVA with repeated measures showed a significant main effect of the filter banks (F [4, 32] = 82.509, p=0.001), a significant main effect of noise type (F [1, 8] = 25.004, p=0.001), and a nonsignificant interaction between filter banks and noise type (F [4, 32] = 2.316, p=0.079). These results indicated that speech intelligibility depends on the different filter banks and different noise type.

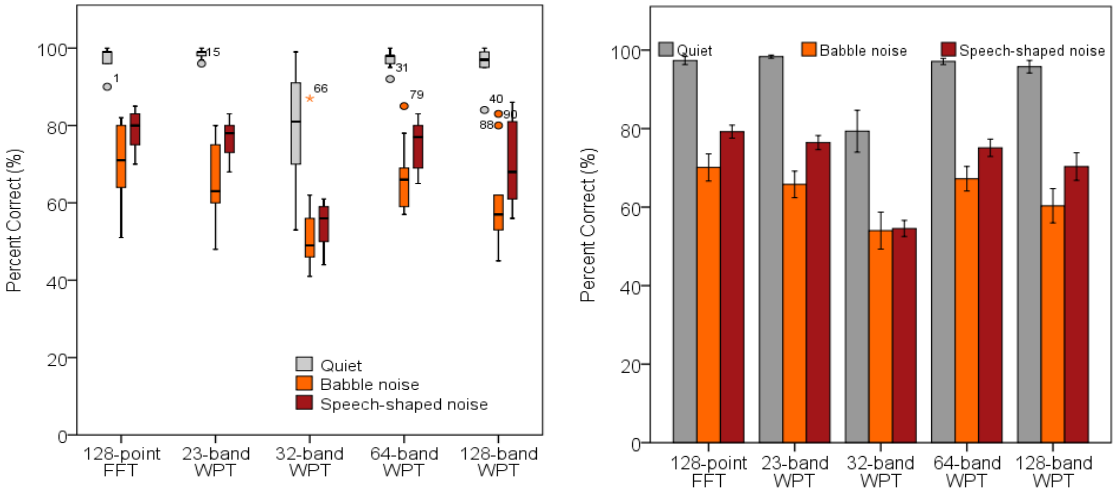


Figure 5.2 Boxplot and mean percentage correct scores for various filter banks in quiet and noisy conditions. The error bars indicate ± 1 standard error of the mean.

Post-hoc tests showed that the 32-band WPT produced significantly worse scores compared to the other filter banks. The 128-band WPT showed significantly lower performance when compared with the 128-point FFT and the 23-band WPT. The

performance of all filter banks in speech-shaped noise was significantly higher than the performance in babble noise (Figure 5.2). It can be seen that the 128-point FFT tended to slightly higher performance than the 23-, 64- and 128-band WPT. However, there was no statistically significant difference between the 128-point FFT, the 23- and 64-band WPT in noisy conditions.

5.2.1.3 Discussion

A. Relationship between frequency spacing of filter banks and the number of channels

The frequency spacing of different filter banks relates to the assignment of the number of channels in the formant regions. The narrower frequency spacing (e.g. 62.5 Hz of the 128-band WPT) is more flexible in terms of allocating the number of channels and in any frequency scale than the wider frequency spacing (e.g. 250 Hz of the 32-band WPT). It can be seen that the channel allocation of 32-band WPT in the F1/F2 region has fewer channels than the others as shown in Table 5.3.

The narrow frequency spacing of filter banks has some advantages for CI design. The narrow frequency spacing can be used to increase the available set of frequency tables in CI processors. This is useful for the clinician who will have suitable options for the frequency-to-electrode allocation for individual CI users, instead of the fixed tables for frequency allocation provided by manufacturers (Fourakis et al., 2007). In addition, the narrowest frequency spacing, 62.5 Hz in this study, may be sufficient for changes in pitch perception, especially for speech signals where the average pitch is close to 125 Hz (Mourad Ghrissi, 2012). The pitch also gives information about sentence prosody (e.g. statements and questions). It is also useful for tonal languages (e.g. Chinese and Thai), where pitch can be used to express semantic and grammatical cues (Mourad Ghrissi, 2012).

B. Relationship between different filter banks and speech intelligibility

Speech recognition performance was improved with the assignment of the more channels to the F1/F2 region. There was no significant difference on speech recognition performance for filter banks of 128-point FFT, 23- and 64-band WPT in both quiet and

noisy conditions, because their filter banks were based on Bark scale and the number of channels in the F1/F2 regions were the same (i.e. 15 channels). The 32-band WPT provided the lowest performance in both quiet and noisy conditions. This may result from the channel allocation of the 32-band WPT (i.e. 11 channels) in the F1/F2 region, which was less than the other filter banks (i.e. 15 channels).

The 128-band WPT provided significantly lower performance than the 128-point FFT and the 23-band WPT in noisy conditions, although it has 15 channels in the F1/F2 regions. This may result from the summing of the power of adjacent bands in all filter banks except the 23-band WPT to generate 22 channels. The 22 channels of the 128-band WPT were generated from 128 bands, whereas those of the 128-point FFT and 64-band WPT were computed from 64 bins/bands. Consequently, the 128-band WPT may provide higher noise power and lower performance than the others.

The obtained results were consistent with other studies in (Skinner et al., 1995; Skinner et al., 1997; Fourakis et al., 2004; Fourakis et al., 2007). Fourakis et al. (2007) suggested that a better performance may be achieved using a strategy whereby at least seven to eight channels are allocated below 1 kHz, with the majority of remaining channels allocated between 1−3 kHz, and the region above 3 kHz allocated only a few channels. In addition, the flexibility of such frequency band assignment should be adjusted in clinical practice to find the optimal frequency-to-electrode mapping in particular CI users.

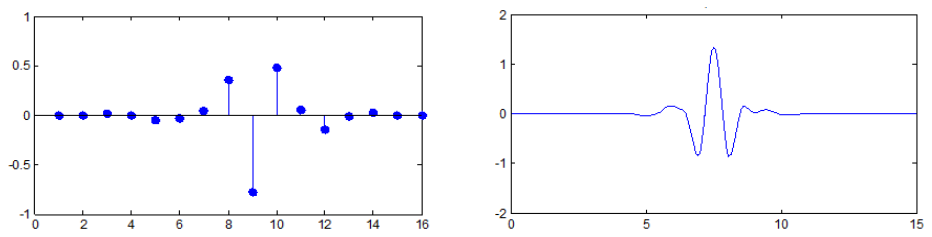
5.2.1.4 Conclusion

Different frequency spacings of the wavelet packet filter were associated with filter spacing (i.e. Bark scale) and the number of channels allocated in the F1/F2 region. The 128-point FFT and WPT filter banks (e.g. 23-, 64- and 128-band WPT) were based on Bark scale and the number of channels allocated in the F1/F2 regions of these filter bank was equal. Such assignment can provide the same speech recognition performance, except for 128-band WPT in noisy conditions. Generally speaking, the number of channels allocated in the F1/F2 region plays a critical role in speech recognition and depends on the filter spacing. The more channels are allocated in the F1/F2 region, the better the speech information that is perceived (McDermott, 1998; Mckay and Henshall, 2002; Fourakis et al., 2007).

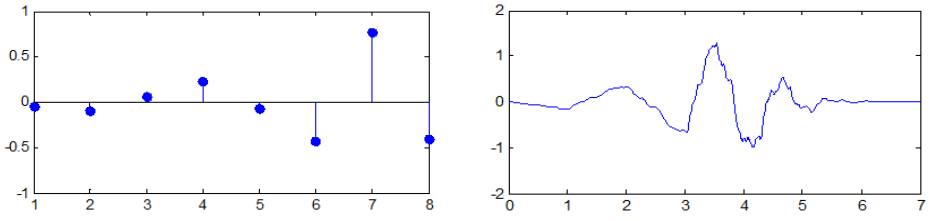
5.2.2 Experiment 2: Effect of perceptually optimised wavelet

The choice of reasonable mother wavelet for the wavelet packet-based speech coding strategy is an important issue (Nogueira et al., 2006; Karmakar et al., 2011). Several mother wavelets are provided in the wavelet toolbox of MATLAB, such as Daubechies, Symlet, and Coiflet wavelets (Appendix C). The mother wavelets of Daubechies and Symlets are widely applied in wavelet packet-based speech coding strategies (Nogueira et al., 2006; Gopalakrishna et al., 2010b). In this thesis, Symlet with order 8 (sym8) is implemented in all experiments.

Some mother wavelets have been designed based on the perceptual frequency scale and the temporal resolution of the auditory system – for instance, the bionic wavelet transform (BWT) was derived from the Morlet mother wavelet, and it has been used as one of the continuous wavelet transforms (CWTs). However, these methods do not provide the requisite structure for wavelet packet filter banks (Karmakar et al., 2011).



(a)



(b)

Figure 5.3 Coefficients of wavelet filters (left) and wavelet functions (right) for the sym8 with filter length of L=16 (a) the perceptually optimized wavelet (pow) with filter length of L=8 (b).

Karmakar et al. (2011) introduced the perceptually optimised wavelet (pow), which was optimally designed based on the Bark scale and the temporal resolution of the auditory system. The advantage of the pow wavelet is clearly visible in terms of the reduction in energy error in each channel in comparison with Daubechies, Symlet, and Coiflet wavelets at the same filter length. Figure 5.3 demonstrates the coefficient of wavelet filter and wavelet function for the sym8 with filter length of L=16 and the pow wavelet with filter length L=8.

The pow wavelet may lead to an increase in the speech recognition performance of wavelet packet-based CI processors. Therefore, this experiment investigates the hypothesis that the pow wavelet will improve speech intelligibility in quiet and noisy conditions.

5.2.2.1 Method

A. Subjects

Eight NH listeners participated in this experiment. All subjects were native speakers of British English (3 males, 5 females, from 18 to 34 years of age) and had normal hearing thresholds (< 20 dB HL). They were staff and students at the University of Southampton and were paid for their participation.

B. Stimuli

All sentences were separately processed offline using wavelet packet-based strategy in quiet and two different noisy conditions, at 5 dB SNR in babble and speech-shaped noises. The wavelet filter banks of 23- and 64-band WPT were used in this study as a result of the findings of Experiment 1. The mother wavelets of Symlet with order 8 (sym8) and pow were examined and compared. There was a total of 12 conditions (2 wavelet packet structures × 2 mother wavelets × 3 noises), as listed in Table 5.4.

C. Procedure

The procedures were the same as in Experiment 1 (Section 5.2.1). Subjects were fully tested in a total of 12 conditions for two sessions on separate days, lasting approximately one hour each.

D. Statistical analysis

The analysis was the same as in Experiment 1 (Section 5.2.1).

Table 5.4 All conditions in this study

Wavelet packet structures (mother wavelet)	Quiet	Noise level at 5 dB SNR	
		Babble noise	Speech-shaped noise
23-band WPT (sym8)	C1	C5	C9
23-band WPT (pow)	C2	C6	C10
64-band WPT (sym8)	C3	C7	C11
64-band WPT (pow)	C4	C8	C12

5.2.2.2 Results

A Shapiro-Wilk test indicated that the data under all the quiet conditions were not normally distributed, while the data in all noisy conditions were normally distributed. Figure 5.4 presents the boxplot and the mean percentage correct scores of both mother wavelets in quiet and noisy conditions. The boxplot in Figure 5.4 shows that the results contained a few outlying data points because some subjects performed poorly and produced low overall scores in the quiet condition.

In the quiet condition, a nonparametric Friedman's ANOVA with repeated measures indicated a nonsignificant main effect of different mother wavelets (χ^2 [3,8]=6.945, p=0.074). In noisy conditions at 5 dB SNR, a three-way ANOVA with repeated measures revealed a significant main effect of the different mother wavelets (F [1,7]=15.935, p=0.005), a nonsignificant main effect of wavelet packet structure (F [1,7] =0.711, p=0.427) and a nonsignificant main effect of noise type (F [1,7]=2.325, p=0.171). There was a significant interaction between mother wavelets and wavelet packet structures (F [1,7]=13.334, p=0.008). However, there was no significant interaction between wavelet packet structures and noises (F [1, 7] =0.098, p=0.763), between mother wavelets and noise type (F [1,7]=0.140, p=0.720), and between wavelet packet structures, mother wavelets and noise (F [1,7]=2.169, p=0.184).

Post-hoc tests indicated that speech intelligibility depends on the different mother wavelets and the optimal mother wavelets associated with wavelet packet structures. The pow wavelet yielded significantly lower speech intelligibility than the sym8

wavelet in almost all noisy conditions. For the 64-band WPT, the pow wavelet produced significantly lower performance than the sym8 in both babble and speech-shaped noise. However, the 23-band WPT with the pow wavelet provided better performance than the sym8 wavelet in speech-shaped noise. The pow and sym8 wavelets for the 23-band WPT had more similar mean scores than for the 64-band WPT.

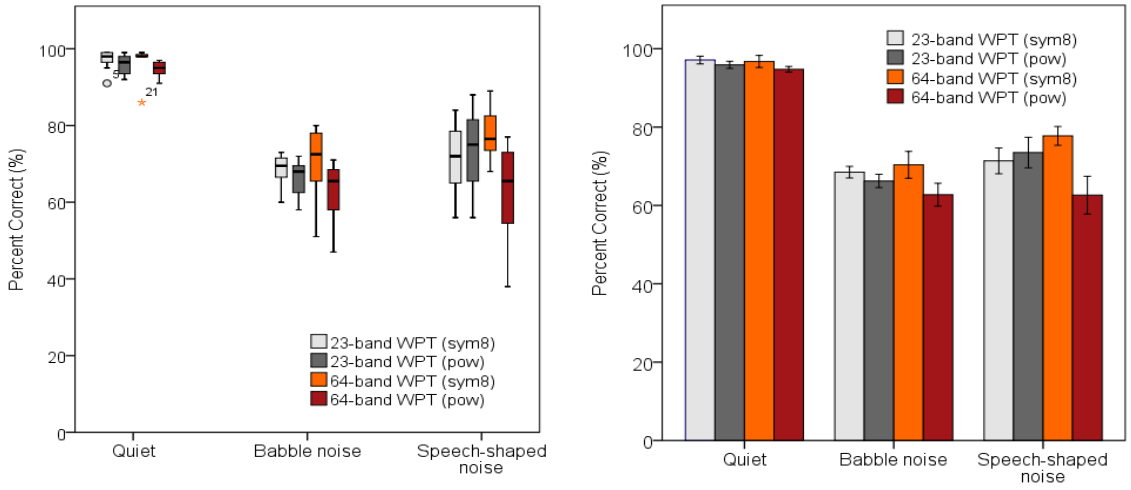


Figure 5.4 Boxplot and mean percentage correct scores for different mother wavelets in quiet and noisy conditions. The error bars indicate ± 1 standard error of the mean.

5.2.2.3 Discussion

The pow wavelet was worse with more frequency bands, especially the 64-band WPT. This is because the pow wavelet was derived from the structure of 21-band WPT and the temporal resolution of the human auditory system. Therefore, the pow may be more appropriate for the structure of 21-band WPT, but not for the structures of 23-or 64-band WPT due to the different structures of wavelet packets.

However, the structures of the 21- and 23- band WPT are very similar because both are constructed directly based on the Bark scale with different frequency ranges in each channel. In contrast, the structure of the 64-band WPT is originally constructed with equal frequency ranges for all 64 subbands and is not based on the Bark scale. Therefore, it is possible that the pow wavelet yielded better performance for the 23-band WPT than for the 64-band WPT.

5.2.2.4 Conclusion

The pow wavelet does not bring benefit in intelligibility performance in both quiet and noisy conditions. The properties of the pow wavelet and wavelet packet structures in CI processors might be more suitable for CI listeners than for NH listeners. In a further design for wavelet packet-based speech processors, the optimal mother wavelet should be derived from their wavelet packet structures (e.g. 23- and 64-band WPT) and the human auditory system to more closely match the behaviour of signals in healthy cochlea. Finally, the optimal mother wavelet can reasonably be selected by comparing the obtained results of these mother wavelets (Sang et al., 2009).

5.2.3 Experiment 3: Effect of frame length

Frame length is defined as the length of time (or the number of samples). Human speech is mixed between voiced and unvoiced sound. The duration of voiced sound is around 40−150 msec whereas unvoiced sound is around 10−50 msec (Shao and Chang, 2007). Therefore, the speech signal is a highly nonstationary signal and its power spectrum changes over time in a duration of above 250 msec. In speech processing the speech signal is segmented into a sufficiently short duration, and then its spectral characteristics are fairly stationary (Loizou, 2007).

Various frame lengths are used in the wavelet packet-based applications of CI processors and speech enhancement. Different frame lengths are used in wavelet packet-based CI processors such as 4 msec (Nogueira et al., 2006) and 16 msec (Gopalakrishna et al., 2010b). Frame lengths implemented in wavelet packet-based speech enhancement include 4 msec (Carnero and Drygajlo, 1999; Shao and Chang, 2007), 8 msec (Shao and Chang, 2011) and 32 msec (Cohen, 2001).

For the above reasons, the frame length is one of the important parameters for wavelet packet filter banks that may affect speech recognition performance in the wavelet packet-based speech coding strategies. This experiment aims to investigate whether the different frame lengths would affect speech recognition. The selection criteria from previous research on speech processing based on wavelet packets are considered, such as 4, 8, 16 and 32 msec.

5.2.3.1 Method

A. Subjects

Seven NH listeners participated in this experiment. All subjects were native speakers of British English (4 males, 3 females, from 18 to 34 years of age) and had normal hearing thresholds (< 20 dB HL). They were staff and students at the University of Southampton and were paid for their participation.

B. Stimuli

All sentences were separately processed offline using wavelet packet-based strategies under quiet and noisy conditions at 5 dB SNR in babble noise. The 23- and 64-band WPT with sym8 were used with frame lengths of 4, 8, 16, 32 and 64 msec (64, 128, 256, 512 and 1024 samples/frame). There were a total of 20 conditions (2 wavelet packet structures × 5 frame lengths × 2 noises), as listed in Table 5.4.

C. Procedures

The procedures were the same as in Experiment 1 (Section 5.2.1). Subjects were fully tested in a total of 20 conditions over two sessions on separate days, lasting approximately 1.5 hours each.

D. Statistical analysis

The analysis was the same as in Experiment 1 (Section 5.2.1).

Table 5.5 All conditions in this study

Wavelet packet structure (frame length)	Quiet	Babble noise at 5 dB SNR
23-band WPT (4 msec)	C1	C11
23-band WPT (8 msec)	C2	C12
23-band WPT (16 msec)	C3	C13
23-band WPT (32 msec)	C4	C14
23-band WPT (64 msec)	C5	C15
64-band WPT (4 msec)	C6	C16
64-band WPT (8 msec)	C7	C17
64-band WPT (16 msec)	C8	C18
64-band WPT (32 msec)	C9	C19
64-band WPT (64 msec)	C10	C20

5.2.3.2 Results

A Shapiro-Wilk test indicated that the data in all the quiet conditions were not normally distributed, whereas the data from all the noisy conditions were normally distributed. Figure 5.5 presents the mean percentage correct scores for the different frame lengths in quiet and noisy conditions.

In quiet conditions, a nonparametric Friedman's ANOVA with repeated measures indicated a nonsignificant main effect of the different frame lengths (χ^2 [9, 7] = 10.364, p=0.322). In babble noise at 5 dB SNR, a two-way ANOVA with repeated measures revealed a nonsignificant main effect of wavelet packet structures (F [1, 6] = 0.038, p=0.851), a significant main effect of the various frame lengths (F [4, 24] = 11.299, p<0.0005), and a significant interaction between wavelet packet structures and frame lengths (F [4, 24] = 3.508, p=0.022).

Post-hoc tests indicated that the frame lengths of 8 and 16 msec have significantly higher speech-intelligibility performance than the others. The 64-band WPT with a frame length of 8 msec provided significantly better performance than the 64-band WPT with a frame length of 4 msec.

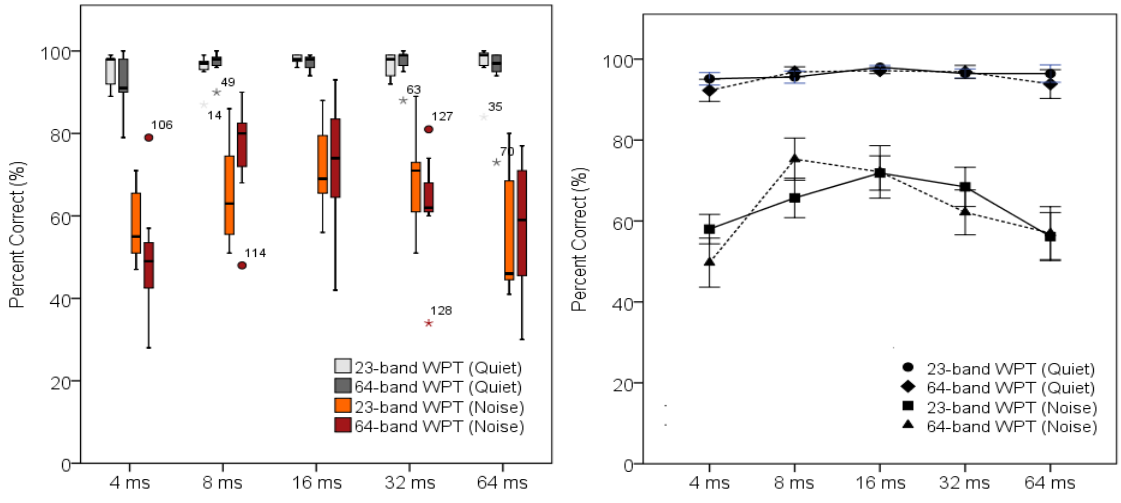


Figure 5.5 Boxplot and mean percent correct scores for the different frame lengths in quiet and noisy conditions. The error bars indicate ± 1 standard error of the mean.

5.2.3.3 Discussion and conclusion

The different frame lengths affected the performance of speech intelligibility in noisy conditions, but not in quiet conditions. The longer the frame length, the higher the computational complexity. A frame length of 8 msec has lower computational complexity than a frame length of 16 msec for the same speech intelligibility in both quiet and noisy conditions.

In addition, the different frame length might have an effect on speech analysis. The frame length of 8 msec may be sufficient to analyse information in the speech signal, particularly unvoiced sound (the majority of consonants). This might be useful for CI users in discrimination between voiced and unvoiced sound. Therefore, the frame length of 8 msec is more suitable than the others in wavelet packet-based speech coding strategies in terms of computational cost and speech analysis (Shao and Chang, 2007).

5.3 Noise reduction algorithms in the wavelet packet-based speech coding strategy

Noise reduction algorithms including time-frequency spectral subtraction (TFSS) and time-adaptive wavelet thresholding (TAWT) for the wavelet packet-based speech coding strategy are investigated in terms of different noise types and SNR levels, as in Experiment 1 and 2 respectively. The experiments were designed to determine whether noise reduction algorithms can improve speech recognition performance for different noise types and SNR levels.

5.3.1 Experiment 1: Comparison of noise reduction algorithms with different noise types

5.3.1.1 Method

A. Subjects

Ten NH listeners participated in this experiment. All subjects were native speakers of British English (6 males, 4 females, from 18 to 22 years of age) and had normal hearing

thresholds (< 20 dB HL). They were staff and students at the University of Southampton and were paid for their participation.

B. Stimuli

All sentences were processed separately offline using a wavelet packet-based strategy with and without noise reduction algorithms under quiet and two different noisy conditions: 5 dB SNR in babble and speech-shaped noises. The noise reduction algorithms including IdBM, TFSS and TAWT are provided in Section 4.3. There were a total of 18 conditions (2 Quiet + (4 algorithms × 2 noise types × 2 wavelet packet structures)).

C. Procedures

The procedures were the same as in Experiment 1 (Section 5.2.1). Subjects were fully tested over two sessions on separate days, lasting approximately 1.5 hours each. After the subjects were completely finished in each condition, they filled in the post-test questionnaire (Appendix F.1) to record information in terms of speech intelligibility.

D. Statistical analysis

The analysis was the same as in Experiment 1 (Section 5.2.1). The data from the post-test questionnaire were analysed using median values for each question.

5.3.1.2 Results

A Shapiro-Wilk test indicated that the data in all conditions were normally distributed. Figure 5.6 presents a boxplot and mean percentage correct scores for different algorithms of noise reduction in all conditions.

A three-way ANOVA with repeated measures was conducted with three main factors: algorithms, noise type, and wavelet packet structures. This revealed a significant main effect of algorithms (F [3, 27] =94.509, p<0.0005), a significant main effect of noise type (F [1, 9] =9.723, p=0.012) and a nonsignificant main effect of wavelet structures (F [1, 9] =0.228, p=0.644). There was a significant interaction between algorithms and noise type (F [3, 27] =7.119, p=0.001). However, there was a nonsignificant interaction between algorithms and wavelet packet structures (F [3, 27] =0.751, p=0.531), a nonsignificant interaction between noise type and wavelet packet

structures (F [1, 9] =0.572, p=0.469), and a nonsignificant interaction between algorithms, noise types, and wavelet packet structures (F [3, 27] =0.928, p=0.416).

Post-hoc tests were used to assess individual pair relationships between algorithms and noise type. The IdBM provided significantly higher scores than the TAWT and TFSS in both babble and speech-shaped noise. The TAWT provided significantly lower scores than the TFSS and the vocoded noisy speech corrupted by speech-shaped noise.

The post-test questionnaires were recorded by all subjects after listening to the vocoded speech in each condition (Appendix F.2). All subjects reported that the overall impression of sound quality for the quiet condition and IdBM was good, while the noisy conditions and noise reduction by TAWT and TFSS were reported to be poor. The listening efforts for the quiet condition and IdBM were negligible while the listening efforts for the noisy conditions and noise reduction by TAWT and TFSS were moderate in order to understand the key words and messages.

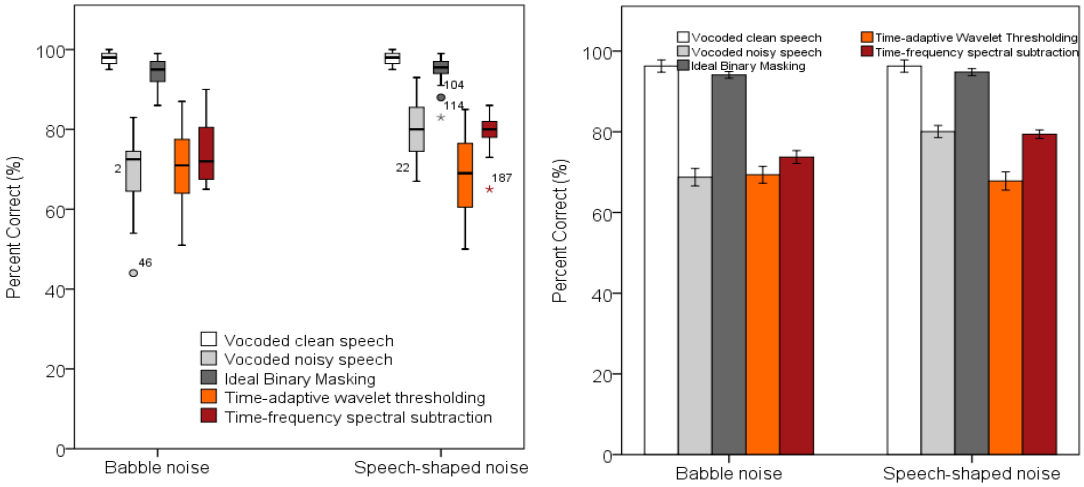


Figure 5.6 Boxplot and mean percentage correct scores for noise reduction algorithms in 5dB SNR babble noise and speech-shaped noise.
The error bars indicate ± 1 standard error.

The subjects felt that when listening over long periods of time, it was moderately easy to listen in quiet conditions and IdBM whereas it was difficult to listen in noisy conditions and with noise reduction by TAWT and TFSS. The articulation of the

vocoded clean speech and IdBM was clearly distinguishable. The articulation of TFSS was fairly clear, whereas the articulation of the vocoded noisy speech and TAWT was not very clear. Generally speaking, the articulation of TFSS was more distinguishable than that of TAWT.

5.3.1.3 Discussion and conclusion

A. Relationship between noise reduction algorithms and different noise types

Noise reduction algorithms including TFSS and TAWT were investigated and compared with IdBM in different noise types. The IdBM can restore speech intelligibility to the same level in as the quiet conditions. This study is consistent with those reported in IdBM studies of CI users by Hu and Loizou (2008) and Kokkinakis et al. (2011). The TFSS and TAWT do not significantly improve speech intelligibility when compared to vocoded noisy speech in both babble and speech-shaped noise at 5 dB SNR. Both TAWT and TFSS in babble noise provided similar scores to those in speech-shaped noise.

The IdBM is the noise reduction algorithm assuming the *priori* SNR is known. Whereas TFSS and TAWT do not assume prior knowledge of speech and noise information, they require the estimation of noise levels. Noise estimation in both TFSS and TAWT may be under- or overestimated and this results in distortion in the enhanced speech. The distortion of the enhanced speech may be more than the noise reduction, and may affect the speech intelligibility performance because speech discrimination becomes more difficult.

Figure 5.7 shows waveforms of vocoded clean speech and vocoded noisy speech with/without noise reduction algorithms for the BKB sentence "*The clown had a funny face*" processed by a wavelet packet-based speech coding strategy with 5 dB SNR babble noise. Figure 5.8 shows electric stimulation patterns (electrodograms), derived using the 12-of-22 strategy of the BKB sentence "*The clown had a funny face*". For all the electrodograms, the y-axis represents the electrode position corresponding to a specific frequency band and the x-axis represents time progression.

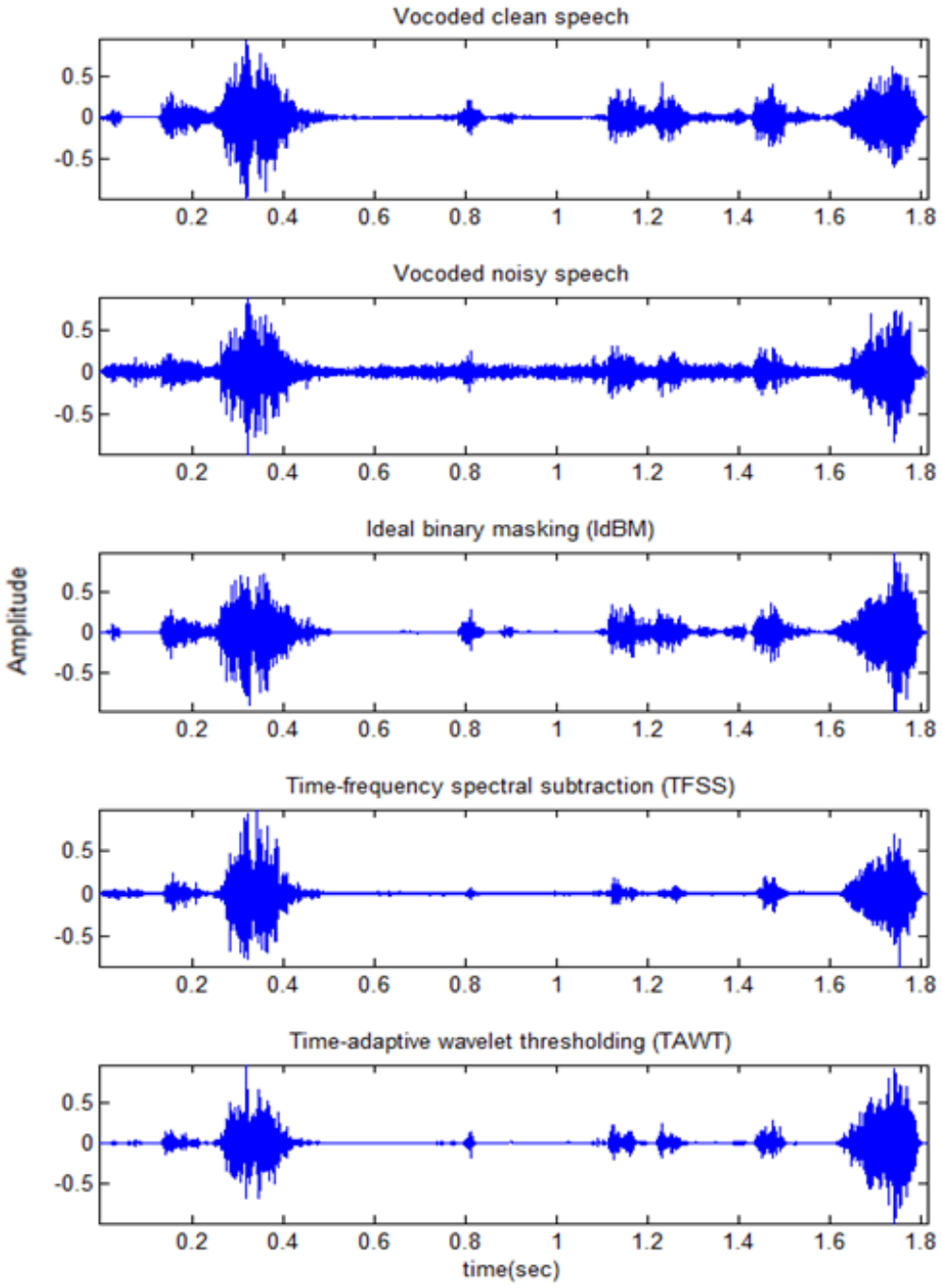


Figure 5.7 Waveforms of the BKB sentence "*The clown had a funny face*" for noise reduction algorithms. (Top to bottom) Plots showing vocoded clean speech, vocoded noisy speech at 5 dB SNR babble noise, vocoded noisy speech with the combination of IdBM and the *n*-of-*m* strategy, vocoded noisy speech with time-frequency spectral subtraction (TFSS) and vocoded noisy speech with time-adaptive wavelet thresholding (TAWT).

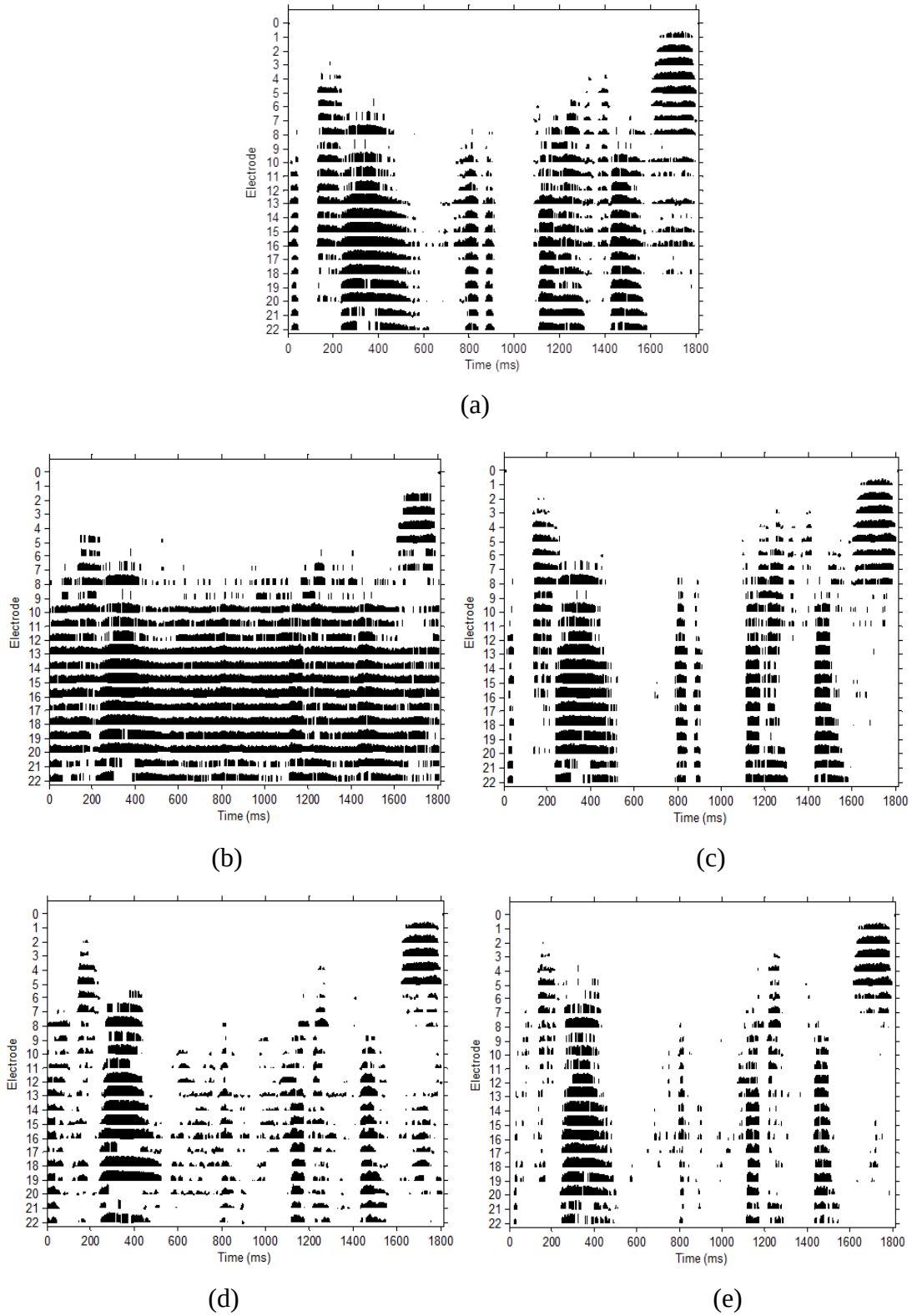


Figure 5.8 Electrodograms of the BKB sentence "*The clown had a funny face*" for noise reduction algorithms. (a) Clean speech. (b) Noisy speech with babble noise at 5 dB SNR. (c) Combination of IdBM and *n*-of-*m* strategy. (d) Time-frequency spectral subtraction (TFSS). (d) Time-adaptive wavelet thresholding (TAWT).

This figure shows the electrodogram of the vocoded clean speech and the vocoded noisy speech with and without noise reduction algorithms at 5 dB SNR babble noise. It can be seen that IdBM can preserve the important characteristics of vocoded clean speech, whereas TAWT and TFSS remove noise and some details of the vocoded clean speech in both waveforms and electrodograms. This may decrease speech intelligibility in CI systems.

B. Intelligibility judgements of noise reduction algorithms

The sentence scores were consistent with the post-test questionnaire results in cases of overall impression of sound quality, listening efforts for understanding messages, ease of listening for long period of time, and distinguishable articulation. The subjects reported that the vocoded clean speech and the IdBM were the same in all cases and the vocoded noisy speech with and without noise reduction by the TAWT and TFSS were the same results in almost all cases, except for distinguishable articulation. The TFSS gives more distinguishable articulation than TAWT.

C. Validity of objective intelligibility measures

Pearson correlation was performed to justify the correlation between objective and subjective intelligibility. Figure 5.9 shows the scatter plots of the NH listeners' mean scores against the predicted values of the NCM and STOI for different noise types. It can be seen that the NCM and STOI produced high correlations at $r = 0.81$ and $r = 0.88$ respectively. This high correlation indicates good validity. Therefore, the NCM and STOI can be pre-evaluated to predict the trend of intelligibility performance for NH listeners.

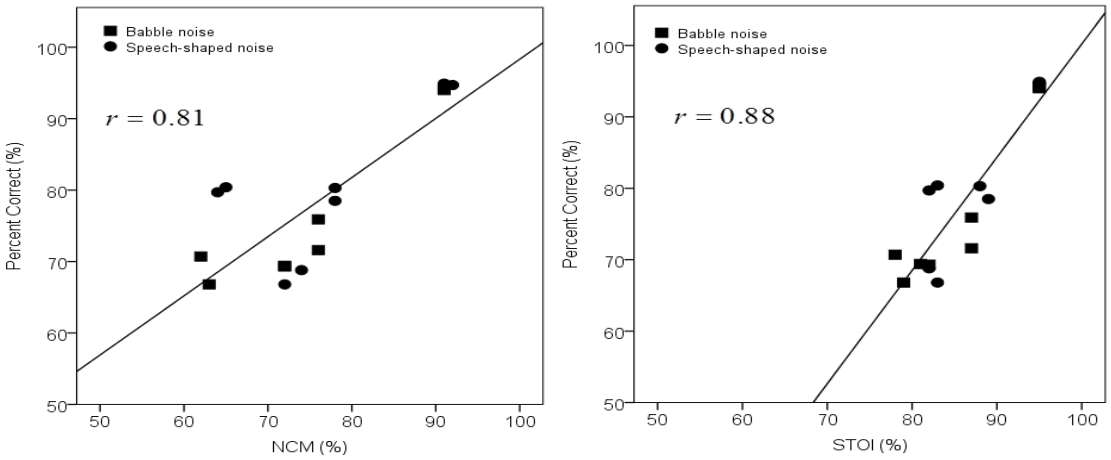


Figure 5.9 Scatter plots of mean scores obtained for sentence processed by noise reduction algorithms with different noise types against the predicted values of NCM and STOI.

5.3.2 Experiment 2: Comparison of noise reduction algorithms with different SNR levels

5.3.2.1 Method

A. Subjects

Fourteen NH listeners participated in this experiment. All subjects were native speakers of British English (8 males, 6 females, from 18 to 24 years of age) and had normal hearing thresholds (< 20 dB HL). They were staff and students at the University of Southampton and were paid for their participation.

B. Stimuli

All sentences were processed separately offline using a wavelet packet-based strategy with sym8 under quiet and noisy conditions. They were corrupted by babble noise at 0, 5 and 10 dB SNR, which are SNR levels where CI users can benefit (Fu et al., 1998). The noisy sentences were also processed using two algorithms for noise reduction (i.e. TFSS and TAWT). There were a total of 18 conditions (3 algorithms × 3 SNR levels × 2 wavelet packet structures).

C. Procedures

The procedures were the same as in Experiment 1 (Section 5.2.1). Subjects were tested over two sessions on separate days, lasting approximately 1.5 hours each. After the subjects were completely finished in each condition, they filled in the post-test questionnaire (Appendix F.1) to record information in terms of speech intelligibility.

D. Statistical analysis

The analysis was the same as in Experiment 1 (Section 5.2.1). The data from the post-test questionnaire were analysed using median values for in each question.

5.3.2.2 Results

A Shapiro-Wilk test indicated that the data in all conditions were normally distributed. Figure 5.10 shows a boxplot and mean percentage correct scores for the two algorithms of noise reduction in noisy conditions.

A three-way ANOVA with repeated measures was conducted with three main factors: algorithms, SNR levels, and wavelet packet structures. The results revealed a nonsignificant main effect of algorithms (F [2, 26] =1.391, p=0.267), a significant main effect of SNR levels (F [2, 26] =77.338, p<0.0005) and a significant main effect of wavelet packet structures (F [1, 13] =8.604, p=0.012). There was a significant interaction between algorithms and SNR levels (F [4, 52] =5.158, p<0.0005). However, there was a nonsignificant interaction between algorithms and wavelet packet structures (F [2, 26] =0.709, p=0.501), a nonsignificant interaction between SNR levels and wavelet packet structures (F [2, 26] =0.550, p=0.583), and a nonsignificant interaction between algorithms, SNR levels, and wavelet packet structures (F [4, 52] =0.172, p=0.952).

Post-hoc tests were used to consider the pair relationships among SNR levels, wavelet packet structures and between algorithms and SNR levels. The mean scores depended on the SNR levels. The higher SNR levels provided higher scores and vice versa. The 64-band WPT yielded slightly higher scores than the 23-band WPT in almost all conditions. The TFSS and TAWT provided significantly higher scores at 0 dB SNR

and significantly lower scores at 10 dB SNR when compared to the vocoded noisy speech at those SNR levels.

The post-test questionnaire results (Appendix F.3) revealed that the overall impressions of sound quality for the vocoded noisy speech at 0, 5 and 10 dB SNR were bad, poor and fair respectively. The overall impression of sound quality for TFSS and TAWT at all SNR levels was poor, except for TFSS at 5 dB SNR, when it was fair. The subjects required moderate listening effort to understand the messages, and they felt that it was difficult to listen for long periods of time for almost all conditions. The articulation of TFSS and TAWT at 0 dB SNR was not very clear, but it was clearer than for the vocoded noisy speech. The articulation of the vocoded noisy speech with and without noise reduction at 5 and 10 dB SNR was fairly clear.

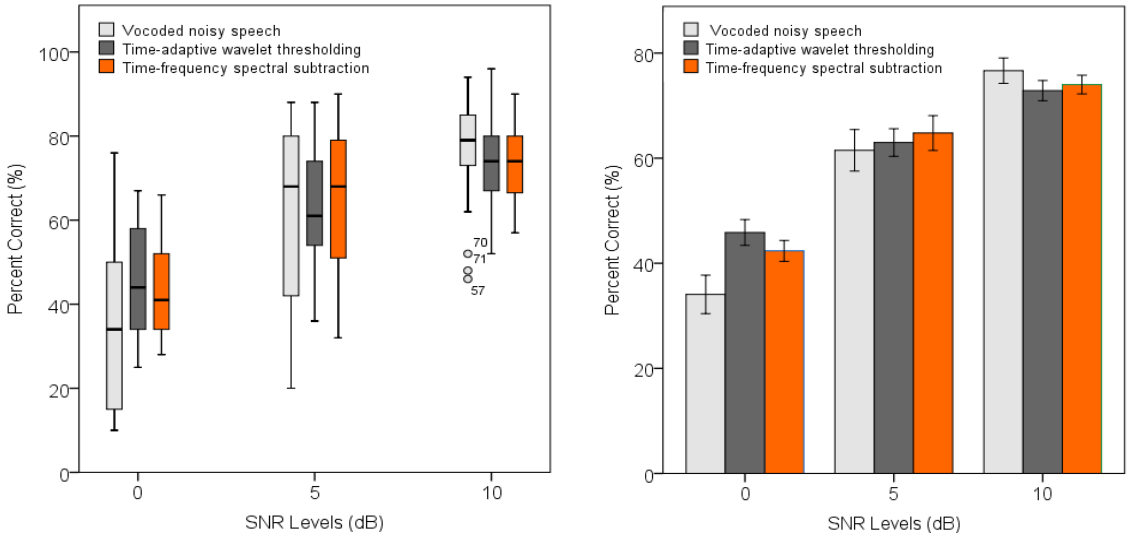


Figure 5.10 Boxplot and mean percentage correct scores for noise reduction algorithms at 0, 5, 10 dB SNR babble noise. The error bars indicate ± 1 standard error.

5.3.2.3 Discussion and conclusion

A. Relationship between noise reduction algorithms and different SNR levels

Noise reduction algorithms including TFSS and TAWT were investigated when speech is corrupted by babble noise at different SNR levels (i.e. 0, 5 and 10 dB SNR). Both TFSS and TAWT provided a significant improvement at 0 dB SNR, no significant

improvements at 5 dB SNR and significantly worse in speech intelligibility at 10 dB SNR when compared to vocoded noisy speech.

Theoretically, vocoded noisy speech with noise reduction algorithms should provide higher scores than vocoded noisy speech without noise reduction algorithms at high SNR levels. However, Figure 5.10 showed the TFSS and TAWT were significantly worse in performance at 10 dB SNR for babble noise. It seems possible that these results are due to noise estimation and difference between NH and HI listeners. It may be related to overestimated noise that distorts the enhanced speech. NH listeners are more sensitive to speech distortion and less sensitive to noise when compared to HI listeners (van Schijndel et al., 2001). NH listeners can reach ceiling performance at higher SNR levels without noise reduction algorithms (Sang, 2012).

In addition, the parameters of noise estimation in noise reduction algorithms, especially the TFSS. These parameters were selected based on predicting objective speech intelligibility for all SNR levels. The parameters should be adjusted to achieve the best results in each SNR level. Other factors related to noise estimation (e.g. noise types, choice of the local thresholds and speech materials) may also influence speech intelligibility. Moreover, the gain function should be more analysed on F1/F2 formant regions, which are important regions for speech intelligibility (Loizou and Gibak, 2011).

If noise estimation is accurate, it leads to good performance in noise reduction algorithms and provides significant improvements in speech intelligibility like IdBM. Although, noise estimation has never been able to accurately track the spectrum of nonstationary noise in practice (Loizou and Gibak, 2011), computing noise estimation without prior knowledge of speech and noise information remains a major challenge to potentially increase speech intelligibility in noisy environments (Kokkinakis et al., 2011).

B. Comparison between TFSS and TAWT

Both TFSS and TAWT are adaptive filters that can be applied to the real-world situation. However, weighting factors (i.e. β and α) for the TFSS were selected in the experiment (Section 4.3.3.2). These factors can be adjusted to optimise performance for each noise type and SNR level, which is a limitation of the TFSS. The trend of TAWT yielded higher scores than TFSS at 0 dB SNR where it is difficult. The TAWT does not

need to adjust any parameters. Therefore, the TAWT is more attractive and suitable than the TFSS in real practice.

C. Validity of objective intelligibility measures

Pearson correlation was performed to justify the correlation between objective and subjective intelligibility. Figure 5.11 shows the scatter plots of the NH listeners' mean scores against the predicted values of NCM and STOI for different SNR levels in babble noise. It can be seen that the NCM and STOI produced high correlations at r = 0.88 and r = 0.91, respectively. The predicted values of the NCM and STOI were validated to predict the vocoded noisy speech with and without noise reduction algorithms for NH listeners. The NCM and STOI were grouped clearly by SNR levels. Higher values of the NCM and STOI are expected for vocoded noisy speech with and without noise reduction algorithms at higher SNR levels.

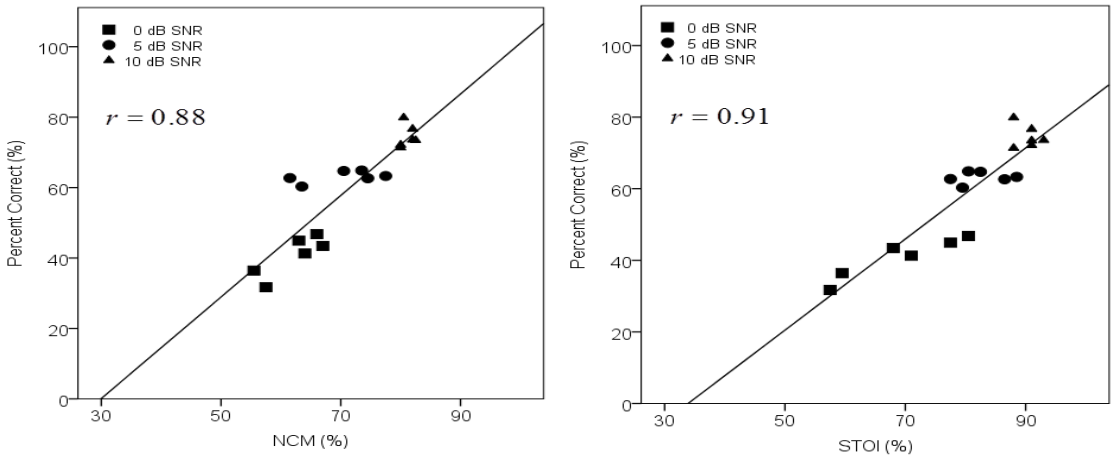


Figure 5.11 Scatter plots of mean scores obtained for sentence processed by noise reduction with different SNR levels against the predicted values of NCM and STOI.

5.4 General conclusions

The present study was designed to evaluate the effects of parametric variation of wavelet packet filter bank and noise reduction algorithms, on speech intelligibility for wavelet packet-based CI processors in both quiet and noisy conditions, using NH listeners. The data collected from all experiments can be concluded as follows:

5.4.1 Effect of parametric variation of wavelet packet filter bank

1. There was no statistically significant difference in speech recognition performance between 23- and 64- band WPT and 128-point FFT (a commercial Nucleus device) in both quiet and noisy conditions (i.e. babble noise and speech-shaped noise at 5 dB SNR). This was because these structures were designed based on the Bark scale and their numbers of channels allocated in the F1/F2 region were the same (i.e. 15 channels). The 23-band WPT can be designed directly for electrodes whereas the others need to aggregate subbands to form the signals sent to electrodes. However, the 64- or 128-band WPT might be more flexible for designing frequency-to-electrode allocations than the 23-band WPT.

In addition, a study of Nogueira et al. (2006) reported that 21-band WPT gave better speech recognition performance than the 128-point FFT when tested with CI users using a 15 dB SNR. The results indicated that a wavelet packet filter bank can be an alternative to the existing speech coding strategy that is used in commercial implants.

2. There was a nonsignificant effect of different mother wavelets (i.e. a perceptually optimised wavelet (pow) and a Symlet with order 8 (sym8)) in quiet conditions but there was a significant effect of different mother wavelets in noisy conditions. The performance of the sym8 was higher than that of the pow in almost all noisy conditions. This might be because the design method of the pow wavelet optimally exploits the structure of the 21-band WPT, but not the structure of the 23- or 64-band WPTs. Therefore, the pow wavelet worse than the other structures of wavelet packet filter bank.

3. There was a nonsignificant effect of different frame lengths in quiet conditions but there was a significant effect of different frame lengths in noisy conditions. The frame lengths of 8 and 16 msec have significantly higher performance than the others. However, the frame length of 8 msec was more appropriate for this speech processor in terms of computational cost and speech analysis.

4. In quiet conditions, it is difficult to assess the three parameters of wavelet packet filter banks to speech recognition performance due to a ceiling effect (i.e. mean scores $\geq$ 90%), except the filter bank of 32-band WPT. This does not determine the performance level of different parameters.

5. In all experiments examining the parametric variation of wavelet pack filter banks, the wavelet packet structure of 23- and 64-band WPT, the Sym8 mother wavelet and a frame length of 8 msec were found to be more suitable that other combinations for this wavelet packet-based CI processor. This optimal set of parameters may optimise speech intelligibility to benefit CI users. In addition, the optimal set of parameters will be useful for future work which may include the development of noise reduction techniques or other tasks designed to achieve additional enhancement in speech intelligibility or CI systems.

5.4.2 Effect of noise reduction algorithms

1. The IdBM as a baseline on denoising performance for NH listeners had the highest speech recognition performance and approached the speech recognition performance of vocoded clean speech (i.e. scores $\geq$ 90%) in both babble noise and speech-shaped noise at 5 dB SNR.

2. The TFSS and TAWT for NH listeners yielded little benefit in speech intelligibility. The TFSS and TAWT did not show significant improvements in speech intelligibility for different noise types (i.e. babble noise and speech-shaped noise) at 5 dB SNR. Both TFSS and TAWT showed a significant improvement at 0 dB SNR, were not significant difference at 5 dB SNR and were significantly worse in speech intelligibility at 10 dB SNR, when compared to vocoded noisy speech in babble noise.

3. The correlation between mean scores of NH listeners and the predicted values of NCM and STOI was strong. The NCM and STOI were found that they can pre-evaluate to predict the trend and pattern of the speech recognition performance in wavelet packet-based speech coding strategies with noise reduction algorithms for NH listeners.

Chapter 6: General discussion

This chapter provides a general discussion of the factors and limitations that influence the speech recognition performance of noise reduction algorithms in wavelet packet-based speech coding strategies. This discussion reports the limitations of WPT for CI processors, objective intelligibility measures, vocoder simulation, and performance evaluation. The limitations of this study are discussed. Finally, some suggestions for future work are given.

6.1 Limitations of WPT

The advantages of WPT are described in Chapters 2 and 3; however, the use of WPT has some limitations that lead to some undesired effects in the development of noise reduction algorithms in wavelet packet-based CI processors.

6.1.1 Problems with WPT

The main problem of WPT is shift variance due to the downsampling operation at each level of decomposition. When a signal is downsampled by 2, the output is only one sample which is selected between two consecutive samples of the signal. A sample is removed, which may contain important information about the speech signal. The downsampling operation results in a shift in time of signals, which produces the differences in the energy distribution of wavelet coefficients. The speech signal may be distorted due to the loss of information, which reduces the speech intelligibility and the perceptual quality of speech signals in both CI processors and noise reduction algorithms.

There is no evidence of this impact on CI design but this problem is referred to in some studies in the area of speech enhancement (Tasmaz and Ercelebi, 2008; Litvin and Cohen, 2011). Other WPTs are proposed to overcome the problem of shift variance, such as the dual-tree complex wavelet packet (Bayram and Selesnick, 2008) and the analytic wavelet packet (Weickert et al., 2009). These WPTs may potentially increase

performance significantly over that of conventional WPTs for CI design and noise reduction algorithms.

6.1.2 Wavelet packet-based speech coding strategies

Based on the results from Experiment 1 (Section 5.2.1) in terms of filter spacing, the 23- and 64-band WPT provided slightly worse performance, but this was not statistically significant when compared to 128-point FFT in noisy conditions. This may produce an energy error in the noise and speech signal in each channel due to the shift variance of WPT (Weickert et al., 2009).

The system complexity can be reduced by processing perceptual wavelet subbands like the 23-band WPT. However, the 64-band WPT has higher frequency resolution than the 23-band WPT and this is more beneficial for noise reduction algorithms before it generates 64 bands into 22 bands (Cohen, 2001). In addition, the signal processing stages of CI design based on wavelet packets (e.g. pre-emphasis and envelope detection) are important issues that may affect outcome.

6.2 Limitations of objective intelligibility measures

Based on Experiment 1 and 2 (Section 5.3), the scores from NCM and STOI were highly correlated with mean scores obtained from NH listeners. Therefore, both NCM and STOI can be used to predict the trend of intelligibility performance for noise reduction in wavelet packet-based speech coding strategy. The present study is consistent with the outcome reported in Sang (2012) for the NCM and STOI in both NH and HI listeners, and in Jianfen et al. (2009) for the NCM in NH listeners.

However, these measures may reliably predict the intelligibility performance, in particular NH listeners, but may not predict or reflect to the recognition performance for CI users. The objective intelligibility measures should include important information of individual CI users such as threshold levels and comfort levels (i.e. dynamic range compression) and the number of active electrodes. The use of information contained in the electrodogram may be sufficient to predict CI user's intelligibility performance.

These allow a great benefit for developing better noise reduction algorithms that require adjustments of the number of parameters and the selection of a noise-reduction algorithm to provide the best performance for a particular CI user.

6.3 Limitations of vocoder simulation

Vocoder simulation with NH listeners may be used to predict the trend and pattern of performance for CI users, but actual testing with CI users may reveal the effect of noise reduction algorithms in wavelet packet-based speech coding strategies. CI designs need to take account of the limitations of vocoder simulation, such as differences in vocoder simulation and processing in a CI device, differences between acoustic and electric hearing and the effects of noise reduction algorithms for NH listeners and CI users. These are critical to the reliability of comparative studies.

6.3.1 Differences between vocoder simulation and processing of a CI device

The vocoder simulation is processed in a similar manner to the signal processing of a CI device. However, some stages of CI devices may not be included in vocoder simulation, such as pre-emphasis and dynamic range compression (Nogueira et al., 2005). These may provide different characteristic of signals and different performance between vocoder simulation and processing of CI devices in both quiet and noisy conditions.

Several studies in vocoder simulation either include (Li, 2009; Chen and Loizou, 2010) or exclude (Mourad Ghrissi, 2012) the stage of pre-emphasis. The pre-emphasis in the vocoder simulation is used to amplify the high-frequency components of speech perception, and it also amplifies noise in noisy speech. Recognition performance may decrease when compared to vocoder simulation without pre-emphasis in noisy conditions.

The dynamic range compression aims to optimally map acoustic amplitudes in speech sounds to electrical amplitudes that reach the audible threshold (T-level) and most comfortable loudness level (C-level). T and C levels are important parameters that would be measured and adjusted for optimal amplitude mapping depending on individual CI users (Zeng, 2004; Loizou, 2006). A few studies in the shape of the

logarithmic mapping function had only a minor effect on consonant and vowel recognition performance in quiet conditions. However, it may be that they provided different performance in a similar study in noisy conditions (Loizou et al., 2000). Another study indicated that increasing input to the dynamic range improved sentence recognition performance in both quiet and noisy conditions (Spahr et al., 2007).

6.3.2 Differences between acoustic hearing and electric hearing

The acoustic hearing of NH listeners and the electric hearing of CI users can be difficult issues when comparing stimuli presented to both groups. The vocoded speech perceived by NH listeners was processed not only by vocoder simulation but also by the external, middle and inner ear, while the vocoded speech perceived by CI users was processed only by CI devices. NH listeners can listen with a healthy auditory system throughout the cochlea, while CI users may have residual auditory nerves throughout the cochlea which are stimulated by the electrode arrays (Zeng, 2004).

The frequency mapping of each electrode relating to the actual position of stimulation in the cochlea may significantly improve CI performance (Stakhovskaya et al., 2007). Channel interactions between the electrodes occur as the current from one electrode spreads to adjacent regions covered by other electrodes in the cochlea. This may distort speech information and degrade speech intelligibility. Channel interaction depends on many factors such as electrode spacing and channel stimulation rate. A wider spacing between electrodes produces a smaller amount of channel interaction and more benefit with a high stimulation rate (Loizou, 2006).

6.3.3 Differences of noise reduction algorithms for NH listeners and CI users

The main concept of noise reduction algorithms is to compromise between noise reduction, speech distortion and the level of residual noise (Virag, 1999). However, NH listeners are less sensitive to noise, but more sensitive to speech distortion when compared to HI listeners (van Schijndel et al., 2001). Another study suggested that HI listeners can bear higher levels of distortion than NH listeners. As the result, noise reduction algorithms with more aggressive gain functions (Hu et al., 2007; Qazi et al., 2012) should be used for CI users to reduce more amount of noise, while noise

reduction with less more aggressive gain functions should be used for NH listeners to preserve the listening quality (Gustafsson et al., 1998).

This reason is consistent with some studies indicating that almost all algorithms for single-microphone noise reduction algorithms provide little benefit or do not improve speech intelligibility for NH listeners in American English (Hu and Loizou, 2007; Li et al., 2011) and other languages (i.e. Chinese and Japanese) (Li et al., 2011). This is due to distortion of enhanced speech resulting from inaccurate noise estimation or excessive noise reduction.

In contrast, some single-microphone noise reduction algorithms have been developed for NH listeners; they show speech intelligibility improvements for CI users. These are algorithms such as spectral subtraction (Yang and Fu, 2005; Verschuur et al., 2006; Kallel et al., 2012), statistical-model based methods (Hu et al., 2007; Li, 2008; Dawson et al., 2011) and subspace algorithms (Loizou et al., 2005). Additionally, another study of single-microphone noise reduction algorithms for HA users included sparse coding shrinkage (Sang, 2012).

General speaking, CI users preferred the more aggressive gain function rather than the less aggressive gain function for noise reduction in their devices. This resulted from impaired auditory factors such as reduced frequency selectivity and reduced temporal resolution. This can lead to significant recognition performance improvement.

6.4 Limitations of performance evaluation

The methodology of performance evaluation for speech intelligibility reflects the reliability and accuracy of outcomes. An appropriate evaluation can examine the effect of interesting parameters to achieve CI development. In contrast, an inappropriate evaluation can result in a misleading interpretation and problems with CI development. Some factors should be considered when interpreting the obtained results, such as speech materials, noise types, SNR levels and variability of subjects.

6.4.1 Choice of speech materials

Speech perception abilities of subjects can be typically evaluated using vowels, consonants, words or sentences (Loizou, 1998). Studies have produced overall scores with different results depending on whether they were looking at consonant, vowel, word or sentence recognition. The consonants, vowels and words are useful in terms of the distinguishable errors of subjects while the sentence recognition is more suitable to representation in real-life communication (Loizou, 2007).

In sentence recognition, subjects need time to distinguish between noise and speech, which is especially unclear at the start of sentences. They were able to use knowledge (e.g. context, grammar and semantics) to identify the correct sentences when they only heard one or two keywords from those sentences (Loizou, 1998; Loizou, 2007). As a result, subjects tend to achieve higher recognition scores in sentence tests than in other tests in noisy conditions.

Different speakers for the same materials may also influence the comparability of results, in terms of gender, nationality, age and a single- or multiple-speaker setup. Moreover, speaking style and rate may also have an impact on performance. Additionally, the number of items in a list of speech materials should provide flexibility to cover an experimental design in all interesting conditions. Having sufficient items avoids repetition in the speech materials.

In the current study examining noise reduction in wavelet packet-based speech coding, consonants, vowels, and words might have been more useful to reveal the capability of wavelet packet-based speech coding and noise reduction techniques than the sentences.

6.4.2 Choice of noise types and SNR levels

The results obtained from some studies indicated that the overall performance from vocoded speech and NH listeners in noisy conditions was close to that of vocoded speech in quiet conditions, due to a ceiling effect. On the other hand, vocoded speech and CI users in noisy conditions at lower SNR levels were closer to floor effects. The ceiling and floor effects make it difficult to interpret the effect of interesting parameters.

These may result from the inappropriateness of either noise type or from the SNR levels.

Different noise types influence speech intelligibility. Different noise types in the same speech materials also give an impact on different intelligibility performance. Numerous noise types in the real world (e.g. babble noise, speech-shaped noise and reverberation noise) are implemented to create degrading speech cues. Babble noise depending on the number of talkers in the mixture is more realistic for CI users in everyday situations and is widely used in studies of speech perception in noise (Simpson and Cooke, 2005; Verschuur, 2007).

SNRs of 0, 5 and 10 dB are levels where CI users can benefit (Fu et al., 1998) and a SNR of 5 dB is normally encountered in everyday situations (Wilson and Dorman, 2008a). Most CI users require approximately 10 to 25 dB higher SNR than NH listeners to achieve similar speech recognition performance (Qazi et al., 2012). In some studies, NH listeners reached ceiling effects with SNR levels such as 10 dB SNR, while CI users reached a mean score of 75% with the same SNR (Zeng, 2004). The use of a lower SNR in some tests for NH listeners may be required to determine actually the performance improvement. The interpretation of obtained results should be cautious due to differences between testing NH listeners and CI users.

6.4.3 Variability of subjects

All subjects require more effort and concentration to understand vocoded speech in noisy conditions than in quiet conditions. It is well known that CI users perform worse than NH listeners in the same noisy conditions. This results from the elevated threshold, loudness recruitment, and poor temporal and frequency resolution. The lack of motivation, attention, confidence and language skill may also have contributed to the lower speech recognition performance.

In most studies, the outcomes of NH listeners can be used to predict the trend and pattern of CI performance. However, there are other discrepancies between testing NH listeners and CI users which may affect overall performance level, such as experience with vocoded speech, learning effects and the age of subjects.

NH listeners usually have no experience with vocoded speech, while CI users have had a period of acclimatisation to vocoded speech and their implant devices over at least one year. However, NH listeners without experience in vocoded speech can achieve higher speech recognition than CI users with prolonged experience (Fu et al., 1998). Some studies indicated that NH listeners take a longer time for training to familiarity with vocoded speech in both quiet and noisy conditions, which may increase speech recognition performance (Dorman et al., 1997). The age difference for NH listeners suggests that older subjects (average age of 70) required more stimulation channels than the younger individuals (average age of 22), i.e. approximately 9 and 6 channels respectively (Sheldon et al., 2008).

6.5 Limitations of this study

6.5.1 Speech materials

Designing listening tests with interesting parameters (e.g. noise types, SNR levels and noise reduction algorithms) in each experiment was limited by the number of BKB sentences lists. A total of 21 BKB sentence lists can only be employed to 10 conditions per session (2 lists per condition). Each experiment has to be done by undertaking listening tests in at least two sessions on separate days. That means that all or some sentence lists may be repeated in different sessions. Some subjects may recognise key words in sentences from the first session. This influences sentence recognition scores. Different speech materials may be used to investigate the differences in obtained results in future research looking at the wavelet packet-based speech coding strategy with noise reduction algorithms.

6.5.2 Learning effect

The variation and difference of speech recognition scores may be dependent on individual subjects. The subjects have unique factors in listening tests such as motivation, attention, confidence and ability. Though the order of conditions and the list-to-condition mapping were randomised, some subjects reported that the former conditions were more difficult for the listening test than the latter conditions. If the

subjects have more training, they have a chance to become more familiar with vocoded speech with and without noise reduction algorithms. This may result in higher speech recognition scores.

The learning effect can be mitigated by randomisation of the experiment, in which the order of conditions and the list-to-condition mapping could be randomised both for individual subjects and between -subjects, to prevent the repeated use of a sentence list and to continuously vary the sentence lists being evaluated. A Latin square randomisation might be used to assign the order of vocoded speech for subjects.

6.5.3 Comparison of previous study

In the previous study, the wavelet packet-based speech coding strategy with different mother wavelets was investigated and compared to a commercial ACE strategy at 15 dB SNR in terms of speech intelligibility for CI users (Nogueira et al., 2006). Another study (Gopalakrishna et al., 2010b) explored wavelet packet-based real-time CI processors in terms of computational complexity, spectral leakage, fixed-point accuracy and tracking temporal envelope features. However, none of the wavelet packet-based CI processors with noise reduction algorithms were investigated and tested with subjects who were both NH listeners and CI users.

It is very difficult to compare between this study and other studies (Nogueira et al., 2006; Gopalakrishna et al., 2010b) due to different parameters in terms of different noise types, SNR levels, speech materials, speech coding strategies, number of channels and evaluation of subjects. In addition, the wavelet packet-based speech coding strategy with noise reduction algorithms was developed in limited time and several parameters could be further explored and studied before it is compared to other strategies used in commercial CIs such as the ACE strategy.

6.5.4 Comparison between NH listeners and CI users

A problem in comparing vocoder simulations to CI speech recognition is the earlier mentioned fact that CI users vary greatly in speech recognition performance due to numerous factors. Listening tests using vocoded speech involve presenting parameter setting to NH listeners. Current vocoder simulation in studies using NH listeners does not necessarily provide conclusions about the absolute performance level of CI users.

These studies can only be used as useful information about tendencies in speech recognition for CI users (Chen and Loizou, 2011). Further studies of wavelet packet-based strategies with noise reduction may benefit CI users, as suggested in Section 6.3.3.

6.5.5 Statistical analysis

Although the number of NH listeners participating in this study is normal and satisfies good academic practice, but the study is underpowered. A larger sample size would provide better statistical power to indicate clearer comparisons and allow an examination of either the effect of parametric variation in wavelet packet filter banks, or the relationship between vocoded speech with and without noise reduction algorithms.

6.6 Conclusion

There are several limitations of the development of noise reduction by a wavelet packet-based strategy, as mentioned earlier. The limitations might confound the obtained results in terms of intelligibility performance. All of them are important issues for CI noise-reduction studies. However, the most important limitations of this study are the issues of WPT, vocoder simulation, and performance evaluation.

Since the problem of WPT is shift variance due to the downsampling stage of decomposition, this problem may reduce the performance of speech coding and denoising in CI processors. Some applications based on WPT (e.g. sound source separation) have reported that this problem can reduce the utility of audio signal processing (Litvin and Cohen, 2011). However, none of the WPT-based CI processors have shown whether this problem affects intelligibility performance. Some researchers have proposed different methods to address shift variance, which leads to the generation of new WPTs. Other WPTs, namely the dual-tree complex wavelet packet (Bayram and Selesnick, 2008) and the analytic wavelet packet (Weickert et al., 2009), may mitigate CI noise-reduction approaches.

The limitations of vocoder simulation include differences in vocoder simulation and processing in a CI device, differences between acoustic and electric hearing, and the

effects of noise reduction algorithms for NH listeners and CI users. These differences may affect the reliability of CI noise-reduction studies. Almost all CI noise-reduction algorithms are evaluated by CI users, who provide more reliable and informative results than NH listeners. The procedure of the listening test used in previous studies allows for all materials to be presented directly to the CI users via either the auxiliary input jack of their CI processors (Loizou et al., 2005; Hu et al., 2007; Li, 2008) or loudspeakers (Yang and Fu, 2005; Verschuur et al., 2006; Dawson et al., 2011). It is not known whether or not any differences in listening tests provide the same benefits to intelligibility performance.

In terms of performance evaluation, the sentence test may be more appropriate for real-life communication, but this test may not reveal informative results for the effect of CI noise reduction. The sets of vowels, consonants or words may be more suitable for analysing spectral and temporal information to evaluate the subject's perception ability (Yao and Zhang, 2002). Another evaluation (i.e. speech reception threshold (SRT)) may reveal the capability of the subject's perception in diverse environments better than fixed SNR. A little bit of SNR levels might provide very different way of intelligibility performance. In addition, SRT can avoid the problem of ceiling/floor effects, which actually helps in the study of understanding speech.

These limitations involve either CI-processor related factors or CI-user related factors, both of which are important to CI noise-reduction studies. They require very expensive and time-consuming measures to evaluate the effect of parametric variations on intelligibility performance. However, this problem can be mitigated with objective intelligibility measures. Objective measures can be used for tuning parameters during the development of wavelet packet-based speech coding and denoising to choose the right set of parameters for CI noise-reduction approaches and to predict the trend of intelligibility performance for CI users.

6.7 Future research

6.7.1 Optimal wavelet functions and wavelet structures

Research in modern wavelets aims to create a set of wavelet function and transform that provide efficient signal analysis, detection, estimation and denoising in numerous applications. There are many choices of wavelet function and transform that can be selected for application in speech and auditory processing. The potentialities and benefits of wavelets are unlimited for development in this research area − for instance, finding an optimal wavelet function, an adaptive optimal wavelet decomposition tree or making a new set of wavelet functions for CI systems.

6.7.2 Noise reduction algorithms

The noise reduction problem remains a challenge. If it is possible, the techniques of noise estimation can improve SNR estimation as the IdBM techniques. This may lead to further improvement in speech intelligibility. Statistical model-based noise reduction, such as Bayesian approaches, can be applied to noise reduction algorithms (e.g. spectral subtraction and wavelet shrinkage). The combination between noise reduction and entropy analysis may increase the benefit of speech enhancement and speech perception. Further comparison of multiple microphones or binaural processing-based noise reduction algorithms should be carried out to find the benefit in terms of speech intelligibility.

6.7.3 Objective speech intelligibility measures

Most of the modern objective intelligibility measures are used for predicting the trend of recognition performance in NH listeners rather than CI users due to lack of useful information of individual CI users. The electrodogram as a representation of CI output may provide important information directly to predict the outcome of speech intelligibility for particular CI users. In addition, objective intelligibility measures should predict performance of speech perception from only noisy speech without the clean speech because the clean speech is often not available in real-world applications.

Chapter 7: Conclusions

This thesis focuses on single-microphone noise reduction strategies for wavelet packet-based CI processors to improve speech intelligibility in noisy conditions. The research contribution can be divided into two parts: the wavelet packet-based CI processor, and noise reduction algorithms. The wavelet packet-based speech coding strategy was developed. The effect of parametric variation of wavelet packet filter bank on speech intelligibility by NH listeners was evaluated to find optimal parameters in terms of filter spacing, optimal wavelet function and frame lengths.

Noise reduction algorithms (i.e. IdBM, TFSS and TAWT) were integrated within wavelet packet-based speech coding strategies and applied directly in time-frequency envelope amplitude. Objective speech intelligibility measures (i.e. NCM and STOI) were employed to predict the trend of speech intelligibility for noise reduction algorithms in all noisy conditions before they were evaluated by NH listeners under different noise types and SNR levels. This research contributes the following conclusions:

1. The wavelet packet-based CI processor can provide an alternative to existing CI systems (e.g. the ACE strategy).
2. Three parameters (i.e. filter spacing, optimal wavelet function and frame length) of the wavelet packet filter bank have influences on speech intelligibility, especially in noisy conditions.
3. The IdBM is an ideal method of noise reduction and its intelligibility performance is nearly 100% or similar to vocoded clean speech. The TFSS and TAWT have produced little benefit in terms of speech intelligibility for NH listeners. The TFSS requires tuning in some parameters to get the best performance in each noisy condition, but not the TAWT.
4. The NCM and STOI can be used for predicting the trend of intelligibility performance for noise reduction algorithms in the wavelet packet-based CI processor for NH listeners, but may not reflect the reliability of recognition performance among CI users due to the lack of impaired auditory necessary information of individual CI users.

In the present study, we believe that the approach of applying wavelet analysis and wavelet shrinkage (e.g. TAWT) is an outstanding candidate for the next generation of modern prosthesis devices such as CI processors.

Appendices

Appendix A: Publication

The 2014 7th International Conference on BioMedical Engineering and Informatics

Improving Speech Intelligibility in Perceptual Wavelet Packet-Based Speech Coding for Cochlear Implants

Siriporn Dachasilaruk, Stefan Bleeck and Paul White
Institute of Sound and Vibration Research, Faculty of Engineering and the Environment
University of Southampton, Southampton, SO17 1BJ, United Kingdom

Abstract—**The purpose of this study was to investigate speech intelligibility for noise reduction algorithms which were integrated into perceptual wavelet packet-based speech coding strategy in cochlear implants (CIs). The algorithms of noise reduction including time-adaptive wavelet thresholding (TAWT) and time-frequency spectral subtraction (TFSS) were selected for this study due to simple and suitable for real-time implementation. The experiments were compared without and with noise reduction algorithms for fourteen normal-hearing (NH) listeners. The speech sentences were corrupted by babble noise in different signal-to-noise ratio (SNR) levels (0, 5 and 10 dB). The experimental results showed that the vocoded noisy speech with TAWT and TFSS provided higher intelligibility at 0 and 5 dB SNR but slightly lower intelligibility at 10 dB SNR when compared to vocoded noisy speech. CI listeners may benefit more than NH listeners in further study.**

Keywords- Cochlear implant; wavelet packet; spectral subtraction; wavelet thresholding; speech intelligibility

I. INTRODUCTION

A cochlear implant (CI) is an electronic prosthesis device implanted into the inner ear to restore partial hearing for profoundly hearing impaired persons by transmitting electric stimulation to auditory nerve. CIs are therefore designed to mimic the function of a normal cochlea. Speech coding strategy is an important part to improve the performance of cochlear devices [1] for effective communication. Speech coding strategies with temporal-envelope information have been developed that provide a higher level of speech-intelligibility than that of spectral-envelope information [2]. There are many speech coding strategies based on temporal envelopes, such as Continuous Interleaved Sampling (CIS), Spectral Peak (SPEAK) and Advanced Combination Encoder (ACE).

The majority of CI users can achieve high performance in speech intelligibility regardless of speech coding strategies they use, because almost all speech coding strategies perform well in quiet environments [3]. However, many CI users complain of severe degradation in speech understanding in noisy environments. Recently CI research effort has increasingly focused on state-of-the art noise reduction strategies to achieve higher speech intelligibility in noisy environments. Several noise reduction strategies have been proposed that use two or more microphones, or else a single microphone.

Multi-microphone noise reduction strategies can bring benefits to CI users. However, implants with two or more microphones are ergonomically difficult, and CI users may not like to wear headphones or a neck loop. Most CI users would find this a cosmetically unappealing prosthesis. Single-microphone noise reduction strategies are therefore more considerate and desirable. These strategies can be divided into two main categories [3]. The first of these is the preprocessing approach, where a noisy speech is processed with a noise reduction algorithm and then the enhanced speech is fed into the CI speech coding strategies. Another category is the noise reduction algorithm's integration into the CI speech coding strategies. This approach is combined to form one part of the speech processors to attenuate directly on noisy envelopes.

The integration of noise reduction algorithms into speech coding strategies has some advantages. This category of algorithm has a small latency caused by preprocessing techniques, low computational complexity and easy to integration into existing CI speech coding strategies. A number of noise reduction algorithms to be integrated into speech coding strategies have been proposed, such as a sigmoidal-shaped function [4], a principle components analysis (PCA) and independent component analysis (ICA) [5], statistical-model based algorithms based on noise estimation [6] and a sparse non-negative matrix factorisation [7].

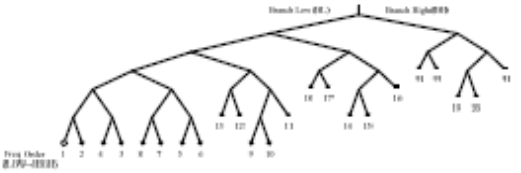
The wavelet packet transform (WPT) is a popular method for dividing the signal into auditory inspired frequency components to match a perceptual auditory scale such as the Bark scale. The perceptual wavelet packet-based speech coding strategy yields lower spectral leakage, better performance in terms of providing good frequency specificity [8], and better speech intelligibility performance than short-time Fourier transform (STFT)-based speech coding strategies for CI users [9]. Therefore, this paper presents noise reduction in perceptual wavelet packet-based speech coding strategies. The paper is organized into 5 sections. Section II describes the CI speech coding strategies based on perceptual WPT. Section III presents the principle of noise reductions including time-adaptive wavelet thresholding (TAWT) [10] and time-

978-1-4799-5838-2/14/$31.00 @2014 IEEE

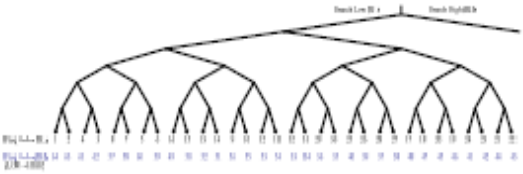
frequency spectral subtraction (TFSS) [11]. Section IV describes the procedures of performance evaluation in terms of speech intelligibility. The experimental results are given in section V. Finally, the conclusion and discussion are provided in section VI.

II. WAVELET PACKET-BASED SPEECH CODING STRATEGY

The stages of a wavelet packet-based speech coding strategy are similar to that of the ACE strategy, which uses an *n*-of-*m* channel selection strategy [12]. The signal is decomposed into *m* channels and only the *n* channels with highest energy are selected for stimulation. Two different structures of WPT [8] are shown in Fig. 1 that is used in this paper. Both are generated from a six level decomposition of wavelet packet. The 23-band WPT is designed to directly approximate the critical bands of the human auditory system using the 22 channels available to the Nucleus-24 device. Consequently, the lowest frequency band shown as a white subband in Fig. 1(a) is not used, because this frequency band plays no significant role in speech perception. The 64-band WPT is treated as the 64 FFT bins according to the Nucleus-24 device. The channels are grouped together to obtain 22 channels with different frequency ranges [12].



(a) 23-band WPT



(b) 64-band WPT

Figure 1. Structures of WPT

The block diagram of the wavelet packet-based speech coding strategy and noise reduction for CI simulation is shown in Fig. 2. The speech signal is recorded by a microphone at 16 kHz sampling rate and is first pre-emphasized by a filter that amplifies high-frequency components. The signal is processed, after pre-emphasis, frame-by-frame by using a sliding window of 128 samples (8 ms) with an overlap of 75%. The overlapping of window is the same as in ACE and it is adapted to the channel stimulation rates in the speech processor program, which is called the MAP. The higher the channel stimulation rates, the more the overlapping of window. The MAP contains a set of parameters that is different for individual CI users such as the channel stimulation rate, threshold and comfort levels.

The signal in each frame is then decomposed using the WPT into different frequency subbands. The power in each band is computed by using the average sum-square of the wavelet coefficients as details in section III. In the 64-band WPT strategy, the 64 frequency bands are computed by summing up power of consecutive frequency-bands with frequency ranges used in the Nucleus device to generate 22 channels. The power per band is weighted following as the ACE strategy. The envelopes are smoothed with a low-pass filter. A maximum of 12 envelopes (12-of-22 channel) is selected and used to modulate white noise, which is filtered by the bandpass filter as the same channels of WPT. A vocoded speech signal is synthesized by summing the modulated signals of each channel.

III. NOISE REDUCTION ALGORITHMS BASED ON ENVELOPE GAIN FUNCTION

Assume that noisy speech $y(n)$ is composed of clean speech $x(n)$ and the additive noise $d(n)$. Then:

$$y(n) = x(n) + d(n) \tag{1}$$

Taking the WPT of both sides gives:

$$Y_{j,n}(k) = X_{j,n}(k) + D_{j,n}(k) \tag{2}$$

where $Y_{j,n}(k)$, $X_{j,n}(k)$ and $D_{j,n}(k)$ are wavelet coefficients of the n^{th} subband at level j for noisy speech, clean speech and noise, respectively. k is the coefficient index in each subband.

Each frame is calculated using WPT. The number of wavelet coefficients in each subband depending on the decomposition level j, is K_j=128/2^j. In a single frame, the energy of each subband can be calculated using the average sum-square of the wavelet coefficients, thus:

$$E_y(i,n) = \frac{1}{K_j}\sum_k \left|Y_{j,n}(k)\right|^2 \tag{3}$$

This stage provides the time-frequency (T-F) envelope amplitude matrix, which represents the number of frames and channels. From (2), the T-F envelope amplitude matrix at the i^{th} frame and n^{th} channel (subband) can be defined as:

$$Y(i,n) = X(i,n) + D(i,n) \tag{4}$$

where $Y(i,n)$, $X(i,n)$ and $D(i,n)$ are the T-F envelope amplitudes matrix for the noisy speech, clean speech and noise, respectively and n=0,1,2,…,N-1 channels (N=22). Noise reduction algorithms are processed in the T-F envelope amplitude matrix as in Fig. 2. The T-F envelope amplitude matrix is modified by multiplying with a gain function (i.e. TAWT and TFSS) to control noise reduction across a wide range of SNR levels.

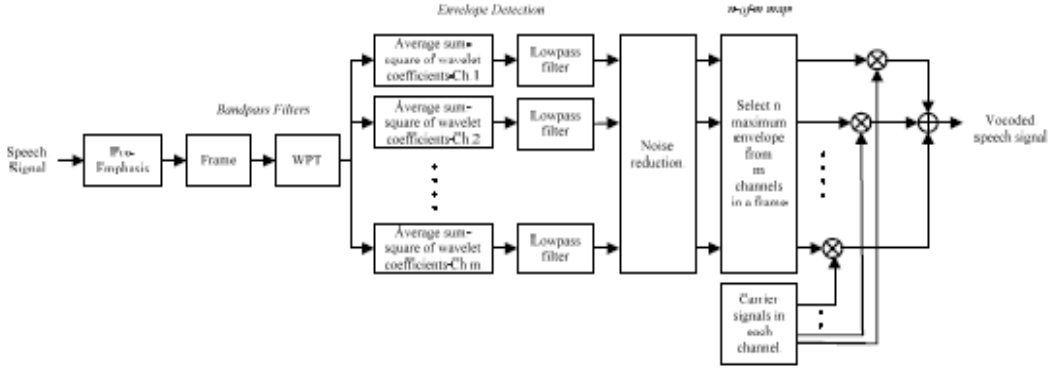


Figure 2. Block diagram of noise reduction in a wavelet packet-based speech coding strategy for CI simulation.

A. *Time-adaptive wavelet thresholding*

The time-adaptive wavelet thresholding (TAWT) algorithm [10] is different from conventional wavelet thresholding [13]. This technique is based on the Teager energy operator (TEO) and the adaptation of the wavelet threshold.

The TEO was modelled by Teager and was further investigated by Kaiser [14]. The TEO is a simple nonlinear function and a very local property of the signal, dependent on the three adjacent samples of the signal with indexes $i-1$, i, and $i+1$. The TAWT algorithm is computed in the following steps. The TEO coefficients $T(i,n)$ can be calculated from samples of three adjacent amplitude envelopes as:

$$T(i,n) = Y^2(i,n) - Y(i+1,n)Y(i-1,n) \tag{5}$$

where $Y(i,n)$ is the T-F envelope amplitude matrix of the noisy speech at the i^{th} frame and the n^{th} channel. The temporal masking $M(i,n)$ is constructed by smoothing the TEO coefficients, defined by:

$$M(i,n) = T(i,n) * h(i,n) \tag{6}$$

where * denotes the convolution operation and $h(i,n)$ is the lowpass filter. The adaptive threshold values $\lambda(i,n)$ are constructed from the temporal masking $M(i,n)$. If $M(i,n)$ below the variance of $M(i,n)$ is set to zero, otherwise temporal masking $M(i,n)$ is normalised as follows:

$$M'(i,n) = \begin{cases} \left[\dfrac{M(i,n)}{\max(M(i,n))}\right], & M(i,n) > \mathrm{var}(M(i,n)) \\ 0, & otherwise \end{cases} \tag{7}$$

The parameter of $M'(i,n)$ is close to 1 for speech regions and close to 0 for noise regions. Therefore the adaptive threshold values $\lambda(i,n)$ can be expressed as:

$$\lambda(i,n) = \lambda_n(1 - M'(i,n)) \tag{8}$$

$$\lambda_n = \sigma_n\sqrt{2\log(N\log_2(N))} \text{ and } \sigma_n = MAD_n / 0.6745 \tag{9}$$

where λ_n represents the channel-dependent threshold values, N is the total frames, σ_n is the noise variances with the median of the absolute deviation (MAD_n) of all the wavelet coefficients $Y(i,n)$ at the n^{th} channel, and 0.6745 is a normalisation factor, which is approximated from fine-scale wavelet coefficients. The enhanced speech $\hat{X}(i,n)$ is modified by the soft thresholding gain function as:

$$\hat{X}(i,n) = \mathrm{sgn}(Y(i,n))\max(|Y(i,n) - \lambda(i,n)|, 0) \tag{10}$$

B. *Time-frequency spectral subtraction*

The spectral subtraction is well-known techniques for speech enhancement. This technique is based on a simple implementation where enhanced speech is obtained by subtracting the noise estimation from the noisy speech. The spectral subtraction proposed by [11] is applied in this study. The gain function of this approach can be created using the relationship between the difference of phases and trigonometric principles as detail in [11]. The gain function, dependent on the estimation of *priori* SNR ξ and *posteriori* SNR $\hat{\gamma}$ parameters, can be expressed as follows:

$$G\left(\hat{\xi},\hat{\gamma}\right)=\sqrt{\left(1-\frac{(\hat{\gamma}+1-\hat{\xi})^2}{4\hat{\gamma}}\right)\Big/\left(1-\frac{(\hat{\gamma}-1-\hat{\xi})^2}{4\hat{\xi}}\right)} \quad (11)$$

The parameters $\hat{\xi}$ and $\hat{\gamma}$ in the gain function $G\left(\hat{\xi},\hat{\gamma}\right)$ are estimated according to:

$$\hat{\gamma}(i,n)=\beta\cdot\hat{\gamma}(i-1,n)+(1-\beta)\cdot\min\left(\hat{\gamma}_I(i,n),20\right) \quad (12)$$

$$\hat{\xi}(i,n)=\alpha\cdot\hat{\xi}_I(i-1,n)+(1-\alpha)\cdot\left(\sqrt{\hat{\gamma}(i,n)}-1\right)^2 \quad (13)$$

$$\hat{\xi}_I(i,n)\triangleq\frac{\hat{X}^2(i,n)}{\hat{D}^2(i,n)} \text{ and } \hat{\gamma}_I(i,n)\triangleq\frac{Y^2(i,n)}{\hat{D}^2(i,n)} \quad (14)$$

where the subscript I indicates the instantaneous values. β and α are weighting factors, which were set to $\beta=0.60$ and $\alpha=0.98$. Both factors control the trade-off between the noise reduction and the speech distortion.

This gain function is employed in the time-frequency spectral subtraction (TFSS) according to the following steps. Initially, an estimate of the noise power spectrum $\hat{D}^2(i,n)$ is averaged from the first five frames. Then $\hat{D}^2(i,n)$ is updated by a noise estimation algorithm [15], which is obtained by the minimum tracking method, since the power spectrum of the noisy speech regularly decays to the noise power level. This method tracks minimum values of a smoothed power spectrum for the noisy speech and multiplies by a constant to compensate for the bias noise estimate. This method has been found to work well for nonstationary environments.

Finally, the T-F envelope amplitude matrix of the enhanced speech is computed by a multiplication of the gain function $G\left(\hat{\xi},\hat{\gamma}\right)$ with the T-F envelope amplitude matrix of the noisy envelopes:

$$\hat{X}(i,n)=G\left(\hat{\xi},\hat{\gamma}\right)\cdot Y(i,n) \quad (15)$$

The *posteriori* SNR $\hat{\gamma}(i,n)$ in (12) is weighted to reduce rapid fluctuations and also to limit the over-suppression of the signal for large values of $\hat{\gamma}(i,n)$. The weighting factor β of $\hat{\gamma}(i,n)$ can improve the estimate of the enhanced speech. The *priori* SNR $\hat{\xi}(i,n)$ in (13) is weighted to control the average of spectral information positioned on past and present frames.

IV. Evaluation of Speech Intelligibility

The performance evaluations reported in this paper are conducted in a speech intelligibility experiment using NH listeners providing subjective information. To evaluate the performance of selected algorithms we conducted a psychophysical experiment using a speech perception paradigm with the BKB (Bamford-Kowal-Bench) sentences test [16].

A. Subjects

Fourteen NH listeners participated in this experiment. All subjects were native speakers of British English (8 males, 6 females, from 18 to 24 years of age) and had normal hearing thresholds (≤ 20 dB HL). They were staff and students at the University of Southampton and were paid for their participation.

B. Speech stimuli

The BKB test consists of 21 lists with each list consisting of 16 sentences (21 lists × 16 sentences = 336 sentences) and 50 key words (3–4 words per sentence). The sentences are composed of no more than seven syllables and their vocabulary reflects the natural language usage of younger and more impaired children. All the BKB sentences were recorded by a male speaker of standard British English at a 22 kHz sampling rate. They were resampled to 16 kHz for the experiment to simulate the speech processing in a CI system.

All sentences were processed separately offline using a wavelet packet-based strategy with Sym8 in noisy conditions. They were corrupted by babble noise at 0, 5 and 10 dB SNR, which are SNR levels where CI users can benefit. The noisy sentences were also processed using two algorithms for noise reduction (i.e. TAWT and TFSS). There were a total of 18 conditions (3 algorithms × 3 SNR levels × 2 wavelet packet structures).

C. Procedure

The experiment was carried out in a sound-treated room. A pure tone audiogram test was carried out to confirm that the subjects had normal hearing thresholds (≤ 20 dB HL, between 250 and 8000 Hz). The speech stimuli were presented using a Dell Latitude E4300 laptop, routed through a Creek Audio OBH-21SE headphone amplifier and presented unilaterally through a Sennheiser HDA280 circumaural headphone. Levels of speech stimuli in all experiments were presented at a comfortable conversional level (65 dB (A)).

Subjects were fully tested in a total of 18 conditions over two sessions on separate days, lasting approximately 1.5 hours each. They used their preferred ear (left or right) that was most comfortable for them to listen to the vocoded speech for the entire test. They were asked to write down the sentences that they heard. In the training session, they were asked to listen to one sentence list in both quiet and noisy conditions in a 5-minute test in order to familiarise themselves with the vocoded speech and the testing procedures. This sentence list was not included in the actual testing.

In the testing session, two lists of BKB sentences (32 sentences) per condition were used to provide one hundred keywords (100 percent). The sentences were scored in terms of the percentage of correct key words per condition,

expressed as "percent correct." No list was repeated across the conditions in each session. The order of conditions and the list-to-condition mapping in each session was randomised across subjects. Subjects were given a 5-minute break every 30 minutes during the test, or whenever they needed to take a rest.

V. Experimental Results

Fig. 3 shows electric stimulation patterns (electrodograms), derived using the 12-of-22 strategy of the BKB sentence "*The clown had a funny face*". For all the electrodograms, the y-axis represents the electrode position corresponding to a specific frequency band and the x-axis represents time progression. This figure shows the electrodogram of the clean speech and the noisy speech with/without noise reduction algorithms at 5 dB SNR babble noise. It can be seen that TAWT and TFSS remove noise and some details of the vocoded clean speech. This may affect to improve speech intelligibility.

Fig. 4 shows mean percentage correct scores for the two algorithms of noise reduction in noisy conditions. A three-way analysis of variance (ANOVA) with repeated measures was conducted with three main factors: algorithms, SNR levels, and wavelet packet structures. The results revealed a nonsignificant main effect of algorithms (F [2, 26] =1.391, p=0.267), a significant main effect of SNR levels (F [2, 26] =77.338, p<0.0005) and a significant main effect of wavelet packet structures (F [1, 13] =8.604, p=0.012). There was a significant interaction between algorithms and SNR levels (F [4, 52] =5.158, p<0.0005). However, there was a nonsignificant interaction between algorithms and wavelet packet structures (F [2, 26] =0.709, p=0.501), a nonsignificant interaction between SNR levels and wavelet packet structures (F [2, 26] =0.550, p=0.583), and a nonsignificant interaction between algorithms, SNR levels, and wavelet packet structures (F [4, 52] =0.172, p=0.952).

Post-hoc tests (Bonferroni) were used to consider the pair relationships among SNR levels, wavelet packet structures and between algorithms and SNR levels. The mean scores depended on the SNR levels. The higher SNR levels provided higher scores and vice versa. The 64-band WPT yielded slightly higher scores than the 23-band WPT in almost all conditions. The TAWT and TFSS provided significantly higher scores at 0 dB SNR and lower scores at 10 dB SNR when compared to the vocoded noisy speech at those SNR levels.

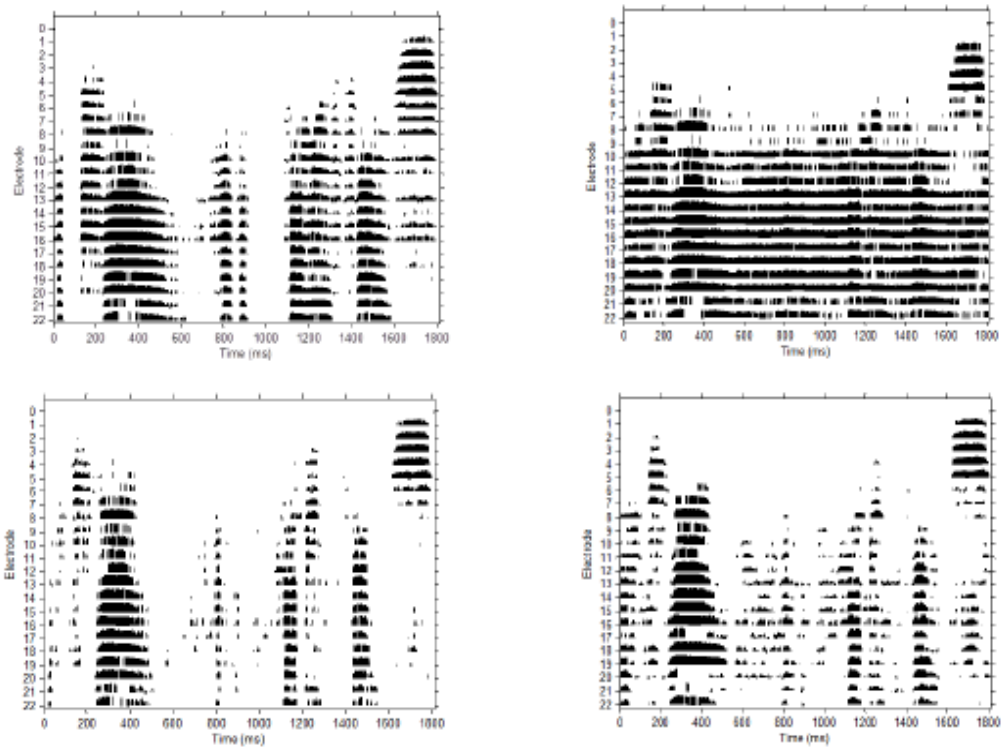


Figure 3. Electrodograms of the BKB sentence "*The clown had a funny face*" for noise reduction algorithms. (Left top panel) Clean speech. (Right top panel) Noisy speech with babble noise at 5 dB SNR. (Left bottom panel) Time-adaptive wavelet thresholding . (Right bottom panel) Time-frequency spectral subtraction.

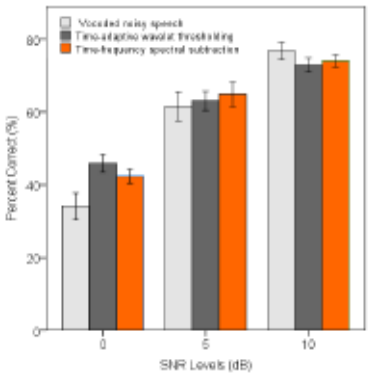


Figure 4. Mean percentage correct scores for two noise reduction algorithms at 0, 5, 10 dB SNR babble noise. The error bars indicate ± 1 standard error

VI. Conclusion and Discussion

Noise reduction algorithms including TAWT and TFSS were investigated when speech is corrupted by babble noise at 0, 5 and 10 dB SNR. Both TAWT and TFSS showed a significant improvement at 0 dB SNR, no significant improvements at 5 dB SNR and significantly worse in speech intelligibility at 10 dB SNR when compared to vocoded noisy speech.

This may result from noise estimation in noise reduction algorithms. Noise estimation in both TAWT and TFSS may be under- or overestimated and this results in distortion in the enhanced speech. The distortion of the enhanced speech may be more than the noise reduction. The distortion of the enhanced speech may affect the speech intelligibility performance because speech discrimination becomes more difficult. If techniques of noise reduction can improve estimation of noise levels, this may lead to improve speech intelligibility.

In addition, NH listeners are more sensitive to speech distortion and less sensitive to noise when compared to hearing-impaired (HI) listeners [17]. NH listeners can reach better intelligibility performance at higher SNR levels without noise reduction algorithms (i.e. 10 dB SNR). Therefore, noise reduction algorithms may work well for CI users, but not work for NH listeners at higher SNR levels.

Acknowledgment

The authors would like to thank all participants for their valuable time to doing this research.

References

[1] P. C. Loizou, "Mimicking the human ear," *IEEE Signal Processing Magazine*, vol. 15, pp. 101-130, 1998.

[2] Z. Fan-Gang, S. Rebscher, W. Harrison, S. Xiaoan, and F. Hailong, "Cochlear implants: system design, integration, and evaluation," *IEEE Reviews in Biomedical Engineering*, pp. 115-142, 2008.

[3] K. Kokkinakis, B. Azimi, Y. Hu, and D. R. Friedland, "Single and Multiple Microphone Noise Reduction Strategies in Cochlear Implants," *Trends Amplif*, vol. 16, pp. 102-116, Jun 2012.

[4] Y. Hu, P. C. Loizou, N. Li, and K. Kasturi, "Use of a sigmoidal-shaped function for noise attenuation in cochlear implants," *Journal of the Acoustical Society of America*, vol. 122, pp. EL128-EL134, Oct 2007.

[5] G. Li, "Speech perception in a sparse domain," PhD Thesis, Institute of Sound and Vibration Research, University of Southampton, 2008.

[6] P. W. Dawson, S. J. Mauger, and A. A. Hersbach, "Clinical evaluation of signal-to-noise ratio-based noise reduction in Nucleus(R) cochlear implant recipients," *Ear Hear.*, vol. 32, pp. 382-390, 2011.

[7] H. M. Hu, N. Mohammadiha, J. Taghia, A. Leijon, M. E. Lutman, and S. Y. Wang, "Sparsity Level in a Non-Negative Matrix Factorization Based Speech Strategy in Cochlear Implants," *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pp. 2432-2436, 2012.

[8] V. Gopalakrishna, N. Kehtarnavaz, and P. C. Loizou, "A Recursive Wavelet-Based Strategy for Real-Time Cochlear Implant Speech Processing on PDA Platforms," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 2053-2063, 2010.

[9] W. Nogueira, A. Giese, B. Edler, and A. Buchner, "Wavelet packet filterbank for speech processing strategies in cochlear implants," *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2006.

[10] S. H. Chen and J. F. Wang, "Speech enhancement using perceptual wavelet packet decomposition and teager energy operator," *Journal of VLSI Signal Processing Systems for Signal Image and Video Technology*, vol. 36, pp. 125-139, 2004.

[11] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, vol. 50, pp. 453-466, Jun 2008.

[12] ACE and CIS DSP Strategies, Software Requirements Specification,Cochlear Corporation, Lane Cove, New South Wales, Australia, 2002.

[13] D. L. Donoho and I. M. Johnstone, "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, vol. 81, pp. 425-455, 1994.

[14] J. Kaiser, "Some useful properties of Teager's energy operators," *ICASSP-93.1993 IEEE International Conference on Acoustics, Speech, and Signal Processing (Cat.No.92CH3252-4)*, pp. 149-152, 1993.

[15] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 504-512, 2001.

[16] J. Bench, A. Kowal, and J. Bamford, "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," *British Journal of Audiology*, vol. 13, pp. 108-12, 1979-Aug 1979.

[17] N. H. van Schijndel, T. Houtgast, and J. M. Festen, "Effects of degradation of intensity, time, or frequency content on speech intelligibility for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 110, pp. 529-542, Jul 2001.

Appendix B: Objective speech intelligibility

B.1 The normalised covariance metric (NCM)

The NCM (Jianfen et al., 2009; Chen, 2011) is calculated as follows (Figure B.1). The stimuli are first decomposed into *N* bands across the signal bandwidth (125-8000 Hz in this study) using Butterworth filters. The envelope of each band is computed using the Hilbert transform and then down-sampled to $2f_{cut}$ Hz, thereby limiting the envelope modulation rate to f_{cut} Hz (200 Hz in this study).

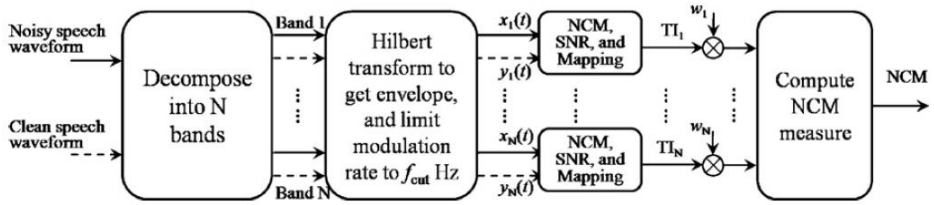


Figure B.1 Computation of NCM measure (Chen, 2011)

Let $x_i(t)$ and $y_i(t)$ be the down-sampled envelope of the clean and noisy speech signals in the i^{th} bands, respectively. The normalised covariance in the i^{th} bands is computed as:

$$r_i = \frac{\sum_t \left(x_i(t) - \mu_i\right)\left(y_i(t) - v_i\right)}{\sqrt{\sum_t \left(x_i(t) - \mu_i\right)^2}\sqrt{\sum_t \left(y_i(t) - v_i\right)^2}} \tag{B.1}$$

where μ_i and v_i are the mean value of $x_i(t)$ and $y_i(t)$, respectively. A value of r_i close to 0 indicates that the clean and noisy speech is uncorrelated, while the value of r_i close to 1 indicates that the clean and noisy speech is related. The signal-to-noise ratio (*SNR*) in each band is defined as:

$$SNR_i = 10\log_{10}\left(\frac{r_i^2}{1 - r_i^2}\right) \tag{B.2}$$

and subsequently limited to the SNR dynamic range of [-15, 15] dB (as done in the computation of SII measure (ANSI, 1997)). The transmission index (TI) in each band

is computed by linearly mapping the SNR values between 0 and 1 using the following equation:

$$TI_i = \frac{SNR_i + 15}{30} \tag{B.3}$$

Finally, the transmission indices are averaged across all frequency bands to produce the NCM index:

$$NCM = \frac{\sum_{i=1}^{N} TI_i \times w_i}{\sum_{i=1}^{N} w_i} \tag{B.4}$$

where the weights w_i are often called the band-importance functions in the computation of the speech intelligibility index (SII) measure. There are several methods for selecting this weight, but the most common weights are the ANSI articulation index (AI) weights as shown in Table 1 (ANSI, 1997). The NCM measure is always limited to the range of [0, 1].

Table B.1 The ANSI AI weights used in the implementation of the NCM (Chen, 2011)

Band	Centre frequency (Hz)	Weight
1	151.3	0.0835
2	208.8	0.0990
3	276.7	0.0913
4	356.9	0.0708
5	451.7	0.0600
6	563.7	0.0493
7	696.0	0.0440
8	852.4	0.0441
9	1037.2	0.0490
10	1255.5	0.0486
11	1513.5	0.0493
12	1818.4	0.0496
13	2178.6	0.0548
14	2604.2	0.0548
15	3107.2	0.0488
16	3701.5	0.0366
17	4403.7	0.0380
18	5233.5	0.0320
19	6214.0	0.0246
20	7372.5	0.0208

B.2 The short-time objective intelligibility measure (STOI)

The STOI (Taal et al., 2011) is a time-frequency intermediate intelligibility measure (Figure B.2) based on a correlation coefficient between the temporal envelopes of the clean and degraded speech in the short-time region. First, the clean and degraded speech are processed in each frame with a length of 25.6 msec, performed by a Hann-window with 50% overlap. Next, the windowed signals are decomposed into DFT-based one-third octave bands. These bands are performed by grouping DFT-bins into 15 one-third octave bands with the lowest and highest frequency band at 150 Hz and 4.3 kHz, respectively.

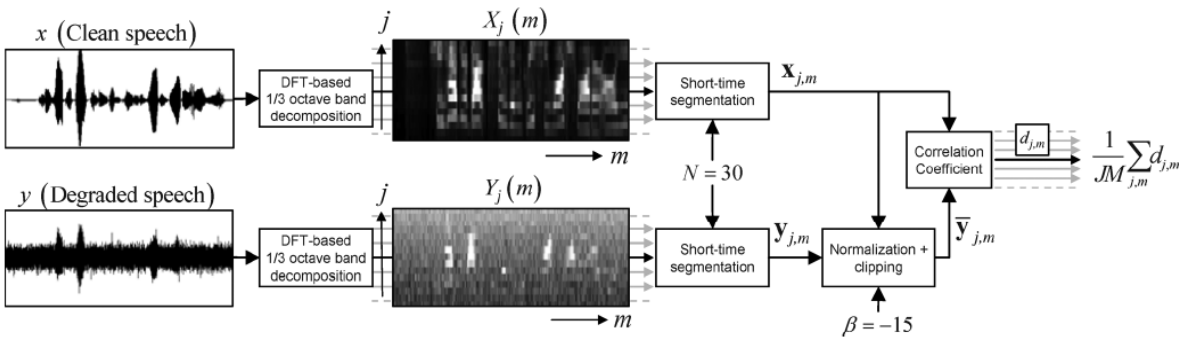


Figure B.2 The computation of the STOI measure (Taal et al., 2011)

Let $\hat{x}(k,m)$ denote the k^{th} DFT-bin of the m^{th} frame of the clean speech. The norm of the j^{th} one-third octave band, referred to as a TF-unit, is then defined as:

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} \left|\hat{x}(k,m)\right|^2} \tag{B.5}$$

where k_1 and k_2 denote the one-third octave band edge, which are rounded to the nearest DFT-bin. The T-Frepresentation of the degraded speech is obtained similarly, and is denoted by $Y_j(m)$. Let $\mathbf{x}_{j,m}$ denote the short-time temporal envelope of the clean speech:

$$\mathbf{x}_{j.m} = \left[X_j(m-N+1), X_j(m-N+2), ..., X_j(m)\right]^T \tag{B.6}$$

where N = 30 which equals an analysis length of 384 msec. Similarly, $\mathbf{y}_{j,m}$ denotes the short-time temporal envelope of the degraded speech, which is normalised and clipped before comparison.

The rationale behind the normalization procedure is to compensate for global level differences which should not have a strong effect on speech intelligibility. The clipping procedure makes sure that the sensitivity of the model towards one TF-unit which is severely degraded is upper bounded. Let $\mathbf{x}(n)$ denote the nth element of $\mathbf{x}$, where $n \in \{1,...,N\}$ and $\|\cdot\|$ represent the l_2 norm. Let $\bar{\mathbf{y}}_{j,m}$ denote the normalised and clipped version of $\mathbf{y}$. Then:

$$\bar{\mathbf{y}}_{j,m}(n) = \min\left(\frac{\|\mathbf{x}_{j,m}\|}{\|\mathbf{y}_{j,m}\|}\mathbf{y}_{j,m}(n), (1+10^{-\beta/20})\mathbf{x}_{j,m}(n)\right) \tag{B.7}$$

where $\beta = -15$ dB refers to the lower signal-to-distortion (SDR) model. SDR is given by:

$$SDR = 10\log_{10}\left(\frac{\mathbf{x}_{j,m}(n)^2}{(\bar{\mathbf{y}}_{j,m}(n) - \mathbf{x}_{j,m}(n))^2}\right) \geq \beta \tag{B.8}$$

The intermediate intelligibility measure is defined as the sample correlation coefficient $d_{j,m}$ between the two vectors, where $\mu_{(\cdot)}$ refers to the sample average of the corresponding vector. Finally, the average of the intermediate intelligibility measure overall bands and frames is computed

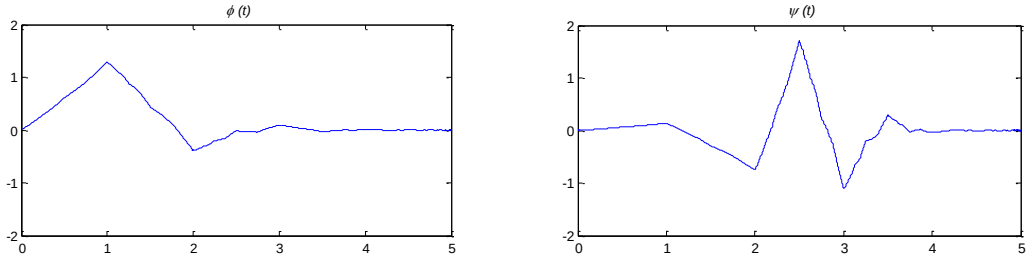
$$d_{j,m} = \frac{(\mathbf{x}_{j,m} - \mu_{x_{j,m}})^T(\bar{\mathbf{y}}_{j,m} - \mu_{\bar{y}_{j,m}})}{\|\mathbf{x}_{j,m} - \mu_{x_{j,m}}\| \ \|\bar{\mathbf{y}}_{j,m} - \mu_{\bar{y}_{j,m}}\|} \tag{B.9}$$

$$d = \frac{1}{JM}\sum_{j,m} d_{j,m} \tag{B.10}$$

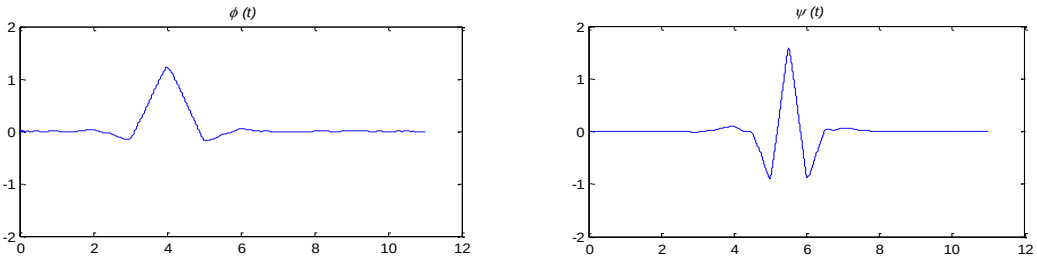
where M represents the total number of frames and J the number of one-third octave band.

Appendix C: Mother wavelets

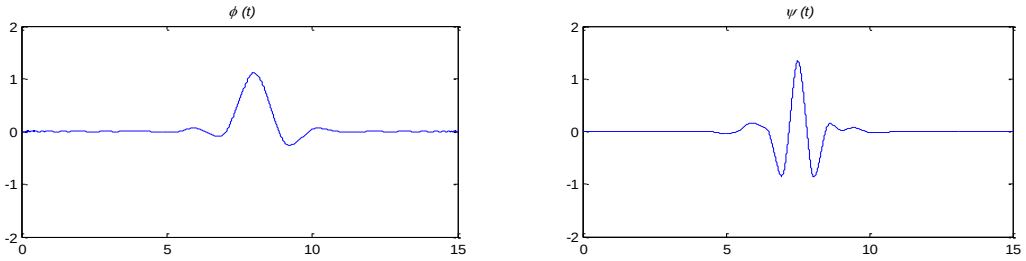
MATLAB Wavelet Toolbox provides a number of mother wavelets with order *N* for WPT, including Haar (*harr*), Daubechies (*dbN*), Symlets (sym*N*), Coiflet (*coifN*), biorthogonal (*biorN*), reverse biorthogonal (*rbioN*) and discrete Meyer (*dmey*). The filter length *L* of mother wavelets is 2*N*, except for Coiflet which is 6*N*. The wavelet functions and the coefficients of wavelet filters for db3, coif2 and sym8 are illustrated in Figure C.1 and C.2, respectively. The example of filter coefficients for decomposition and reconstruction are shown in Table C.1.



(a) db3

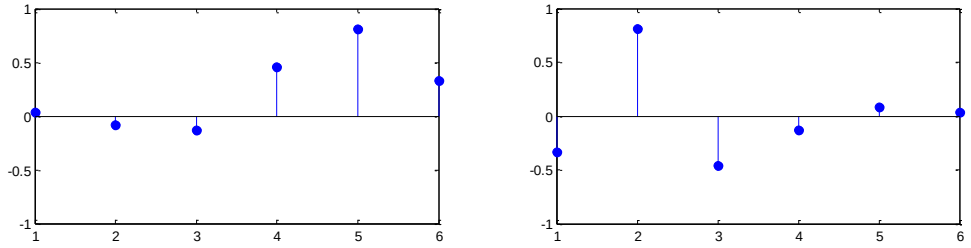


(b) coif2

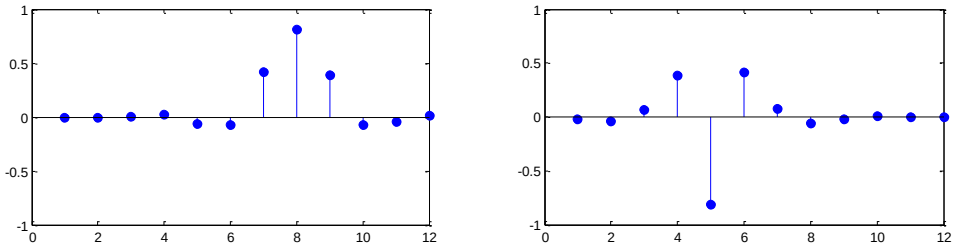


(c) sym8

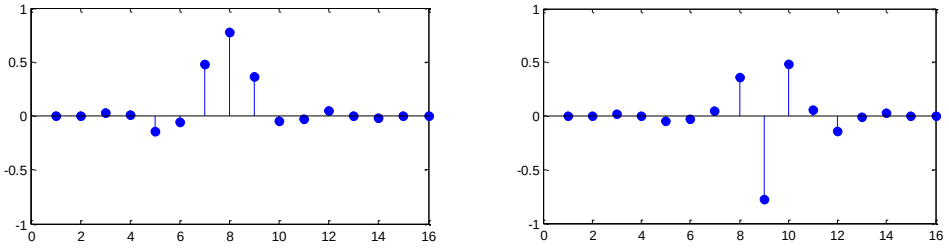
Figure C.1 Example of the scaling functions $\phi(t)$ (left) and wavelet functions $\psi(t)$ (right) with order *N*.



(a) db3



(b) coif2



(c) sym8

Figure C.2 Example of coefficients of lowpass filters (left) and highpass filter (right)

Table C.1 Filter coefficients for decomposition and reconstruction

dB3	Decomposition		Reconstruction	
n	Lowpass filter $h(n)$	Highpass filter $g(n)$	Lowpass filter $\tilde{h}(n)$	Highpass filter $\tilde{g}(n)$
0	0.0352	-0.3327	0.3327	0.0352
1	-0.0854	0.8069	0.8069	0.0854
2	-0.1350	-0.4599	0.4599	-0.1350
3	0.4599	-0.1350	-0.1350	-0.4599
4	0.8069	0.0854	-0.0854	0.8069
5	0.3327	0.0352	0.0352	-0.3327

Table C.1 (Continued) Filter coefficients for decomposition and reconstruction

Coif2	Decomposition		Reconstruction	
n	Lowpass filter $h(n)$	Highpass filter $g(n)$	Lowpass filter $\tilde{h}(n)$	Highpass filter $\tilde{g}(n)$
0	-0.0007	-0.0164	0.0164	-0.0007
1	-0.0018	-0.0415	-0.0415	0.0018
2	0.0056	0.0674	-0.0674	0.0056
3	0.0237	0.3861	0.3861	-0.0237
4	-0.0594	-0.8127	0.8127	-0.0594
5	-0.0765	0.4170	0.4170	0.0765
6	0.4170	0.0765	-0.0765	0.4170
7	0.8127	-0.0594	-0.0594	-0.8127
8	0.3861	-0.0237	0.0237	0.3861
9	-0.0674	0.0056	0.0056	0.0674
10	-0.0415	0.0018	-0.0018	-0.0415
11	0.0164	-0.0007	-0.0007	-0.0164

Sym8	Decomposition		Reconstruction	
n	Lowpass filter $h(n)$	Highpass filter $g(n)$	Lowpass filter $\tilde{h}(n)$	Highpass filter $\tilde{g}(n)$
0	-0.0034	-0.0019	0.0019	-0.0034
1	-0.0005	-0.0003	-0.0003	0.0005
2	0.0317	0.0150	-0.0150	0.0317
3	0.0076	0.0038	0.0038	-0.0076
4	-0.1433	-0.0491	0.0491	-0.1433
5	-0.0613	-0.0272	-0.0272	0.0613
6	0.4814	0.0519	-0.0519	0.4814
7	0.7772	0.3644	0.3644	-0.7772
8	0.3644	-0.7772	0.7772	0.3644
9	-0.0519	0.4814	0.4814	0.0519
10	-0.0272	0.0613	-0.0613	-0.0272
11	0.0491	-0.1433	-0.1433	-0.0491
12	0.0038	-0.0076	0.0076	0.0038
13	-0.0150	0.0317	0.0317	0.0150
14	-0.0003	0.0005	-0.0005	-0.0003
15	0.0019	-0.0034	-0.0034	-0.0019

Appendix D: Speech processors

D.1 Design parameters for cochlear implant devices

Table D.1 shows the detail specific parameters for currently available clinical cochlear implant devices from the three major manufacturers (Fan-Gang et al., 2008).

Table D.1 Parameters for cochlear implant devices from the three major manufactures.

	Current electrode arrays					Experimental electrode arrays		
Parameters	Advanced Bionics HiFocus 1J	Advanced Bionics Helix	Cochlear Contour Advance	Med-El Combi 40+	Med-El FlexSoft	Cochlear Hybrid	Cochlear Hybrid-L	Med-El FlexEAS
Active Length	17mm	13.25mm	15.5mm	26.4mm	26.4mm	6mm	15mm	20.9mm
Total Length	20mm	20mm	25mm	31.5mm	31.5mm	6mm/10mm	16mm	25mm
Carrier Material	Silicon rubber (LSR-70)		Silicon rubber (LSR-30)	Silicon rubber (LSR-40)		Silicon rubber (LSR 30)		Silicon rubber (LSR-40)
Carrier Diameter (base to tip)	0.8-0.4mm	1.16-0.66mm 1.2-0.7mm	0.8-0.5mm	0.8x0.78mm at base 0.58x0.48mm at apex			0.35x0.25 mm at tip	0.8x0.78mm at base 0.58x0.35mm at apex
Number of Electrodes	16	16	22	12 pairs	7 basal pairs + 5 apical singles	6	22	7 basal pairs + 5 apical singles
Spacing	1.1mm	0.85mm	0.75mm	2.4mm	2.4mm	0.75mm	0.75mm	1.9mm
Shape	Straight	Pre-curved	Pre-curved	Straight	Straight	Straight	Straight	Straight
Stylet	No	Yes	Yes	No	No	No	No	No

D.2 Example of speech processor programs (MAP)

```
                map_name: 'ACE'
       audio_sample_rate: 16000
       channel_stim_rate: 500
           analysis_rate: 500
             block_shift: 32
               num_bands: 22
            num_selected: 12
          interval_length: 1
       implant_stim_rate: 6000
                 implant: [1x1 struct]
            phase_width: 25
              phase_gap: 8
                 period: 166.6000
          rf_frame_width: 62
               processes: {8x1 cell}
wav_sample_rate_tolerance: 1.0500
            block_length: 128
                  window: [128x1 double]
              buffer_opt: []
           window_length: 128
                num_bins: 65
                bin_freq: 125
               bin_freqs: [1x65 double]
              char_freqs: [22x1 double]
             sample_rate: 500
                equalise: 1
               band_bins: [22x1 double]
                 weights: [22x65 double]
            power_gains: [22x1 double]
        crossover_freqs: [23x1 double]
            band_widths: [22x1 double]
               gains_dB: 0
             base_level: 0.0156
              sat_level: 0.5859

                      Q: 20
                sub_mag: -1.0000e-010
      lgf_dynamic_range: 31.4806
              lgf_alpha: 416.2063
    channel_order_type: 'base-to-apex'
         channel_order: [22x1 double]
            electrodes: [22x1 double]
                 modes: 103
          special_idle: 1
      threshold_levels: [22x1 double]
        comfort_levels: [22x1 double]
            full_scale: 1
                volume: 100
           volume_type: 'standard'
```

D.3 ACE Strategy

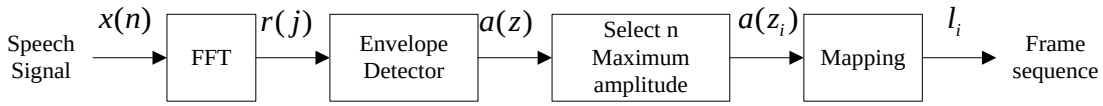


Figure D.1 Block diagram of ACE strategy

The ACE strategy is used in the Nucleus-24 processor made by the Cochlear Corporation, and a basic block diagram (Cochlear, 2002; Nogueira et al., 2005) is shown in Figure D.1. The speech signal at 16 kHz sampling rate is processed in each frame (8 ms and L=128 samples) and then performed by hanning window with overlap depending on the parameters of the channel stimulation rate in CI user's MAP. The windowed signal is transformed by FFT. The windowed function used is:

$$w(j) = 0.5\left(1.0 - \cos\left(\frac{2\pi j}{L}\right)\right), \qquad j = 0,1,2,...,L-1 \tag{D.1}$$

The 128-point FFT provides 128 spectral coefficients or 128 bins. Due to the symmetry properties of FFT, the first 64 bins are then used and the second 64 bins are discarded without loss of information. The 64 FFT bins with linear spacing are rearranged to mimic the critical band of the auditory system by summing the powers of adjacent bins to provide m channels (typically 20 or 22) with different frequency ranges. The frequency range in each channel is defined by the frequency table of Cochlear Corporation. Generally, the apical one-third of the channels are allocated with linear spacing to frequencies up to 1 kHz, while the basal two-thirds of the channels are allocated with logarithm spacing to frequencies above 1 kHz. The real part of the j^{th} FFT bin is denoted by $x(j)$ and the imaginary part by $y(j)$. The power of the bin is:

$$r^2(j) = x^2(j) + y^2(j), \qquad j = 0,1,...,L/2 \tag{D.2}$$

The power of the envelope of channel z is calculated as a weighted sum of the FFT bin powers. Where $g_z(j)$ are set to the gain g_z for a specific number of bins and otherwise zeros, the envelope of channel z is:

$$a(z) = \sqrt{\sum_{j=0}^{L/2} g_z(j) r^2(j)}, \quad z = 1,...,M \tag{D.4}$$

The envelope channels $a(z_i)$ with the largest amplitude are selected for stimulation. The mapping block is done by using the loudness growth function (LGF), which is a logarithmically-shaped function that maps the acoustic envelope amplitude $a(z_i)$ to an electrical magnitude

$$p(z_i) = \begin{cases} \dfrac{\log(1+\rho(a(z_i)-s)/(m-s))}{\log(1+\rho)}, & s \le a(z_i) \le m, \\ 0, & a(z_i) < s, \\ 1, & a(z_i) \ge m, \end{cases} \tag{D.5}$$

$$l_i = T + (C-T)p_i \tag{D.6}$$

The magnitude $p(z_i)$ is a fraction in the range 0 to 1 that represents the proportion of the output range (from the threshold *T* to the comfort level *C*). An input at the base-level *s* is mapped to an output at threshold level, and no output is produced for an input of lower amplitude. The parameter *m* is the input level at which the output saturates; inputs at this level or above result in stimuli at comfort level. If there are less than *N* envelopes above base level, they are mapped to the threshold level. The parameter ρ controls the steepness of the LGF. Finally, the channels z_i are stimulated sequentially with a stimulation order from high to low frequencies (base-to-apex) with levels.

Appendix E:

E.1 A geometric approach to power spectral subtraction

Let $y(n) = x(n) + d(n)$ be the noisy speech $y(n)$ consisting of the clean speech $x(n)$ and noise $d(n)$. Taking the short-time Fourier transform (STFT) of both sides gives:

$$Y(\omega) = X(\omega) + D(\omega) \tag{E.1}$$

Equation (E.1) is multiplied by its conjugate $Y^*(\omega)$. This can be written as follows:

$$|Y(\omega)|^2 = |X(\omega)|^2 + |D(\omega)|^2 + 2|X(\omega)||D(\omega)|\cos(\theta_X - \theta_D) \tag{E.2}$$

Equation (E.1) can be rewritten in polar form by its magnitude and phase as:

$$a_Y e^{j\theta_Y} = a_X e^{j\theta_X} + a_D e^{j\theta_D} \tag{E.3}$$

where $\{a_Y, a_X, a_D\}$ are the magnitude spectrums and $\{\theta_Y, \theta_X, \theta_D\}$ are the phases of the noisy speech, clean speech, and noise spectrum, respectively.

The noisy speech $Y(\omega)$ in Equation (E.1) can be represented geometrically in the complex plane as the sum of the clean speech $X(\omega)$ and noise $D(\omega)$ as in Figure E.1 (a). The cross term in Equation (E.2) conducts to the error of the noise estimate. If the phase difference between clean speech and noise $(\theta_X - \theta_D)$ is 90, then $|Y(\omega)|^2 = |X(\omega)|^2 + |D(\omega)|^2$. This cross term can lead to an underestimation (i.e. $(\theta_X - \theta_D) < 90$) and overestimation (i.e. $(\theta_X - \theta_D) > 90$) of noise in the power spectral subtraction.

The gain function G of the spectral subtraction can be generated from the triangle (Figure E.1 (b)) using the Sine Rule with $\overline{AB} \perp \overline{BC}$. The gain function G can be given by:

$$\overline{AB} = a_Y \sin(\theta_D - \theta_Y) = a_X \sin(\theta_D - \theta_X) \tag{E.4}$$

$$a_Y^2(1 - c_{YD}^2) = a_X^2(1 - c_{XD}^2) \tag{E.5}$$

$$G = \frac{a_X}{a_Y} = \sqrt{\frac{1-c_{YD}^2}{1-c_{XD}^2}} \tag{E.6}$$

where $c_{YD} \triangleq \cos(\theta_Y - \theta_D)$ and $c_{XD} \triangleq \cos(\theta_X - \theta_D)$. Since no methods accurately determine these phases (i.e. c_{YD} and c_{XD}), the explicit relationship between the phases can be represented using the trigonometric principle. Equation (E.2) can be rewritten in terms of the magnitude spectrums $\{a_Y, a_X, a_D\}$ as:

$$a_Y^2 = a_X^2 + a_D^2 + 2a_X a_D \cos(\theta_X - \theta_D) \tag{E.7}$$

The cosine rule for the triangle as in Figure E.1 (b) gives the following relationships as:

$$a_X^2 = a_Y^2 + a_D^2 - 2a_Y a_D \cos(\theta_Y - \theta_D) \tag{E.8}$$

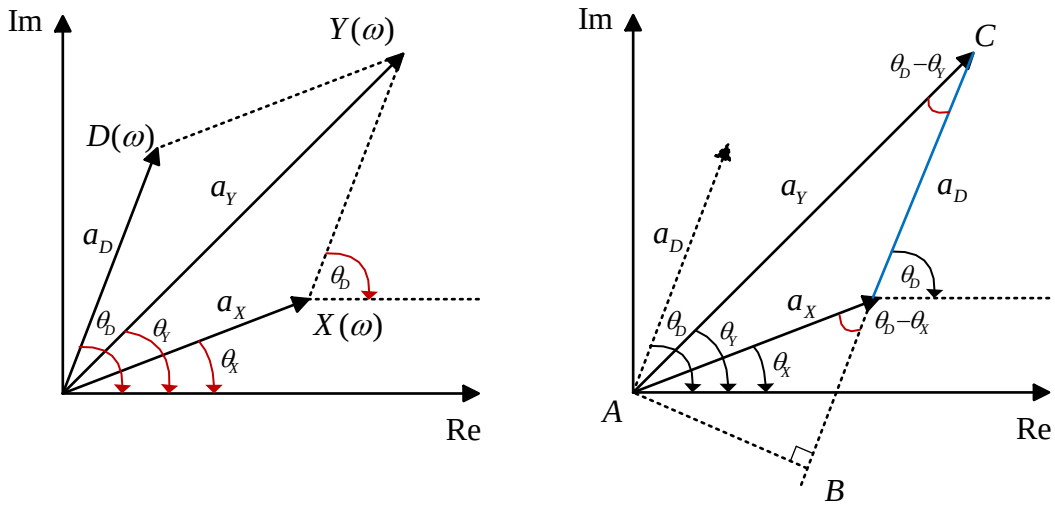
Dividing both sides of Equation (E.7) and (E.8) by a_D^2 and using the definitions of $\xi \triangleq a_X^2 / a_D^2$ and $\gamma \triangleq a_Y^2 / a_D^2$, then c_{YD} and c_{XD} can then be given by:

$$c_{XD} = \frac{a_Y^2 - a_D^2 - a_X^2}{2a_X a_D} = \frac{\gamma - 1 - \xi}{2\sqrt{\xi}} \tag{E.9}$$

$$c_{YD} = \frac{a_Y^2 + a_D^2 - a_X^2}{2a_Y a_D} = \frac{\gamma + 1 - \xi}{2\sqrt{\gamma}} \tag{E.10}$$

Then the gain function as in Equation (E.6) can be rewritten as:

$$G = \sqrt{\frac{1-c_{YD}^2}{1-c_{XD}^2}} = \sqrt{\left(1 - \frac{(\gamma+1-\xi)^2}{4\gamma}\right) \Big/ \left(1 - \frac{(\gamma-1-\xi)^2}{4\xi}\right)} \tag{E.11}$$



(a) (b)

Figure E.1 The geometric viewpoint of spectral subtraction in the complex plane. (a) represents noisy speech $Y(\omega)$ as the sum of clean speech $X(\omega)$ and noise $D(\omega)$. (b) represents the triangle of the geometric relationship between the phases of noisy/clean speech and noise. Adapted from Lu and Loizou (2008).

Appendix F:

F.1 Post-test questionnaire

Adapted from Dehaish et al. (2008)

QUESTIONS	CONDITIONS				
	C1	C2	C3	…	Cn
1. **Articulation:** Were the sounds distinguishable? (1) Yes, very clear (2) Yes, clear enough (3) Fairly clear (4) No, not very clear (5) No, not at all					
2. **Listening effort:** How would you describe the effort you were required to make in order to understand the message? (1) Complete relaxation possible; no effort required (2) Attention necessary; no appreciable effort required (3) Moderate effort required (4) Considerable effort required (5) No meaning understood with any feasible effort					
3. **Ease of listening:** Would it be easy to listen to this voice for long periods of time? (1) Very easy (2) Easy (3) Neutral (4) Difficult (5) Very difficult					
4. **Overall impression:** How do you rate the quality of the sound you just heard? (1) Excellent (2) Good (3) Fair (4) Poor (5) Bad					

F.2 The post-test questionnaire results

Table F.2 The post-test questionnaire results for noise reduction in wavelet packet-based speech coding strategy with different noise types.

Lists	Detail		Q1 - Articulation						Q2 - Listening effort					
			[1]	[2]	[3]	[4]	[5]		[1]	[2]	[3]	[4]	[5]	
C1	23-WPT	Quiet		●						●				
C2	64-WPT			●						●				
C3	23-WPT	BB				●					●			
C4	64-WPT					●					●			
C5	23-WPT	SS				●					●			
C6	64-WPT					●					●			
C7	23-WPT	BB-IdBM		●						●				
C8	64-WPT			●						●				
C9	23-WPT	BB-TAWT				●					●			
C10	64-WPT					●					●			
C11	23-WPT	BB-TFSS			●						●			
C12	64-WPT				●						●			
C13	23-WPT	SS-IdBM		●						●				
C14	64-WPT			●						●				
C15	23-WPT	SS-TAWT				●					●			
C16	64-WPT					●					●			
C17	23-WPT	SS-TFSS			●						●			
C18	64-WPT				●						●			

Note: the abbreviations used in table are as follows: C-condition, Q-Question, BB-Babble noise, SS-Speech-shaped noise, IdBM-Ideal binary masking, TAWT-Time-adaptive wavelet thresholding and TFSS-Time-frequency spectral subtraction.

Table F.2 (Continued) The post-test questionnaire results for noise reduction in wavelet packet-based speech coding strategy with different noise types.

Lists	**Detail**		Q3 - Ease of listening						Q4 - Overall impression					
			[1]	**[2]**	**[3]**	**[4]**	**[5]**		**[1]**	**[2]**	**[3]**	**[4]**	**[5]**	
C1	23-WPT	Quiet		●						●				
C2	64-WPT			●						●				
C3	23-WPT	BB				●						●		
C4	64-WPT					●						●		
C5	23-WPT	SS				●						●		
C6	64-WPT					●						●		
C7	23-WPT	BB-IdBM		●						●				
C8	64-WPT			●						●				
C9	23-WPT	BB-TAWT				●						●		
C10	64-WPT					●						●		
C11	23-WPT	BB-TFSS				●						●		
C12	64-WPT					●						●		
C13	23-WPT	SS-IdBM		●						●				
C14	64-WPT			●						●				
C15	23-WPT	SS-TAWT				●						●		
C16	64-WPT					●						●		
C17	23-WPT	SS-TFSS				●						●		
C18	64-WPT					●						●		

Note: the abbreviations used in table are as follows: C-condition, Q-Question, BB-Babble noise, SS-Speech-shaped noise, IdBM-Ideal binary masking, TAWT-Time-adaptive wavelet thresholding and TFSS-Time-frequency spectral subtraction.

F.3 The post-test questionnaire results

Table F.3 The post-test questionnaire results for noise reduction in wavelet packet-based CI processors with different SNR levels.

Lists	Detail		Q1 - Articulation						Q2 - Listening effort					
			[1]	[2]	[3]	[4]	[5]		[1]	[2]	[3]	[4]	[5]	
C1	23-WPT	0-BB					●					●		
C2	64-WPT						●					●		
C3	23-WPT	5-BB			●						●			
C4	64-WPT				●						●			
C5	23-WPT	10-BB			●						●			
C6	64-WPT				●						●			
C7	23-WPT	0-TAWT				●						●		
C8	64-WPT					●						●		
C9	23-WPT	0- TFSS				●						●		
C10	64-WPT					●						●		
C11	23-WPT	5-TAWT			●							●		
C12	64-WPT				●							●		
C13	23-WPT	5-TFSS			●						●			
C14	64-WPT				●						●			
C15	23-WPT	10-TAWT			●							●		
C16	64-WPT				●						●			
C17	23-WPT	10-TFSS			●						●			
C18	64-WPT				●						●			

Note: the abbreviations used in table are as follows: C-condition, Q-Question, the different SNR levels (i.e. 0, 5 and 10 dB), BB-Babble noise, SS-Speech-shaped noise, TAWT-Time-adaptive wavelet thresholding and TFSS-Time-frequency spectral subtraction.

Table F.3 (Continued) The post-test questionnaire results for noise reduction in wavelet packet-based CI processors with different SNR levels.

Lists	**Detail**		Q3 - Ease of listening						Q4 - Overall impression					
			[1]	**[2]**	**[3]**	**[4]**	**[5]**		**[1]**	**[2]**	**[3]**	**[4]**	**[5]**	
C1	23-WPT	0-BB					●						●	
C2	64-WPT					●							●	
C3	23-WPT	5-BB				●						●		
C4	64-WPT					●						●		
C5	23-WPT	10-BB			●						●			
C6	64-WPT				●						●			
C7	23-WPT	0-TAWT				●						●		
C8	64-WPT					●						●		
C9	23-WPT	0- TFSS				●						●		
C10	64-WPT					●						●		
C11	23-WPT	5-TAWT				●						●		
C12	64-WPT					●						●		
C13	23-WPT	5-TFSS				●					●			
C14	64-WPT					●						●		
C15	23-WPT	10-TAWT				●						●		
C16	64-WPT				●							●		
C17	23-WPT	10-TFSS				●						●		
C18	64-WPT				●							●		

Note: the abbreviations used in table are as follows: C-condition, Q-Question, the different SNR levels (i.e. 0, 5 and 10 dB), BB-Babble noise, SS-Speech-shaped noise, TAWT-Time-adaptive wavelet thresholding and TFSS-Time-frequency spectral subtraction.

References

[1] Agbinya, J. I. (1996). "Discrete wavelet transform techniques in speech processing." 1996 IEEE Tencon - Digital Signal Processing Applications Proceedings, Vols 1 and 2**,** 514-519.

[2] Ansi (1997). "Methods for the Caculation of the Speech Intelligibility Index." American National Standard Institute, New York.

[3] Ashino, R., Mandai, T. and Morimoto, A. (2010). "Blind Source Separation of Spatio-Temporal Mixed Signals Using Phase Information of Analytic Wavelet Transform." International Journal of Wavelets Multiresolution and Information Processing, 8**,** 575-594.

[4] Bahoura, M. and Rouat, J. (2001). "Wavelet speech enhancement based on the Teager Energy operator." IEEE Signal Processing Letters, 8**,** 10-12.

[5] Bahoura, M. and Rouat, J. (2006). "Wavelet speech enhancement based on time-scale adaptation." Speech Communication, 48**,** 1620-1637.

[6] Baumann, U. and Nobbe, A. (2006). "The cochlear implant electrode-pitch function." Hear Res, 213**,** 34-42.

[7] Bayram, I. and Selesnick, I. W. (2008). "On the dual-tree complex wavelet packet and M-band transforms." Ieee Transactions on Signal Processing, 56**,** 2298-2310.

[8] Behrenbruch, C. P. and Lithgow, B. J. "SNR Improvement, Filtering and Spectral Equaliation in Cochlear Implants using Wavelet Techniques." 2nd International Conference on Bioelectromagnetism, 1998. 61-62.

[9] Bench, J., Kowal, A. and Bamford, J. (1979). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children." British Journal of Audiology, 13**,** 108-112.

[10] Berouti, M., Schwartz, R. and Makhoul, J. (1979). "Enhancement of speech corrupted by acoustic noise." ICASSP 79.1979 IEEE International Conference on Acoustics, Speech and Signal Processing**,** 208-211.

[11] Boll, S. F. (1979). "Suppression of Acoustic Noise in Speech Using Spectral Subtraction." Ieee Transactions on Acoustics Speech and Signal Processing, 27**,** 113-120.

[12] Buechner, A., Frohne-Buechner, C., Boyle, P., Battmer, R. D. and Lenarz, T. (2009). "A high rate n-of-m speech processing strategy for the first generation Clarion cochlear implant." International Journal of Audiology, 48**,** 868-875.

[13] Carnero, B. and Drygajlo, A. (1999). "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms." IEEE Transactions on Signal Processing, 47**,** 1622-1635.

[14] Chang, S., Kwon, Y. ; Sang-Il Yang ; Kim, I-Jae. "Speech enhancement for non stationary noise environment by adaptive wavelet packet." IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002.

References

[15] Cheikhrouhou, I., Atitallah, R. B., Ouni, K., Hmida, A. B., Mamoudi, N. and Ellouze, N. "Speech Analysis using Wavelet Transforms Dedicated to Cochlear Prosthesis Stimulation Strategy." ISCCSP 2004, 2004. IEEE, 639-642.

[16] Chen, F. (2011). "The relative importance of temporal envelope information for intelligibility prediction: A study on cochlear-implant vocoded speech." Med.Eng Phys., 33**,** 1033-1038.

[17] Chen, F. and Loizou, P. C. (2010). "Contribution of Consonant Landmarks to Speech Recognition in Simulated Acoustic-Electric Hearing." Ear and Hearing, 31**,** 259-267.

[18] Chen, F. and Loizou, P. C. (2011). "Predicting the Intelligibility of Vocoded Speech." Ear and Hearing, 32**,** 331-338.

[19] Chen, J. D., Benesty, J., Huang, Y. and Doclo, S. (2006). "New insights into the noise reduction Wiener filter." IEEE Transactions on Audio Speech and Language Processing, 14**,** 1218-1234.

[20] Chen, S. H. and Wang, J. F. (2004). "Speech enhancement using perceptual wavelet packet decomposition and teager energy operator." Journal of Vlsi Signal Processing Systems for Signal Image and Video Technology, 36**,** 125-139.

[21] Chen, S. H., Wu, H. T., Chang, Y. K. and Truong, T. K. (2007). "Robust voice activity detection using perceptual wavelet-packet transform and Teager energy operator." Pattern Recognition Letters, 28**,** 1327-1332.

[22] Christiansen, C., Pedersen, M. S. and Dau, T. (2010). "Prediction of speech intelligibility based on an auditory preprocessing model." Speech Communication, 52**,** 678-692.

[23] Cochlear (2002). "ACE™ and CIS DSP Strategies." *Software Requirements Specification,Part Number: N95287F Issue 1.* Lane Cove, New South Walse, Australia.

[24] Cochlear (2007). "Selection stimulation rate with the Nucleus freedom system." White paper prepared by Cochlear Ltd.

[25] Cohen, L. "Enhancement of speech using Bark-scaled wavelet packet decomposition." 7th European conference speech, communication and technology, EUROSPEECH-2001, 2001 2001 Aalborg, Denmark. 1933-1936.

[26] Coifman, R. R., Mayer, Y. and Wickerhauser, M. V. (1992). "Wavelet Analysis and Signal Processing " *In Wavelets and Their Applications.* Jones and Barlett.

[27] Daubechies, I. (1992). *Ten lectures on wavelets,* Montpelier, Vermont, USA, Capital city press.

[28] Dawson, P. W., Mauger, S. J. and Hersbach, A. A. (2011). "Clinical evaluation of signal-to-noise ratio-based noise reduction in Nucleus(R) cochlear implant recipients." Ear Hear., 32**,** 382-390.

[29] De Melo, T. M., Bevilacqua, M. C. and Costa, O. A. (2012). "Speech perception in cochlear implant users with the HiRes 120 strategy: a systematic review." Brazilian Journal of Otorhinolaryngology, 78**,** 129-133.

[30] Dehaish, S., Wang, S. and Brinton, J. (2008). *Performance evaluation of speech processor for cochlear implant.* MSc, University of Southampton.

[31] Derbel, A., Kallel, F., Samet, M. and Hamida, A. B. (2008). "Bionic Wavelet Transform based on Speech Processing dedicated to a Fulliy Programmable Stimulation Strategy for Cochlear Prostheses." Asian Journal of Scienfic Research, 1**,** 293-309.

[32] Dimitriadis, D., Maragos, P. and Potamianos, A. (2011). "On the Effects of Filterbank Design and Energy Computation on Robust Speech Recognition." IEEE Transactions on Audio Speech and Language Processing, 19**,** 1504-1516.

[33] Donoho, D. L. (1995). "De-Noising by Soft-Thresholding." Ieee Transactions on Information Theory, 41**,** 613-627.

[34] Donoho, D. L. and Johnstone, I. M. (1994). "Ideal Spatial Adaptation by Wavelet Shrinkage." Biometrika, 81**,** 425-455.

[35] Dorman, M. F., Loizou, P. C., Fitzke, J. and Tu, Z. M. (1998). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels." Journal of the Acoustical Society of America, 104**,** 3583-3585.

[36] Dorman, M. F., Loizou, P. C. and Rainey, D. (1997). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs." Journal of the Acoustical Society of America, 102**,** 2403-2411.

[37] Dorman, M. F., Loizou, P. C., Spahr, A. J. and Maloff, E. (2002). "A comparison of the speech understanding provided by acoustic models of fixed-channel and channel-picking signal processors for cochlear implants." Journal of Speech Language and Hearing Research, 45**,** -.

[38] Drake, L. A., Rutledge, J. C. and Cohen, J. (1993). "Wavelet Analysis in Recruitment of Loudness Compensation." IEEE Transactions on Signal Processing, 41**,** 3306-3312.

[39] Ephraim, Y. and Cohen, I. (2004). "Recent advancements in speech enhancement." CRC Electronic Engineering Handbook**,** -.

[40] Ephraim, Y. and Malah, D. (1984). "Speech Enhancement Using A Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator." Ieee Transactions on Acoustics Speech and Signal Processing, 32**,** 1109-1121.

[41] Ephraim, Y. and Malah, D. (1985). "Speech Enhancement Using A Minimum Mean-Square Error Log-Spectral Amplitude Estimator." Ieee Transactions on Acoustics Speech and Signal Processing, 33**,** 443-445.

[42] Erkelens, J. S., Hendriks, R. C., Heusdens, R. and Jensen, J. (2007). "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors." IEEE Transactions on Audio Speech and Language Processing, 15**,** 1741-1752.

[43] Evans, N. W. D., Mason, J. S. D., Liu, W. M. and Fauve, B. (2006). "An assessment on the fundamental limitations of spectral subtraction." 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, Vols 1-13**,** 145-148.

[44] Fan-Gang, Z., Rebscher, S., Harrison, W., Xiaoan, S. and Haihong, F. (2008). "Cochlear implants: system design, integration, and evaluation." IEEE Reviews in Biomedical Engineering**,** 115-142.

[45] Fernandes, F. C. A., Van Spaendonck, R. L. C. and Burrus, C. S. (2003). "A new framework for complex wavelet transforms." IEEE Transactions on Signal Processing, 51**,** 1825-1837.

[46] Festen, J. M. (1987). "Speech-Reception Threshold in a Fluctuating Background Sound and its Possible Relation to Temporal Auditory Resolution." The Psychophysics of Speech Perception, 39**,** 461-466.

[47] Festen, J. M. and Plomp, R. (1990). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing." J Acoust Soc Am, 88**,** 1725-1736.

[48] Firszt, J. B., Holden, L. K., Skinner, M. W., Tobey, E. A., Peterson, A., Gaggl, W., Runge-Samuelson, C. L. and Wackym, P. A. (2004). "Recognition of speech presented at soft to loud levels by adult cochlear implant recipients of three cochlear implant systems." Ear and Hearing, 25**,** 375-387.

[49] Fourakis, M. S., Hawks, J. W., Holden, L. K., Skinner, M. W. and Holden, T. A. (2004). "Effect of frequency boundary assignment on vowel recognition with the Nucleus 24 ACE speech coding strategy." J Am Acad Audiol, 15**,** 281-299.

[50] Fourakis, M. S., Hawks, J. W., Holden, L. K., Skinner, M. W. and Holden, T. A. (2007). "Effect of frequency boundary assignment on speech recognition with the Nucleus 24 ACE speech coding strategy." Journal of the American Academy of Audiology, 18**,** 700-717.

[51] French, N. R. and Steinberg, J. C. (1947). "Factors Governing the Intelligibility of Speech Sounds." Journal of the Acoustical Society of America, 19**,** 90-119.

[52] Fu, Q. J. and Nogaki, G. (2005). "Noise susceptibility of cochlear implant users: the role of spectral resolution and smearing." J Assoc Res Otolaryngol, 6**,** 19-27.

[53] Fu, Q. J. and Shannon, R. V. (1998). "Effects of amplitude nonlinearity on phoneme recognition by cochlear implant users and normal-hearing listeners." J Acoust Soc Am, 104**,** 2570-2577.

[54] Fu, Q. J., Shannon, R. V. and Wang, X. (1998). "Effects of noise and spectral resolution on vowel and consonant recognition: acoustic and electric hearing." Journal of the Acoustical Society of America, 104**,** 3586-3596.

[55] Fugal, D. L. (2009). *Conceptual Wavelets in Digital Signal Processing*, Space & Signals Technologies LLC.

[56] Ghanbari, Y. and Karami-Mollaei, M. R. (2006). "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets." Speech Communication, 48**,** 927-940.

[57] Glasberg, B. R. and Moore, B. C. J. (1990). "Derivation of Auditory Filter Shapes from Notched-Noise Data." Hearing Research, 47**,** 103-138.

[58] Goldsworthy, R. L. and Greenberg, J. E. (2004). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations." Journal of the Acoustical Society of America, 116**,** 3679-3689.

[59] Gopalakrishna, V., Kehtarnavaz, N. and Loizou, P. (2010a). "Real-Time Implementation of Wavelet-Based Advanced Combination Encoder on Pda Platforms for Cochlear Implant Studies." 2010 Ieee International Conference on Acoustics, Speech, and Signal Processing**,** 1670-1673.

[60] Gopalakrishna, V., Kehtarnavaz, N. and Loizou, P. C. (2010b). "A Recursive Wavelet-Based Strategy for Real-Time Cochlear Implant Speech Processing on PDA Platforms." IEEE Transactions on Biomedical Engineering, 57**,** 2053-2063.

[61] Greenwood, D. D. (1990). "A Cochlear Frequency-Position Function for Several Species - 29 Years Later." Journal of the Acoustical Society of America, 87**,** 2592-2605.

[62] Guan, T., Yu, S. and Ye, D. (2005). *Application of wavelet in speech processing of cochlear implant*.

[63] Gustafsson, S., Jax, P. and Vary, P. (1998). "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics." Proceedings of the 1998 Ieee International Conference on Acoustics, Speech and Signal Processing, Vols 1-6**,** 397-400.

[64] Harma, A., Karjalainen, M., Savioja, L., Valimaki, V., Laine, U. K. and Huopaniemi, J. (2000). "Frequency-warped signal processing for audio applications." Journal of the Audio Engineering Society, 48**,** 1011-1031.

[65] Hermes, D. J. and Vangestel, J. C. (1991). "The Frequency Scale of Speech Intonation." Journal of the Acoustical Society of America, 90**,** 97-102.

[66] Hillenbrand, J., Getty, L. A., Clark, M. J. and Wheeler, K. (1995). "Acoustic Characteristics of American English Vowels." Journal of the Acoustical Society of America, 97**,** 3099-3111.

[67] Houtgast, T. and Steeneken, H. J. M. (1985). "A Review of the Mtf Concept in Room Acoustics and Its Use for Estimating Speech-Intelligibility in Auditoria." Journal of the Acoustical Society of America, 77**,** 1069-1077.

[68] Hu, H., Lutman, M. E., Mohammadiha, N., Sang, J., Bleeck, S. and Wang, S. (2013). "Sparse non-negative matrix fatoization strategy for cochlear implants." Transactions on Biomedical Engineering.

[69] Hu, Y. and Loizou, P. C. (2002). "A subspace approach for enhancing speech corrupted by colored noise." IEEE Signal Processing Letters, 9**,** 204-206.

[70] Hu, Y. and Loizou, P. C. (2004). "Speech enhancement based on wavelet thresholding the multitaper spectrum." IEEE Transactions on Speech and Audio Processing, 12**,** 59-67.

[71] Hu, Y. and Loizou, P. C. (2007). "A comparative intelligibility study of single-microphone noise reduction algorithms." Journal of the Acoustical Society of America, 122**,** 1777-1786.

[72] Hu, Y. and Loizou, P. C. (2008). "A new sound coding strategy for suppressing noise in cochlear implants." Journal of the Acoustical Society of America, 124**,** 498-509.

[73] Hu, Y., Loizou, P. C., Li, N. and Kasturi, K. (2007). "Use of a sigmoidal-shaped function for noise attenuation in cochlear implants." Journal of the Acoustical Society of America, 122**,** El128-El134.

[74] Itu-T (2000). "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codes." ITU-T Recommendation P.862.

[75] Jianfen, M., Yi, H. and Loizou, P. (2009). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions." Journal of the Acoustical Society of America, 125**,** 3387-3405.

[76] Kadambe, S. and Boudreauxbartels, G. F. (1992). "Application of the Wavelet Transform for Pitch Detection of Speech Signals." IEEE Transactions on Information Theory, 38**,** 917-924.

[77] Kaiser, J. (1993). "Some useful properties of Teager's energy operators." ICASSP-93.1993 IEEE International Conference on Acoustics, Speech, and Signal Processing (Cat.No.92CH3252-4)**,** 149-152.

[78] Kallel, F., Frikha, M., Ghorbel, M., Ben Hamida, A. and Berger-Vachon, C. (2012). "Dual-channel spectral subtraction algorithms based speech enhancement dedicated to a bilateral cochlear implant." Applied Acoustics, 73**,** 12-20.

[79] Kamath, S. and Loizou, P. (2002). "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise." 2002 Ieee International Conference on Acoustics, Speech, and Signal Processing, Vols I-Iv, Proceedings**,** 4164-4164.

[80] Karmakar, A., Kumar, A. and Patney, R. K. (2011). "Synthesis of an Optimal Wavelet Based on Auditory Perception Criterion." Eurasip Journal on Advances in Signal Processing**,** -.

[81] Kasturi, K. and Loizou, P. C. (2007). "Effect of filter spacing on melody recognition: acoustic and electric hearing." Journal of the Acoustical Society of America, 122**,** 29-34.

[82] Kates, J. M. and Arehart, K. H. (2005). "Coherence and the speech intelligibility index." Journal of the Acoustical Society of America, 117**,** 2224-2237.

[83] Kiefer, J., Hohl, S., Sturzebecher, E., Pfennigdorff, T. and Gstoettner, W. (2001). "Comparison of speech recognition with different speech coding strategies (SPEAK, CIS, and ACE) and their relationship to telemetric measures of compound action potentials in the nucleus CI 24M cochlear implant system." Audiology, 40**,** 32-42.

[84] Kingsbury, N. (2001). "Complex wavelets for shift invariant analysis and filtering of signals." Applied and Computational Harmonic Analysis, 10**,** 234-253.

[85] Kokkinakis, K., Azimi, B., Hu, Y. and Friedland, D. R. (2012). "Single and Multiple Microphone Noise Reduction Strategies in Cochlear Implants." Trends Amplif, 16**,** 102-116.

[86] Kokkinakis, K., Hazrati, O. and Loizou, P. C. (2011). "A channel-selection criterion for suppressing reverberation in cochlear implants." Journal of the Acoustical Society of America, 129**,** 3221-3232.

[87] Li, G. (2008). *Speech perception in a sparse domain.* PhD Thesis, Institute of Sound and Vibration Research, University of Southampton.

[88] Li, J. F., Yang, L., Zhang, J. P., Yan, Y. H., Hu, Y., Akagi, M. and Loizou, P. C. (2011). "Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English." Journal of the Acoustical Society of America, 129**,** 3291-3301.

[89] Li, M., Mcallister, H. G., Black, N. D. and De Perez, T. A. (2000). "Wavelet-based nonlinear AGC method for hearing aid loudness compensation." IEE Proceedings-Vision Image and Signal Processing, 147**,** 502-507.

[90] Li, M., Mcallister, H. G., Black, N. D. and De Perez, T. A. (2001). "Perceptual time-frequency subtraction algorithm for noise reduction in hearing aids." IEEE Transactions on Biomedical Engineering, 48**,** 979-988.

[91] Li, N. (2009). *Contribution of acoustic landmarks to speech recognition in noise by cochlear implant users.* PhD Dissertation, Electrical engineering, University of Texas.

[92] Litvin, Y. and Cohen, I. (2011). "Single-Channel Source Separation of Audio Signals Using Bark Scale Wavelet Packet Decomposition." Journal of Signal Processing Systems for Signal Image and Video Technology, 65**,** 339-350.

[93] Lockwood, P. and Boudy, J. (1992). "Experiments with a Nonlinear Spectral Subtractor (Nss), Hidden Markov-Models and the Projection, for Robust Speech Recognition in Cars." Speech Communication, 11**,** 215-228.

[94] Loizou, P. and Gibak, K. (2011). "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions." IEEE Transactions on Audio, Speech and Language Processing**,** 47-56.

[95] Loizou, P. C. (1998). "Mimicking the human ear." IEEE Signal Processing Magazine, 15**,** 101-130.

[96] Loizou, P. C. (1999). "Signal-processing techniques for cochlear implants - A review of progress in deriving electrical stimuli from the speech signal." IEEE Engineering in Medicine and Biology Magazine, 18**,** 34-46.

[97] Loizou, P. C. (2006). "Speech processing in vocoder-centric cochlear implants." *In:* MOLLER, A. R. (ed.) *Cochlear and Brainstem Implants.* Basel, Switzerland: Karger, 109-143.

[98] Loizou, P. C. (2007). *Speech Enhancement: Theory and Practice,* Boca Raton, FL, CRC Press, LLC.

[99] Loizou, P. C., Cohen, I., Gannot, S. and Paliwal, K. (2007). "Special issue on speech enhancement." Speech Communication, 49**,** 527-529.

[100] Loizou, P. C., Lobo, A. and Hu, Y. (2005). "Subspace algorithms for noise reduction in cochlear implants." J.Acoust.Soc.Am., 118**,** 2791-2793.

[101] Loizou, P. C., Poroy, O. and Dorman, M. (2000). "The effect of parametric variations of cochlear implant processors on speech understanding." Journal of the Acoustical Society of America, 108**,** 790-802.

[102] Lu, Y. and Loizou, P. C. (2008). "A geometric approach to spectral subtraction." Speech Communication, 50**,** 453-466.

[103] Ma, J. F. and Loizou, P. C. (2011). "SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech." Speech Communication, 53**,** 340-354.

[104] Mallat, S. (2009). *A wavelet tour of signal processing: the sparse way*, United States, Elsevier Inc.

[105] Martin, R. (2001). "Noise power spectral density estimation based on optimal smoothing and minimum statistics." IEEE Transactions on Speech and Audio Processing, 9**,** 504-512.

[106] Marzinzik, M. and Kollmeier, B. (2002). "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics." IEEE Transactions on Speech and Audio Processing, 10**,** 109-118.

[107] Mcaulay, R. J. and Malpass, M. L. (1980). "Speech Enhancement Using a Soft-Decision Noise Suppression Filter." IEEE Transactions on Acoustics Speech and Signal Processing, 28**,** 137-145.

[108] Mcdermott, H. J. (1998). "How cochlear implants have expanded our understanding of speech perception." IEEE Engineering in Medicine and Biology Magazine, 20**,** 2251-2256.

[109] Mckay, C. M. and Henshall, K. R. (2002). "Frequency-to-electrode allocation and speech perception with cochlear implants." Journal of the Acoustical Society of America, 111**,** 1036-1044.

[110] Meyer, Y. (1993). *Wavelets: algorithms and applications*, Pennsylvania, USA, the society for industrial and applied mathematics.

[111] Miller, J. D. (1989). "Auditory-Perceptual Interpretation of the Vowel." Journal of the Acoustical Society of America, 85**,** 2114-2134.

[112] Moon, I. J. and Hong, S. H. (2014). "What Is Temporal Fine Structure and Why Is It Important?" The Korean Audiological Society, 18**,** 7.

[113] Moore, B. C. J. (2008). "Basic auditory processes involved in the analysis of speech sounds." Philosophical Transactions of the Royal Society B-Biological Sciences, 363**,** 947-963.

[114] Mourad Ghrissi, A. C. (2012). "Comparison of IIR Filterbanks and FFT Filterbanks in Cochlear Implant Speech Processing Strategies." J.Electrical Systems**,** 76-84.

[115] Nidcd (2014). "Cochlear Implants." publication no. 09-4798 [online]. http://www.nidcd.nih.gov/health/hearing/pages/coch.aspx: National Institutes of Health.

[116] Nie, K., Lan, N. and Gao, S. (1998). "Implementation of CIS speech processing strategy for cochlear implants by using wavelet transform." Proceeding of ICSP 98**,** 1395-1398.

[117] Nogueira, W., Buchner, A., Lenarz, T. and Edler, B. (2005). "A psychoacoustic "NofM"-type speech coding strategy for cochlear implants." Eurasip Journal on Applied Signal Processing, 2005**,** 3044-3059.

[118] Nogueira, W., Giese, A., Edler, B. and Buchner, A. (2006). "Wavelet packet filterbank for speech processing strategies in cochlear implants." IEEE International Conference on Acoustic, Speech, and Signal Processing**,** -.

[119] Paglialonga, A., Tognola, G., Baselli, G., Parazzini, A., Ravazzani, P. and Grandori, F. "Speech processing for cochlear implants with the discrete wavelet transform: feasibility study and performance evaluation." Proceeding of the 28th IEEE EMBS Annual International Conference, 2006. 3763-3766.

[120] Paglialonga, A., Tognola, G., Sibella, F., Parazzini, A., Ravazzani, P., Grandori, F. and Baselli, G. (2008). "Influence of cochlear implant-like operating conditions on wavelet speech processing." Computers in Biology and Medicine, 38**,** 799-804.

[121] Peng, Z. K. and Chu, F. L. (2004). "Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography." Mechanical Systems and Signal Processing, 18**,** 199-221.

[122] Pinter, I. (1996). "Perceptual wavelet-representation of speech signals and its application to speech enhancement." Computer Speech and Language, 10**,** 1-22.

[123] Qazi, O. U. R., Van Dijk, B., Moonen, M. and Wouters, J. (2012). "Speech Understanding Performance of Cochlear Implant Subjects Using Time-Frequency Masking-Based Noise Reduction." Ieee Transactions on Biomedical Engineering, 59**,** 1364-1373.

[124] Qin, M. K. and Oxenham, A. J. (2003). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers." Journal of the Acoustical Society of America, 114**,** 446-454.

[125] Reyes, N. R., Zurera, M. R., Ferreras, F. L. and Amores, P. J. (2003). "Adaptive wavelet-packet analysis for audio coding purposes." Signal Processing, 83**,** 919-929.

[126] Rhebergen, K. S. and Versfeld, N. J. (2005). "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners." Journal of the Acoustical Society of America, 117**,** 2181-2192.

[127] Sang, J. (2012). *Evaluation of the sparse coding shrinkage noise reduction algorithm for the hearing impaired.* PhD, University of Southampton.

[128] Sang, Y. F., Wang, D., Wu, J. C., Zhu, Q. P. and Wang, L. (2009). "Entropy-Based Wavelet De-noising Method for Time Series Analysis." Entropy, 11**,** 1123-1147.

[129] Santos, J. F., Cosentino, S., Hazrati, O., Loizou, P. C. and Falk, T. H. (2012). "Performance Comparison of Intrusive Objective Speech Intelligibility and Quality Metrics for Cochlear Implant Users." 13th Annual Conference of the International Speech Communication Association 2012 (Interspeech 2012), Vols 1-3**,** 1722-1725.

[130] Scalart, P. and Vieira, J. (1996). "Speech enhancement based on a priori signal to noise estimation." 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, Conference Proceedings, Vols 1-6**,** 629-632.

[131] Selesnick, I. W., Baraniuk, R. G. and Kingsbury, N. G. (2005). "The dual-tree complex wavelet transform." IEEE Signal Processing Magazine, 22**,** 123-151.

[132] Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J. and Ekelid, M. (1995). "Speech recognition with primarily temporal cues." Science, 270**,** 303-304.

[133] Shao, Y. and Chang, C. H. (2006). "A Kalman filter based on wavelet filter-bank and psychoacoustic modeling for speech enhancement." 2006 Ieee International Symposium on Circuits and Systems, Vols 1-11, Proceedings**,** 121-124.

[134] Shao, Y. and Chang, C. H. (2007). "A generalized time-frequency subtraction method for robust, speech enhancement based on wavelet filter banks modeling of human auditory system." IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 37**,** 877-889.

[135] Shao, Y. and Chang, C. H. (2011). "Bayesian Separation With Sparsity Promotion in Perceptual Wavelet Domain for Speech Enhancement and Hybrid Speech Recognition." IEEE Transactions on Systems Man and Cybernetics Part A-Systems and Humans, 41**,** 284-293.

[136] Sheikhzadeh, H. A., H. R. (2001). "An Improved wavelet-based speech enhancement system." *7th European Conference on Speech Communication and Technology*. Aalborg Congress and Culture Centre, Aalborg, Denmark.

[137] Sheldon, S., Pichora-Fuller, M. K. and Schneider, B. A. (2008). "Effect of age, presentation method, and learning on identification of noise-vocoded words." J Acoust Soc Am, 123**,** 476-488.

[138] Shukla, P. D. (2003). *COMPLEX WAVELET TRANSFORMS AND THEIR APPLICATIONS*. Master of Philosophy, University of Strathclyde.

[139] Simpson, S. A. and Cooke, M. (2005). "Consonant identification in N-talker babble is a nonmonotonic function of N." J Acoust Soc Am, 118**,** 2775-2778.

[140] Sinha, D. and Tewfik, A. H. (1993). "Low Bit Rate Transparent Audio Compression using Adapted Wavelets." IEEE Transactions on Signal Processing, 41**,** 3463-3479.

[141] Skinner, M. W., Holden, L. K. and Holden, T. A. (1995). "Effect of frequency boundary assignment on speech recognition with the speak speech-coding strategy." Ann Otol Rhinol Laryngol Suppl, 166**,** 307-311.

[142] Skinner, M. W., Holden, L. K. and Holden, T. A. (1997). "Parameter selection to optimize speech recognition with the Nucleus implant." Otolaryngology-Head and Neck Surgery, 117**,** 188-195.

[143] Skinner, M. W., Holden, L. K., Whitford, L. A., Plant, K. L., Psarros, C. and Holden, T. A. (2002). "Speech recognition with the nucleus 24 SPEAK, ACE, and CIS speech coding strategies in newly implanted adults." Ear and Hearing, 23**,** 207-223.

[144] Smith, J. O. and Abel, J. S. (1999). "Bark and ERB bilinear transforms." IEEE Transactions on Speech and Audio Processing, 7**,** 697-708.

[145] Spahr, A. J., Dorman, M. F. and Loiselle, L. H. (2007). "Performance of patients using different cochlear implant systems: Effects of input dynamic range." Ear and Hearing, 28**,** 260-275.

[146] Spriet, A., Van Deun, L., Eftaxiadis, K., Laneau, J., Moonen, M., Van Dijk, B., Van Wieringen, A. and Wouters, J. (2007). "Speech understanding in background noise with the two-microphone adaptive beamformer BEAM (TM) in the nucleus Freedom (TM) cochlear implant system." Ear and Hearing, 28**,** 62-72.

[147] Srinivasan, P. and Jamieson, L. H. (1998). "High-quality audio compression using an adaptive wavelet packet decomposition and psychoacoustic modeling." IEEE Transactions on Signal Processing, 46**,** 1085-1093.

[148] Stakhovskaya, O., Sridhar, D., Bonham, B. H. and Leake, P. A. (2007). "Frequency map for the human cochlear spiral ganglion: implications for cochlear implants." J Assoc Res Otolaryngol, 8**,** 220-233.

[149] Steeneken, H. J. M. and Houtgast, T. (1980). "Physical Method for Measuring Speech-Transmission Quality." Journal of the Acoustical Society of America, 67**,** 318-326.

[150] Steeneken, H. J. M. and Houtgast, T. (1982). "Some Applications of the Speech Transmission Index (Sti) in Auditoria." Acustica, 51**,** 229-234.

[151] Stegmann, J. and Schroder, G. (1997). "Robust voice-activity detection based on the wavelet transform." 1997 Ieee Workshop on Speech Coding for Telecommunications, Proceedings**,** 99-100.

[152] Stickney, G. S., Zeng, F. G., Litovsky, R. and Assmann, P. (2004). "Cochlear implant speech recognition with speech maskers." J Acoust Soc Am, 116**,** 1081-1091.

[153] Taal, C. H., Hendriks, R. C., Heusdens, R. and Jensen, J. (2011). "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech." IEEE Transactions on Audio Speech and Language Processing, 19**,** 2125-2136.

[154] Tan, B. T., Lang, R., Schroder, H., Spray, A. and Dermody, P. (1994). *Applying Wavelet Analysis to Speech Segmentation and Classification*, BELLINGHAM, SPIE - INT SOC OPTICAL ENGINEERING.

[155] Tasmaz, H. and Ercelebi, E. (2008). "Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE-STSA estimation in various noise environments." Digital Signal Processing, 18**,** 797-812.

[156] Teager, H. M. and Teager, S. M. (1990). "Evidence for Nonlinear Sound Production Mechanisms in the Vocal-Tract." Speech Production and Speech Modelling, 55**,** 241-261.

[157] Van Schijndel, N. H., Houtgast, T. and Festen, J. M. (2001). "Effects of degradation of intensity, time, or frequency content on speech intelligibility for normal-hearing and hearing-impaired listeners." Journal of the Acoustical Society of America, 110**,** 529-542.

[158] Vanhoesel, R. J. M. and Clark, G. M. (1995). "Evaluation of a Portable 2-Microphone Adaptive Beamforming Speech Processor with Cochlear Implant Patients." Journal of the Acoustical Society of America, 97**,** 2498-2503.

[159] Verschuur, C. (2007). *Acoustic models of consonant recognition in cochlear implant users.* PhD, University of Southampton.

[160] Verschuur, C., Lutman, M. and Wahat, N. H. (2006). "Evaluation of a non-linear spectral subtraction noise suppression scheme in cochlear implant users." Cochlear.Implants.Int., 7**,** 193-196.

[161] Veselinovic, D. and Graupe, D. (2003). "A wavelet transform approach to blind adaptive filtering of speech from unknown noises." IEEE Transactions on Circuits and Systems Ii-Analog and Digital Signal Processing, 50**,** 150-154.

[162] Vetterli, M. and Herley, C. (1992). "Wavelets and Filter Banks - Theory and Design." IEEE Transactions on Signal Processing, 40**,** 2207-2232.

[163] Virag, N. (1999). "Single channel speech enhancement based on masking properties of the human auditory system." IEEE Transactions on Speech and Audio Processing, 7**,** 126-137.

[164] Wang, D. L. (2005). "On ideal binary mask as the computational goal of auditory scene analysis." Speech Separation by Humans and Machines**,** 181-197.

[165] Weickert, T., Benjaminsen, C. and Kiencke, U. (2009). "Analytic Wavelet Packets-Combining the Dual-Tree Approach With Wavelet Packets for Signal Analysis and Filtering." Ieee Transactions on Signal Processing, 57**,** 493-502.

[166] Wilson, B. S. and Dorman, M. F. (2008a). "Cochlear implants: A remarkable past and a brilliant future." Hear Res, 242**,** 3-21.

[167] Wilson, B. S. and Dorman, M. F. (2008b). "Cochlear implants: Current designs and future possibilities." Journal of Rehabilitation Research and Development, 45**,** 695-730.

[168] Wouters, J. and Vanden Berghe, J. (2001). "Speech recognition in noise for cochlear implantees with a two-microphone monaural adaptive noise reduction system." Ear and Hearing, 22**,** 420-430.

[169] Yang, L. P. and Fu, Q. J. (2005). "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise." Journal of the Acoustical Society of America, 117**,** 1001-1004.

[170] Yang, X. W., Wang, K. and Shamma, S. A. (1992). "Auditory Representations of Acoustic-Signals." IEEE Transactions on Information Theory, 38**,** 824-839.

[171] Yao, J. and Zhang, Y. T. (2002). "The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations." IEEE Trans Biomed Eng, 49**,** 1299-1309.

[172] Zeng, F. G. (2004). "Trends in cochlear implants." Trends Amplif, 8**,** 1-34.

[173] Zwicker, E. and Terhardt, E. (1980). "Analytical Expressions for Critical-Band Rate and Critical Bandwidth as a Function of Frequency." Journal of the Acoustical Society of America, 68**,** 1523-1525.