UNIVERSITY OF
Southampton
Institute of Sound and
Vibration Research

Speech perception in a sparse domain

Guoping Li

Institute of Sound and Vibration Research

University of Southampton

A thesis submitted for the degree of

Doctor of Philosophy

March 2008

Acknowledgements

Foremost, I would like to thank Prof. Mark Lutman, my supervisor. Without his encouragement and support, I would not be able to enjoy so much from my PhD. I feel my mind has been completely free and I was allowed to explore whatever I was interested. This is the best gift a supervisor can give to a student. I also take this opportunity to thank Prof. Martin Cooke. He kindly provided the speech materials and his results on glimpsing model. I would also like to thank Dr. Christopher James' help on my understanding of independent component analysis and his tutorial paper on ICA. South England Cochlear Implant Centre, SOCIC, helped contact CI patients for my experiments. Specially thanks to Roberta Buhagiar, Sarah Merritt and Julie Eyles' help on the patient's data. My thesis would be impossible without the support from ISVR. The discussion with Pof. Paul white on PCA helped me to develop the idea of SPARSE. Dr. Stefan Bleeck not only helped me keep fit by playing squash with him but also helped on the programming of the interface. Without his help, the experiment probably would have taken much longer. Dr. Shouyan Wang make me more critical about my ideas and forced me to read more about information theory. Dr. Ben Lineton let me khow to enjoy much of the landscape of Hampshire and England by cycling. I certainly will keep the cycling experiences as important part of my PhD. Also special thanks also to Cochlear Company, they provided both finance and equipments to my research. Special thanks to Herbert Mauch, he came to help us setup the NIC streaming experiments. And Dr. Matthijs Killian encouraged us on the patent application. Also thanks my colleagues, Mona Mahmoud, Sheetal Athalye, Hadeel

Alsaleh, Srikanta Mishra, Gilles Pigasse. It is so convenient we can learn Arabic, French and Hindu in one office. Also thanks my friends Jin Yan, Sun Dan, Hou Yan, Lv Jin, Wei Kun, Hai Rong. With your friendship, my PhD life became so colourful and I did not feel the boring sided of PhD.

Finally, special thanks to my parents for trusting, supporting and encouraging me to pursue the PhD. I also have to say thanks to my wife Zhihong. Your love makes me feel so happy every day. The PhD is impossible without your support. Some of your artful ideas even contributed to the key concepts of this thesis. This thesis is dedicated to you, my love.

Abstract

Environmental statistics are known to be important factors shaping our perceptual system. The visual and auditory systems have evolved to be efficient for processing natural images or speech. The common characteristics between natural images and speech are that they are both highly structured, therefore having much redundancy. Our perceptual system may use redundancy reduction and sparse coding strategies to deal with complex stimuli every day. Both redundancy reduction and sparse coding theory emphasise the importance of high order statistics signals.

This thesis includes psycho-acoustical experiments designed to investigate how higher order statistics affect our speech perception. Sparseness can be defined by the fourth order statistics, kurtosis, and it is hypothesised that greater kurtosis should be reflected by better speech recognition performance in noise. Based on a corpus of speech material, kurtosis was found to be significantly correlated to the glimpsing area of noisy speech, an established measure that predicts speech recognition. Kurtosis was also found to be a good predictor of speech recognition and an algorithm based on increasing kurtosis was also found to improve speech recognition score in noise. The listening experiment for the first time showed that higher order statistics are important for speech perception in noise.

It is known the hearing impaired listeners have difficulty understanding speech in noise. Increasing kurtosis of noisy speech may be particularly helpful for them to achieve better performance. Currently, neither hearing aids nor cochlear implants help hearing impaired users

greatly in adverse listening enviroments, partly due to having a reduced dynamic range of hearing. Thus there is an information bottleneck, whereby these devices must transform acoustical sounds with a large dynamic range into the smaller range of hearing impaired listeners. The limited dynamic range problem can be thought of as a communication channel with limited capacity. Information could be more efficiently encoded for such a communication channel if redundant information could be reduced. For cochlear implant users, unwanted channel interaction could also contribute lower speech recognition scores in noisy conditions.

This thesis proposes a solution to these problems for cochlear implant users by reducing signal redundancy and making signals more sparse. A novel speech processing algorithm, SPARSE, was developed and implemented. This algorithm aims to reduce redundant information and transform signals input into more sparse stimulation sequences. It is hypothesised that sparse firing patterns of neurons will be achieved, which should be more biological efficient based on sparse coding theory. Listening experiments were conducted with ten cochlear implant users who listened to speech signals in modulated and speech babble noises, either using the conventional coding strategy or the new SPARSE algorithm. Results showed that the SPARSE algorithm can help them to improve speech understanding in noise, particularly for those with low baseline performance. It is concluded that signal processing algorithms for cochlear implants, and possibly also for hearing aids, that increase signal sparseness may deliver benefits for speech recognition in noise. A patent based on the algorithm has been applied for.

List of Abbreviations

ACE	Advanced Combined Encoder
CASA	Computational auditory scene analysis
CI	Cochlear implant
CIS	Continuous interleaved sampling
ICA	Independent component analysis
PCA	Principal component analysis
SNR	Signal-to-noise-ratio
SPEAK	Spectral peak
STFT	Short-time Fourier Transform
VCV	Vowel-Consonant-Vowel

Contents

1 Introduction **1**

1.1 Contribution to knowledge 1
1.2 Introduction 2
1.3 Auditory scene analysis and Gestalt grouping rules 4
1.4 Gestalt rules and structured stimuli 7
1.5 Structured world and redundant stimuli 9
1.6 Redundancy and perception 10
1.6.1 Redundancy and visual perception 11
1.6.2 Redundancy and speech perception 13
1.6.2.1 Glimpsing speech 15
1.7 Redundancy exploration via neurons 16
1.7.1 Barlow's redundancy exploration 17
1.7.2 Sparse coding 18
1.7.3 Redundancy and higher order statistics 22
1.8 Thesis outline 23

2 Sparseness and glimpsing **25**

2.1 Introduction 25
2.2 Methods 26
2.3 Speech materials 27
2.4 Results of quantitative analysis of kurtosis 28
2.5 Discussion 30

3 Exploring redundancy: a macroscopic approach for hearing research **33**
3.1 Introduction . . . 33
3.2 Microscopic approach for hearing research . . . 34
3.2.1 Traditional approach for hearing research . . . 34
3.2.2 Traditional approach to speech perception . . . 36
3.2.3 Computational auditory scene analysis . . . 37
3.3 Auditory communication system and information theory . . . 38
3.4 Statistics for perception . . . 43
3.4.1 State space and perception . . . 44
3.4.2 Principal component analysis . . . 48
3.4.3 Independent component analysis . . . 51
3.4.4 Projection pursuit . . . 57
3.5 Conclusion . . . 58

4 Sparseness and speech perception **60**
4.1 Introduction . . . 60
4.2 Generation of speech and noise material . . . 61
4.3 Subjects and procedure . . . 63
4.4 Results . . . 64
4.5 Discussion . . . 65
4.5.1 Implications for auditory neuroscience . . . 65
4.5.2 Further research . . . 66
4.5.3 Implications for hearing aids and cochlear implants . . . 67

5 Speech perception in different spaces **69**
5.1 Introduction . . . 69
5.2 Time frequency domain . . . 73
5.3 PCA and ICA transformation . . . 75
5.4 Sparse domain transform . . . 77
5.5 Discussion . . . 80

6 Sparse stimuli for cochlear implants **82**
6.1 Introduction . . . 82
6.2 Introduction to cochlear implants . . . 87
6.2.1 The principle of cochlear implants . . . 87
6.2.2 The task of speech processor . . . 88
6.2.3 Review of speech processing algorithms . . . 91
6.2.4 Sparse stimuli for cochlear implants: SPARSE . . . 93
6.3 Sparse stimuli for cochlear implants . . . 94
6.3.1 Combined compact coding (PCA)and sparse coding (ICA) 94
6.3.1.1 PCA processing for speech envelopes . . . 97
6.3.2 ICA for cochlear implants: SPARSE algorithm . . . 100
6.3.3 Derivation of the SPARSE algorithm . . . 105
6.4 Discussion and Conclusions . . . 109

7 Experimental evaluation of SPARSE algorithm for cochlear implants **112**
7.1 Introduction . . . 112
7.2 SPARSE algorithm parameters . . . 113
7.3 SPARSE evaluation . . . 114
7.3.1 Speech materials . . . 114
7.3.2 Increased sparseness . . . 116
7.4 Subjective experiments . . . 117
7.4.1 Experiments I: Normal hearing subjects and sparse stimuli 117
7.4.1.1 Results and Discussion . . . 118
7.4.2 Experiment II: Cochlear implant users and sparse stimuli . 118
7.4.2.1 Speech materials and subjects . . . 118
7.4.3 Results and Discussion . . . 120

8 Conclusions **125**
8.1 Overview . . . 125
8.2 Discussion . . . 126
8.2.1 Investigation of higher order statistics in speech perception 126
8.2.2 Information capability matching . . . 127
8.2.3 Kurtosis and SNR . . . 128

8.2.4 Parameters of SPARSE 128
8.3 Conclusion and future work 128

9 Appendix A: Results of subjective experiments **130**

10 Appendix B: NIC streaming **139**

References **151**

List of Figures

1.1 Principle of cochlear implants 5
1.2 Cocktail party problem 6
1.3 The Gestalt rules in visual perception 7
1.4 Gestalt rules in sound perception 8
1.5 Structure and natural world 10
1.6 Redundancy and communication bandwidth 12
1.7 Redundancy in image and speech 13
1.8 Neurons explore the redundancy of the environment 16
1.9 Examples of three levels of kurtosis representing neural response patterns 20
1.10 Compact coding and sparse coding 21

2.1 Speech has higher kurtosis than gaussian noise 27
2.2 Kurtosis values and the number of talkers 28
2.3 Correlation between kurtosis and glimpsing area 29
2.4 Correlation between kurtosis and average speech recognition score 30

3.1 Information theory for telecommunication and the auditory system 41
3.2 Example of state space and sparse coding by neurons 45
3.3 PCA and state space of two picture stimuli 49
3.4 PCA and dimension reduction 51
3.5 An example of state space transform with PCA, unable to reduce the dimension. 53
3.6 Two signals before mixing. 54
3.7 Example of ICA separating two mixtures. 55

3.8 Convergence of project pursuit algorithms 57

4.1 Examples of signals with increasing kurtosis. 61
4.2 Setup of the mixing process. 62
4.3 Different level of kurtosis 63
4.4 Increased kurtosis predicts improved speech recognition score . . . 64
4.5 Logistic regression for speech recognition score and kurtosis. . . . 65

5.1 Perception in the state space 70
5.2 Basis function of FFT 74
5.3 Patten playback of the spectrum 75
5.4 Redundancy reduction of image by PCA 76
5.5 Redundancy reduction of speech by PCA 78
5.6 Basis functions of PCA 79
5.7 Basis functions of ICA 80

6.1 Illustration of the bottleneck problem between acoustical information and electrical stimulation 83
6.2 Diagram of proposed processing scheme 84
6.3 A schematic overview of cochlear implant working principle 89
6.4 CI speech processor mimics the function of normal hearing 90
6.5 Standard CIS processor 92
6.6 The ‘sparse’ direction of cochlear implant developments 95
6.7 Flowchart of SPARSE algorithm 96
6.8 Example of spectrum analysis by PCA 98
6.9 Eigenvalues and reconstruction of a signal by PCA 100
6.10 Kurtosis and reconstruction of ACE output by PCA 101
6.11 The spectrogram of ACE and PCA output 102
6.12 The application of ICA in the SPARSE algorithm 103
6.13 Example of spectrum analysis by ICA 104
6.14 Example of sparse spectrum by ACE and SPARSWE 106
6.15 Kurtosis and reconstruction of ACE output by SPARSE 107
6.16 Shrinkage function, a soft thresholding method 109
6.17 Example of the SPARSE algorithm in 5 dB noisy condition 110

7.1 Example of VCV word and the corresponding masking noises . . . 115
7.2 Increased sparseness with SPARSE 116
7.3 Normal hearing subjects with SNR = 5 dB 119
7.4 The improved speech recognition score by the SPARSE algorithm for normal hearing listeners 120
7.5 The improvement speech recognition by the SPARSE algorithm for CI users 121
7.6 Speech recognition score for CI users whose baseline performance is lower than 70% 122

9.1 Normal hearing subject recognition in babble noise 131
9.2 CI users recognition score in noise 134
9.3 CI users recognition score in noise 135

10.1 NIC streaming process 140

List of Tables

9.1 Normal hearing listeners in quiet 132
9.2 Normal hearing listeners in babble noise 132
9.3 Normal hearing listeners in modulated babble noise 133
9.4 CI users listening VCV words in quiet 136
9.5 CI users in babble noise 137
9.6 CI users in modulated babble nois 138

Chapter 1

Introduction

1.1 Contribution to knowledge

This thesis investigates the application of sparse coding to speech perception and its implication for cochlear implants. Sparseness is quantified through the fourth order statistic, kurtosis and it relates to the core concept of redundancy reduction and sparse coding. This thesis also investigates the principle of speech perception and its relation with the high order statistical characteristics of speech. It shows that like other sensations, environmental stimuli statistics have an important effect on speech perception. High order statistics are used to quantify the key property of speech: sparseness.

This thesis first builds the link between speech perception and high order statistics through computing the kurtosis and glimpsing areas of speech. Glimpsing (Cooke, 2006) is thought to be an efficient strategy to glimpse the speech against the noise by taking advantage of areas with higher local signal-noise ratio. Kurtosis is the fourth order statistics which can measure the sparse distribution of the data. The relationship between kurtosis and speech perception score is investigated. Positive correlation was found between the kurtosis and speech recognition score, suggesting that the kurtosis could be explored by the auditory system. It also showed that the sparse structure of the data is important for speech perception.

Further this thesis shows that speech recognition performance of normal hearing subjects could be improved by making the noisy signals more sparse. In the

experiment, with projection pursuit algorithm (Stone, 1993), two noisy signal mixtures are projected into a more sparse space, the output is more sparse and easier to recognize.

Based on the fact that the higher order statistical characteristics of speech have an important effect on speech perception, this thesis proposes that sparse stimuli for cochlear implant users might provide a better interface between auditory neurons and the acoustical space. Based on the concept of sparse coding and redundancy reduction (Barlow, 2001; Field, 1994; Hyvarinen *et al.*, 2005), this thesis develops an algorithm, SPARSE,which can transform the spectrum envelope to a sparse representation and use it to stimulate auditory neurons. Subjective experiments show that the algorithm is helpful to cochlear implant users whose baseline performance is low. This algorithm could be potentially applied for hearing aid users as well. The principles of the SPARSE algorithm have been included in a patent application filed by the author and colleague (UK patent application number 0717210.9, filed on 5 September 2007).

One of the main contributions of this thesis is including the concept of sparse coding and redundancy exploration into research on speech perception and applying it in cochlear implant speech processing algorithms. This study shows that sparse coding is important for speech perception and it could help cochlear implant users to achieve better performance on speech recognition. The other important contribution of this thesis is that it supports the idea that environmental statistics do have an important role in our auditory perceptual system. By introducing the sparse coding theory into auditory perception research, new insights on speech perception could be obtained.

1.2 Introduction

Speech perception research has been an exciting and mysterious research field. Our ears collect acoustical sound information and send it to the brain. How the acoustical sound is transmitted and encoded have been the main research questions. Many classic theories and knowledge about speech perception have been developed based on psycho-acoustical experiments, computational models and neurophysiologic experiments.

Understanding of the auditory system was especially interesting for research on the cochlear implant, which has helped many profoundly deaf people to regain the sense of hearing by stimulating the auditory nerves through electrodes implanted in the inner ear. Most cochlear implant users benefit greatly and some of them even can communicate using the telephone without lip reading. This was unimaginable even just decades ago.

Fig. 1.1 shows the function of a cochlear implant. It bypasses the outer ear, middle ear, and hair cells of the inner ear. Sounds can be picked up by the microphone and processed with a speech processor, electrical stimuli then are sent to the auditory neurons through the electrodes implanted in the cochlea during surgery.

When such a biological system can be functionally replaced by a man made device, we would have been very confident to say that the principles of peripheral auditory system are well understood. In fact, as research goes deeper into the field of cochlear implants and speech perception, it seems now we have more questions and puzzles about hearing. And some of them even go back to the original fundamental questions of hearing science: what is the role of our ears?

The ear has traditionally been viewed as a frequency analyzer (Helmholtz, 1863; Moore, 2003a). But later people soon realized far more is involved in speech processing than mere time-frequency analysis. The main function of cochlear implants is spectral analysis, mimicking the function of inner ear. It stimulates the auditory nerves with the power spectral envelopes of speech. Cochlear implant research has been very successful and it is certainly a very successful neural prosthetic device. However, it works well in quiet but not in noisy environments (Loizou *et al.*, 2000; Moore, 2003b). Also the performance of each individual is highly variable. While looking for enhanced speech processing algorithms to improve the speech recognition performance in such adverse environments, it is worthwhile taking a step back to reconsider the function of ears. And this could bring us to a new stage of signal processing for cochlear implants.

If the ear only does a faithful analysis of the spectro-temporal characteristics of acoustic waveforms, similar to what a cochlear implant does, it could actually make it difficult to understand speech in a noisy environment (e.g. in a cocktail

party). In daily life, however, we are able to understand speech in a quite complex sound environment, with music on, or a background of many people talking.

Fig. 1.2 shows the cocktail party problem. Two people are talking at the same time, but normal hearing subjects are still able to focus one talker's speech. The spectrum of the mixture quite different from the individual talkers. This cocktail party problem was posed long time ago by Helmholtz (1877): how one hears the quality of an instrument playing among others. Cherry (1953) and Bregman (1990) tried to find out what cues are important for separation through psychoacoustic experiments. All the information Cherry found is a mixture of different factors (Arons, 1992), such as voices from different directions, pitch, mean speeds, lip-reading, gestures and different accents. He concluded that:

> *"The result is a babble, but nevertheless the message may be separated".*

1.3 Auditory scene analysis and Gestalt grouping rules

After almost another half a century, Bregman (1990) concluded some unified grouping principle for hearing, also called *Auditory Scene Analysis(ASA).*

In order to investigate speech perception in a more complex environment, Bregman (1990) conducted experiments on the auditory system with complex multiple sounds. The stimuli used in these experiments were more than one sound. The main aim of such research was to investigate how we group different frequency components of sounds, given a mixture of different sounds.

Bregman concluded some principles on how we can focus on one sound in a cocktail party like environment. He introduced some Gestalt grouping rules (Bregman, 1990), which might be used by the auditory system. These rules can define how we group a mixture of different elements into correct objects. For example, the sounds with common onset or common modulation are more likely to be grouped as one sound. There are also rules like similarity, good continuation, and so on. Fig. 1.3 shows that our visual grouping can be explained by similar

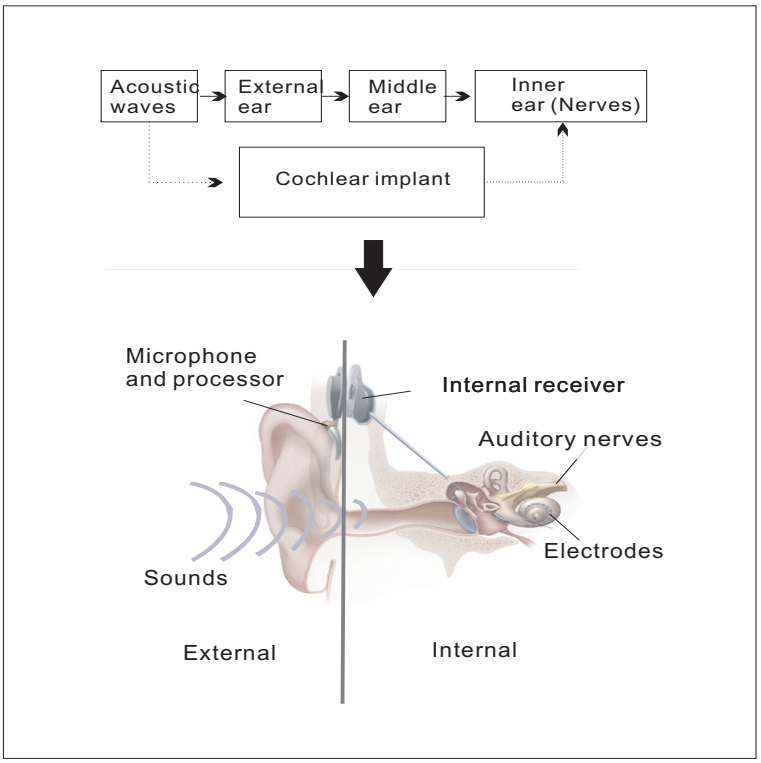


Figure 1.1: *Principle of cochlear implants. Cochlear implants use electrical stimuli to stimulate the auditory neurons of the profoundly deaf. The firing patterns stimulated by the electrical stimuli can be interpreted by the brain as sounds. In the upper panel, the solid lines are the routes of sound for normal hearing. The dashed line shows that the cochlear implant work as an intermediate stage to encode the sound waves as neuron firing patterns. In the lower panel, it shows how a cochlear implant works. Cochlear implant are made up of: (a) external parts including microphone, speech processor and an external transmitter (b)internal parts including internal receiver and electrodes array. The microphone picks up the sound and extract some useful information through certain algorithm in the speech processor. It sends electrical stimuli to auditory neurons via electrodes. The users will get a sense of hearing with such electric pulse stimuli.*

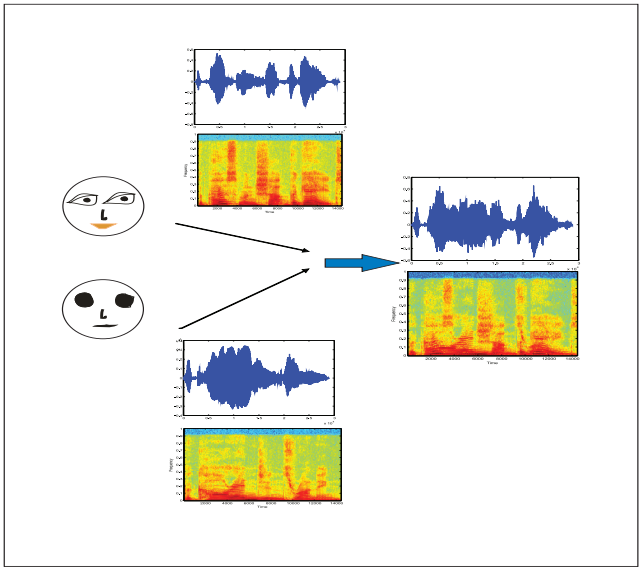


Figure 1.2: *Two or more talkers are speaking simultaneously. It is difficult, if purely based on the spectrogram of the mixture, to group different frequency components into one sound. Such a mixture will be difficult for automatic speech recognition machines. But we are able to cope with such situation on a daily basis.*

Gestalt principles. For example, we group the circles to rows because the rows distance is much smaller than the column distance. Also the illusions of the white triangular shape appearing in the lower graph can be explained by the good continuity principle. With the same good continuity principle, we also see the double pictures for the lower right part (two faces and a cup like structure).

These are some of several Gestalt grouping rules in vision research. Bregman concluded that similar rules can be used to explain the grouping of sounds. As shown in Fig. 1.4, sounds are allocated to different groups by the proximity of their physical distance (time) and whether the frequency band is continuous (good continuation) (Bregman, 1990). The rules used are almost the same as in Fig. 1.3.

Based on research of auditory scene analysis, computational models (Brown & Cooke, 1994; Cooke & Ellis, 1998; Cooke & Brown, 1993) were built to model this process, which first decompose the signal into time-frequency elements based on auditory filter models. Their features, such as periodicities, frequency transitions, onsets and offsets are extracted. The time-frequency elements can then

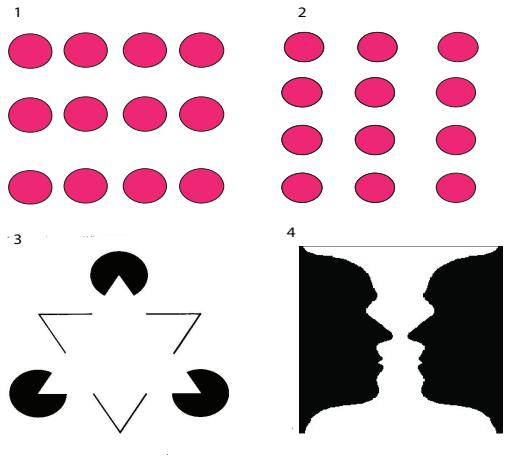


Figure 1.3: *The Gestalt rules in visual perception. The upper panel will be perceived as rows for '1' and columns for '2' according to the proximity of Gestalt grouping rules. '3' and '4' are vision illusions which can be explained based on good continuity.*

be grouped based on the application of Gestalt rules to the features. Waveforms can be re-synthesized from a group of such elements.

These models helped to establish possible perception principles in our auditory system used for grouping. However, these Gestalt rules are not explicit and more descriptive than operational. These rules can be thought as simply summarizing description of the phenomena of human perception. The Gestalt psychology was criticized by some (Bruce *et al.*, 1996) as:

> *"The physiological theory of the Gestaltists has fallen by the wayside, leaving us with a set of descriptive principles, but without a model of perceptual processing. Indeed, some of their 'law' of perceptual organization today sound vague and inadequate. What is meant by a 'good' or 'simple' shape, for example?"*

1.4 Gestalt rules and structured stimuli

Gestalt rules, although vague, may reflect the characteristics of the stimuli: highly structured. The key grouping elements, such as onset, common modulation, will

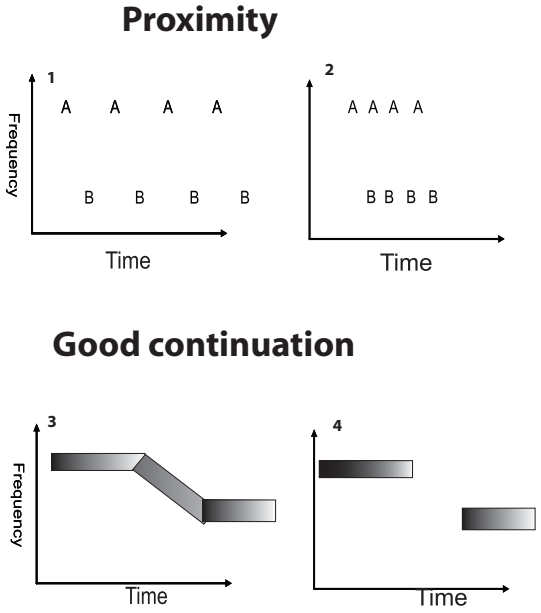


Figure 1.4: *Gestalt rules in sound perception. The first is proximity rules. In the upper panel, for '2', A's are more near to each other along the time dimension. So 2 is more likely to be heard as :'A-A-A..A', 'B-B-B..B'; For '1', the sound is more likely to be heard as :A-B-A-B...A-B, . In the lower panel, '3' could be heard as a continuous sound as A and B are connected by glissandi, while 4 will be heard as two different sounds as abrupt changes will be interpreted as a new source. (Redrawn from Purwins* et al. *(2000)).*

not even exist if a stimulus has no structure.

After all, the sensory organs have to deal with mixed stimuli in daily life. Our sensory organ has to be able to separate different stimuli from a given mixture. The ideal single 'pure' sensory stimuli, such as pure tones used in the laboratory conditions, seldom exist in real life environments.

Our perceptual system must have developed efficient methods to deal with complex stimuli. These Gestalt grouping rules reflect that there are certain *structures* in the stimuli (Attneave, 1954). It is believed that signals from the natural world are highly structured (Attneave, 1954; Barlow, 2001). Here the structure is contrary to randomness. For example, we hardly see any random gaussian like pictures, or listen to pure gaussian noise in the natural environment.

The world around us is highly structured, not completely random. It is the structure that makes it possible for us to have all these grouping rules. These structured stimuli must have an important effect on our perceptual system. Fig. 1.5 shows examples of structured images/shapes. The elliptical and triangular shape are abstracted from the natural environment. The picture taken in a forest shows that the structure of trees, which is hidden in a random like picture. A computer generated random gaussian picture is shown on the lower right. There is hardly any structure within the random image.

For perception, it is important to explore structures of stimuli. Complete randomness provides no structure and also few rules can be used to distinguish or group such objects. And fortunately, the non-random characteristics of the world provides us a good chance to perceive and understand it.

1.5 Structured world and redundant stimuli

When we have observed the structures or have knowledge of these structures, we can then use certain rules to distinguish different objects, such as a circle or triangle. Also importantly, these external sensory stimuli also become partially *redundant* for us because we would be able to predict one portion of the signal based on the other spatial or temporal portion of a given signal. Indeed it is the knowledge of structure that make a signal become redundant. If we know the structure of an object, the representation can be more economical as we

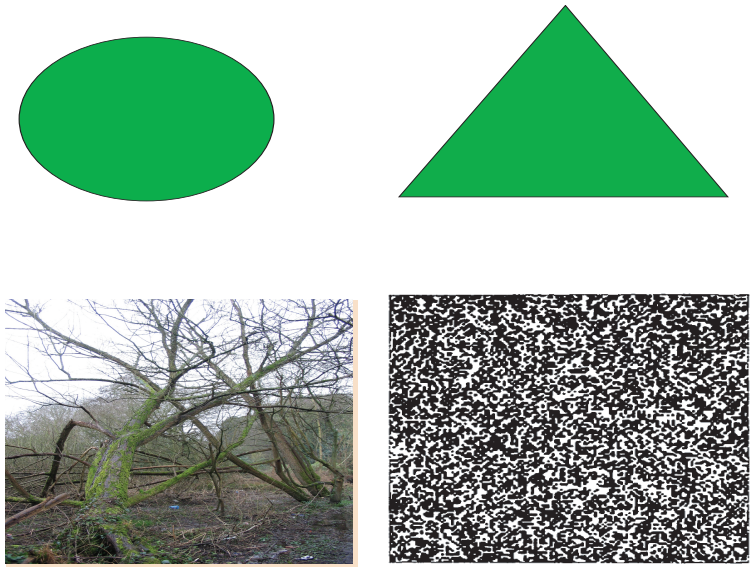


Figure 1.5: *Structure and natural world. Elliptical and triangular shapes are quite structured shapes abstracted from natural world. Given a structure like a equilateral triangular shape, we only need the length of one side to recover the whole shape. Similarly, we need not remember each point of on the circle to recover a circle. The natural image of trees, seemingly quite random, however, shows that there are always some structure in the natural world (Trunk and branches of the trees). The completely random picture generated by computer (Attneave, 1954), however, shows no structure.*

can encode the signal with less symbols. And this economical representation has also been called a redundancy reduction strategy (Barlow, 1959, 2001). For example, if we know a triangular shape is equilateral, we only need to refer the length and orientation of one side to redraw the shape, without the need to remember the length of three sides. It is perhaps the structural characteristics of the environment that makes humans able to learn about nature and science.

1.6 Redundancy and perception

Our perceptual systems are evolved to explore the redundancy of the environment, perceiving the causes of the structure. The structure can be different. For an

image, it can have different shape, textures; or a sound, it can have different harmonics or fundamental frequencies. It is difficult to describe different structures in general. But the common ground behind structured stimuli is that they all have redundancy. And this redundancy feature of stimuli is especially important for perception. Investigation of perception from an information processing point of view could provide some new insight on how our perceptual system processes environmental stimuli every day.

1.6.1 Redundancy and visual perception

Attneave (1954) argued that Gestalt grouping rules indeed reflect the redundant nature of the natural environment. As he pointed out:

> *"It is not surprising that the perceptual machinery should 'group' those portions of its input which share the same information: any system handling redundant information in an efficient manner would necessarily do something of the sort."*

One important contribution from Attneave is that he introduces the information theory into the psychological experiments on visual perception. The information theory proposed by Claude Shannon (Shannon, 1948) provides a mathematical description of redundancy and communication efficiency. He developed a mathematical theory that quantified the variables involved: the amount of information transmitted and the capacity of a channel to carry information. One of the key issues in information theory is that the amount of information carried by a signal is often less than the maximum amount that could be transmitted by the communication channel (also known as the 'bandwidth') (Barlow, 2001). Fig. 1.6 shows the relationship between bandwidth, redundancy and information. Part of the capacity of the communication channel is occupied by redundant parts of the signal.

Attnneave showed there is much redundancy in the image through guessing experiments by asking subjects to guess the colour of pixels in the image (Attneave, 1954). The fact that errors made are much fewer than chance proves that the image is predictable, and information theory shows that predictability

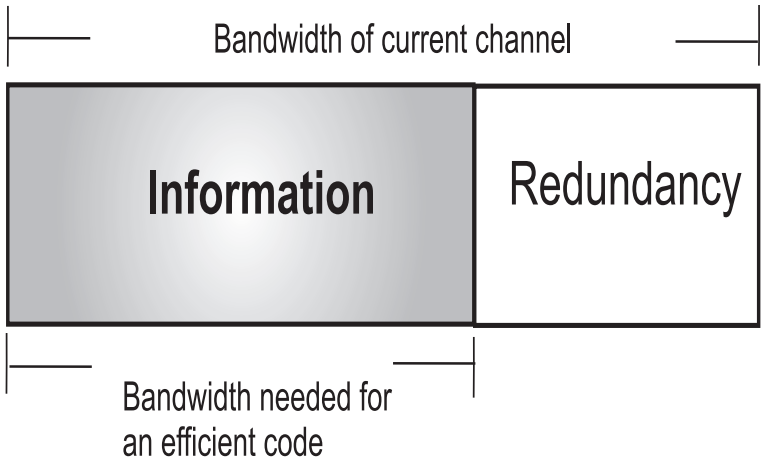


Figure 1.6: *Redundancy and communication bandwidth. The graph shows the bandwidth carrying a signal is typically occupied by both information and redundant parts of the signal. The bandwidth can be more efficiently used if the bandwidth is occupied by the information only. Redrawn from Hoyer & Hyvarinen (2002).*

is essentially the same thing as redundancy. Attneave's idea of taking perception as an information-handling process is quite appealing, as it can help make internal perception describable in a mathematical sense. Perception, according to Attneave, is to detect the structure of the stimuli and uses the redundancy.

> *"It appears likely that a major function of the perceptual machinery is to strip away some of the redundancy of stimulation, to describe or encode incoming information in a form more economical than that in which it impinges on the receptors."*

His key observation on visual perception was that the borders of images are the perceptually important information and the rest of the images can be thought as redundant information. As we can see a cat can be perceived by using only the edges (see Fig. 1.7). And interestingly, Shannon *et al.* (1995) showed that the envelope of speech carries important information for speech. And people can understand speech based on the envelopes only given four or more channels, showing that speech is also redundant.

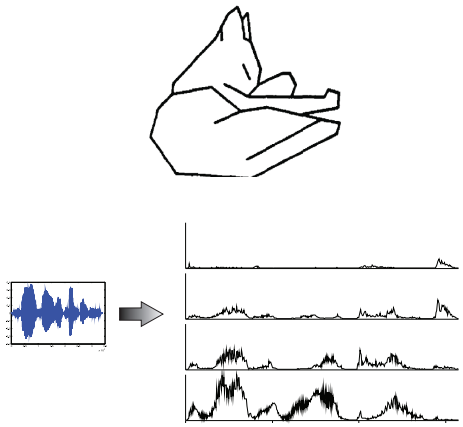


Figure 1.7: *Redundancy in image and speech. Attneave has shown that the cat can be perceived by the edges of the image. Shannon* et al. *(1995) showed that normal hearing people could understand speech with only temporal information in four broad band frequency channels without spectral fine structures. Both Attneave and Shannon's investigation on perception show that sensory stimuli are redundant.*

1.6.2 Redundancy and speech perception

It has long been known that redundancy can help to correct errors in telecommunication systems. And speech is such a typical information carrier. The redundancy of speech can help deliver the speech information in adverse environments.

In telecommunication, if a code 001 is sent and the receiver only gets 0?1, losing the middle '0', guesses have to be made for the missing digit. Alternatively, redundancy can be added to make the communication more reliable in noisy communication channels. If 0010 was transmitted, instead of 001, defining when the end is zero, the second digits also should be 0. Thus even if we lost the second digit due to communication errors, we still get the whole sequence by checking the last digit. Here adding redundancy is a way to build a more reliable communication system in noisy conditions.

Speech can be thought as an information carrier. The rich redundancy of speech makes it a reliable tool for human communication in different environments. Coker & Umeda (1974) defined speech redundancy as

"any characteristic of the language that forces spoken messages to

> *have, on average, more basic elements per message, or more cues per basic element, than the barest minimum (necessary for conveying the linguistic message).”*

And they concluded that the purpose of redundancy in speech communication is similar to telecommunication: to provide a basis for error correction and resistance in noisy environments (Greenberg *et al.*, 2004). As Coke and Umeda (1974) futher pointed out:

> *“if redundancy is a property of a language and has to be learned, then it has a purpose.”*

The redundancy of speech was investigated mainly to test how robust the speech can be. Specially after the telephone was invented, people were interested in how to transfer speech with smaller bandwidth, which was a big concern due to the limited communication capacity of the analogue telephone system. Fletcher in Bell Laboratories investigated how to reduce the bandwidth of the speech signals without compromising the ability to communicate using the telephone (Allen, 1994; Fletcher, 1953; Fletcher & Galt, 1950). As a result of these studies, speech outside the range of 300-3400Hz can be discarded without affecting the speech intelligibility for telephone communication.

The redundancy of speech was also shown by subjective experiments listening to information reduced speech (Distorted): making holes in the speech spectrum (Kasturi *et al.*, 2002) and gaps in the temporal waveform (Strange *et al.*, 1983). The conclusion is similar in that listeners are able to understand speech only based on partial of speech. Experiments were also tried to test how robust the speech can be by representing speech with different elements of speech, such as envelope and fine structure of the speech (Drullman, 1995; Shannon *et al.*, 1995), or only using few frequency channels to represent the whole speech (Warren *et al.*, 1995), or just chopping the signal to make some slices zero or filling it with noise (Miller & Licklider, 1950).

All these experiments showed that speech is redundant and only few components are needed to allow people to understand speech. These experiments showed that there is much redundancy in speech and the redundancy can be

thought as a fundamental property of speech. Speech production systems put a physical constraint on what speech would sounds like, just as light reflection rules in the natural world define how a picture would appear. These constraints make speech have unique structure and so there is deemed to be redundancy in the speech.

1.6.2.1 Glimpsing speech

The ability of understanding speech based on partial information was explained by the *glimpsing* theory (Cooke, 2003, 2006). Glimpsing is a more familiar term in vision. We can recognize an object based on fragmentary evidence. Similarly, when noisy speech is portrayed in the time-frequency domain, as in a spectrogram there is potential to extract the speech from the noise even when large parts of the speech are masked by noise. In the glimpsing approach it is postulated that we listen selectively to the instances in noisy speech that have better signal-to-noise ratio (SNR).

Cooke (2006) tested the glimpsing theory of speech perception by comparing scores for vowel-consonant-vowel (VCV) (e.g. /aga/) word recognition in babble modulated noise [1] with different glimpsing areas. The glimpsing areas can be controlled by adding different babble noises, which have different numbers of talkers. His results showed a high correlation ($r = 0.955$) between the glimpse area and speech recognition scores for normal hearing subjects. Thus, the actual performance of listeners was well described by the glimpsing model, providing strong support for the principles behind the model: the glimpsing area is important for speech recognition and normal hearing subjects can take advantage of these areas.

The glimpsing model is based on the facts (1) that the speech signal itself is sparsely distributed in the time and frequency domain, since the sound signal is highly modulated with many silences due to physical constraints on the speech production system, and (2) that the clean speech signal is redundant, which make

[1] The Babble modulated noise is produced by multiplying the envelope of babble waveforms and speech shaped noise. The speech shaped noise was created by processing white noise with a filter whose magnitude response was equal to the long-term magnitude spectrum of the entire set of sentences from TIMIT database (Simpson & Cooke, 2005).

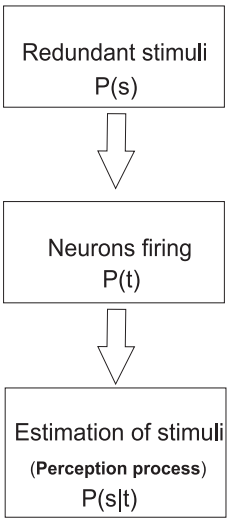


Figure 1.8: *Neurons explore the redundancy of the environment. Stimuli from the natural world are redundant and in order to perceive the stimuli, the brain needs to estimate the stimuli based on neuron firing patterns. Bayes's theorem could be used to derive P(s|t); Rieke et al. (1999).*

the signal more robust in noisy or adverse environments (i.e., the information of speech can be recognised from only parts of its physical representation).

As Cooke suggested, glimpsing is based on the sparseness and redundancy of speech. It is the sparseness of speech that makes the glimpsing possible. And it is the redundancy of speech that makes us able to understand speech through glimsping areas, discarding the non-glimpsing areas.

The glimpsing theory clearly suggested that the speech is redundant and the amount of glimpsing areas can influence our speech recognition in noise.

1.7 Redundancy exploration via neurons

Redundancy is an important characteristic of speech. But how does our auditory system explore it efficiently to facilitate our daily speech communication? A look into how our sensory neurons fire given the redundant stimuli might give the answer. Our perception is largely based on the information provided by the neurons (see Fig. 1.8). As Barlow (1972) put it :

> *"Perception corresponds to the activity of a small selection from the very numerous high-level neurons, each of which corresponds to a pattern of external events of the order of complexity of the events symbolized by a word."*

The neurons must be able to explore the redundancy of the environment, as we know redundancy is important for speech communication and understanding. The redundancy of the stimuli will have important effects on the sensory neuron firing patterns. Our perception is strictly an estimating process of the real stimuli based on the firing patterns of the neurons, and hence is fundamentally statistical rather than deterministic. So exploring redundancy with statistical methods might be an efficient way to investigate the auditory system.

1.7.1 Barlow's redundancy exploration

Barlow (1959, 1989, 2001) investigated how neurons explore the redundancy of environments. His work holds the same view as Mach (1886), Pearson (1892), Helmholtz(1925) and Craik(1943), that the statistics of the sensory stimuli are important for perception and cognition. He also observed that many sensory neurons at later stages of processing are generally less active (Barlow, 1972):

> *"The sensory system is organized to achieve as complete a representation of the sensory stimulus as possible with the minimum number of active neurons."*

Neurons can compress information into a channel with reduced capacity. And this compression is necessary when the channel capacity is limited. Atick (1992) pointed out that there could be an information bottleneck along a sensory pathway. In order to fit the huge dynamic range of the input data, the nervous system must have to perform data compression. One such compression strategy is redundancy reduction (Attneave, 1954; Barlow, 1961). The sensory system is organized to achieve as complete a representation of sensory stimuli as possible with the minimum number of active neurons.

The idea of redundancy reduction focuses the ideas of 'economic' thought, using as few neurons as possible to have a complete representation of the sensory

stimulus. And later Barlow revised his ideas by stating that the redundancy reduction idea needs to be changed to redundancy exploration, as compression is not ideal for the brain to analyze *compact coding* (Barlow, 2001; Field, 1994). A compact code performs a transform that represents the input with a reduced number of vectors with minimal RMS (Root-Mean-Square) error (Field, 1994). Compact coding needs a high active ratio of neurons (see Fig. 1.10), and a high active ratio of neurons is not biologicaly efficient (Lennie, 2003) [1].

As Barlow (2001) later pointed out:

> *"It is the knowledge and recognition of the redundancy that is important, not its redundancy. Because once we recognize the redundancy, our encoding will be much simplified, focusing more on the non-regular parts."*

1.7.2 Sparse coding

Our sensory systems are continuously stimulated by outside events but we are usually not aware of this because only a tiny fraction of the information reaches our consciousness. A common feature of neuronal sensory systems is that they reduce the amount of redundant information in successive processing stages. In other words, sensory systems filters relevant information hierarchically so that higher processing stages receive more relevant information. This reduction of redundancy is an efficient coding strategy to maximize the information conveyed to the brain, without consuming excessive neural resources and without overloading the brain with excessive input (Barlow, 1959, 2001, 1972).

Sparseness can be conceptualized differently according to context. In neural systems, sparseness is often described by referring to neuronal activity: only a few neurons out of many are active at the same time (population sparse), or over the course of time, each neuron will be rarely active (life sparse) (Olshausen & Field, 2004; Olshausen & O'Connor, 2002). Population sparse is used in this thesis, referring to the distribution of samples in a population. A population is considered

[1] However, when there is a bottleneck along the transmission path, the redundancy reduction is necessary and could be optimized to transfer the maximal information without occupying too much bandwidth

sparse if samples are scattered or its distribution is super gaussian; having more values in the centre than in the tails compared to a normal distribution.

Sparse coding strategy for neurons takes advantage of redundancy in the environment. It states information is represented by a relatively small number of simultaneously active neurons out of a large population. This coding scheme is also called efficient coding. Compared with compact coding (see Fig. 1.10), sparse coding emphasizes the sparseness of the firing patterns. And sparseness here can be measured using kurtosis (Field, 1987):

$$k = \frac{1}{n}\sum_{i=1}^{n}\frac{(r_i - \mu)^4}{\sigma^4} - 3 \tag{1.1}$$

where k is kurtosis, r is the amplitude of signal, μ is the mean and σ is the standard deviation. For a gaussian (non-sparse) distribution $k = 0$, whereas for non-gaussian signals the kurtosis may be super-gaussian ($k > 0$) or sub-gaussian ($k < 0$).

Fig. 1.9 shows three different levels of sparseness. Kurtosis can be seen as an indicator of how sparse a signal can be. Compared to a gaussian signal (kurtosis=0) with equal variance, a signal with higher kurtosis is more sparse than that of low kurtosis. (A gaussian signal has maximum entropy and no redundancy.)

Sparse coding has been identified as an important encoding principle for sensory neurons to encode environment stimuli. It has also been found that natural scenes are highly structured and can be modelled (to a first approximation) as a sparse collection of local features (e.g. edges) (Bell & Sejnowski, 1996, 1997).

Many advantages have been proposed for sparse coding (Olshausen & Field, 2004). Sparse coding:

1. is useful for forming associations for memory;
2. makes the structure of the input more explicit;
3. gives a simple representation;
4. is energy efficient.

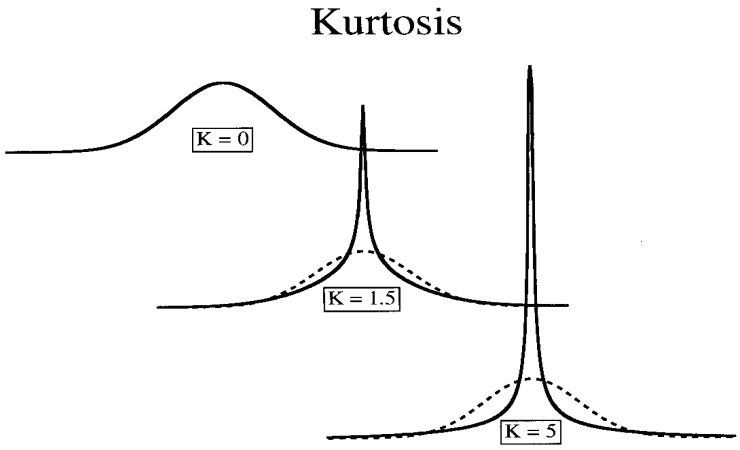


Figure 1.9: *Examples of three levels of kurtosis representing neural response patterns (Redrawn from Field (1994)). In a signal with more sparse distribution, the probability distribution is more peaky with a higher kurtosis. Each of the distributions has the same variance. With sparse input, the neuron response pattern will also be sparse and most of time the neurons are likely to be in a state of rest.*

The relationship between sparse firing of neurons and the statistics of natural scenes has been intensively investigates through computational models. With natural stimuli as training data (e.g. natural sounds or images), the optimization rules are to represent the image or sound as sparsely as possible, and the characteristics of the transforms are very much like the receptive field of sensory neurons (Bell & Sejnowski, 1997; Olshausen & Field, 1996). Modern information theory optimization algorithms were used to sparsely analyze natural images. They obtained decompositions which are very similar to the receptive field of visual cortex (V1).

Lewicki (2000, 2002) tried a similar idea on auditory modelling. He derived optimized filters based on the statistics of natural sounds. The method is based on the assumption that the auditory filter evolves and is shaped by the statistics of the environment. Lewicki was able to find a time-frequency analysis which maximizes statistical independence among the outputs of filters. The characteristics of these filters, also called kernal functions, resemble many characteristic of cat auditory nerve fibres, and it also bears similarity to the auditory filters that have

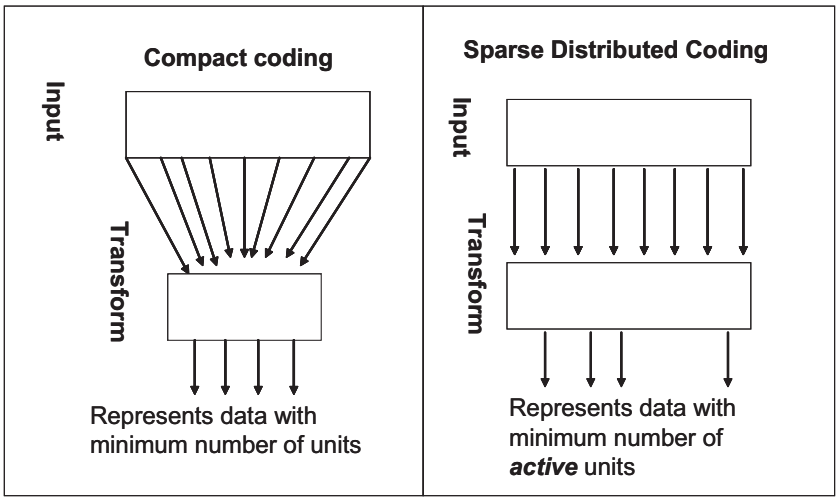


Figure 1.10: *Compact coding and sparse coding (Redrawn from Field (1994)). The figure shows the difference between compact coding and sparse coding. Compact coding shows that the dimensionality has been reduced, allowing the input to be represented with a minimum number of active cells. For sparse coding, the dimensionality is not reduced. Overall, the response probability for any particular cell is relatively low (Rieke et al., 1999).*

been characterized psychophysically in humans and other animals. This suggests that independence or sparse coding could appear at a much earlier stage (Peripheral auditory filters) for auditory encoding than the encoding in vision (Cortex) (Olshausen & O'Connor, 2002). Olshausen & O'Connor (2002) suggested that the sparse coding would most probably happen at the point of expansion in the representation: The visual system has a early bottleneck, where information from more than 100 million photoreceptors is funneled into 1 million optic nerve fibers. The representation is then expanded by a factor of 50 in the cortex, where sparse coding has been observed. For auditory system, The 3000 inner hair cells of the cochlea immediately expand to 30,000 auditiory nerve fibers. And such expansion may provide physical neurobiological structure bases for sparse coding in the earlier periphery auditory filtering.

1.7.3 Redundancy and higher order statistics

From redundancy exploration to sparse encoding of natural stimuli, the key concept is that our perceptual system has been shaped by environmental statistics. This shaping can be considered to occur on different time scales: long term evolutionary changes, during development in infancy and modified throughout life, or even on a moment by moment basis as the sound environment changes. And the statistics of the sensory stimuli we receive everyday from the environment are important for perception and cognition. This idea is certainly not new. The statistics of the sensory stimuli have long been recognized to be important for perception and cognition (Helmholtz, 1925; Match, 1886; Pearson, 1892). The common ground for redundancy reduction and sparse coding is that the perceptual system is an information processing system. Our perceptual system can be investigated by exploring the statistical structure of the environment. Smaragdis (2001) proposed the redundancy reduction principle could be an unifying principle for computational audition. And he also showed that it is possible to explain perceptual grouping by investigating the statistics of mixed sounds through redundancy exploration ideas.

Since statistical structure of the input stimuli is so important, it would be interesting to test the relationship between environmental statistics and our per-

ceptual performance. In order to explore this relationship, high order statistics have to be used. Although second order correlations can provide a good view of spectrum analysis [1], high order statistics have to be used to explore inter-dependency. In order to make the redundancy more explicit, high order statistics can lead to factorial coding like transformation, which can make hidden factors more explicit. Redundancy can also be viewed as a form of dependency. To remove the dependency, high order statistics have to be calculated (Hyvarinen & Oja, 2001; Stone, 1993).

As it is clear that environmental statistics have a big impact on perception and neuron coding, it would be interesting to see how the high order statistics of the environment affect the speech recognition tasks via psycho-acoustical experiments and how we can apply this concept to help cochlear implant users.

1.8 Thesis outline

This chapter has introduced the concepts of redundancy reduction, sparseness and glimpsing theory for speech perception in noise. The goals of this thesis are multiple. First is to bring high order statistics to the field of speech perception research. This is quite similar to the earlier computational vision research. Second is to develop speech processing algorithms based on sparseness coding. The useful of sparseness provides a fundamental basis for the concepts of the new algorithms.

Cochlear implants could also be an ideal test ground for the sparse coding principle. The electrodes of cochlear implants stimulate auditory neurons with electrical pulses. If sparse coding principles hold for auditory neurons, the design of electrical stimuli should consider the sparse firing property of neurons. As the neuron firing is highly synchronized with the electrical stimuli (Hartmann *et al.*, 1984; Kiang & Moxon, 1972), the stimuli should be designed to be sparse in order to get sparse firing patterns.

Also sparse coding based speech enhancement algorithms have been proposed (Liu *et al.*, 2005; Potamitis *et al.*, 2001), but these algorithms were not specially

[1] Fourier analysis of correlation is the same as spectrum analysis

designed for cochlear implants. The experiments on the enhanced speech processing algorithm based on sparseness may not only help cochlear implant users get better performance but also provide an alternative solution for the design of electrical stimuli for cochlear implants. The sparseness characteristic of the stimuli may be important for speech perception for cochlear implant users.

Chapter 2 shows sparseness has an important effect on speech perception for normal hearing through psycho-acoustical experiments.

Chapter 3 introduces a macroscopic approach for the research. Mathematical tools, PCA and ICA, are introduced and they can reduce redundancy of speech and make it sparse.

Chapter 4 uses PCA and ICA related methods to make noisy signal mixtures more sparse. The improvement of speech recognition is observed by testing subjects listening to speech materials with increasing sparseness.

Chapter 5 applies the idea of PCA and ICA to speech analysis and shows some examples of the analysis. The transformation is studied in detail.

Chapter 6 proposes a speech processing algorithm, SPARSE, based on sparse coding principles.

Chapter 7 shows the results of subjective experiments with the normal hearing and cochlear implant users.

Chapter 8 is the conclusion and discussion of potential further research.

Chapter 2

Sparseness and glimpsing

2.1 Introduction

The sparse feature of neuron coding can be described by the fourth-order statistic, kurtosis (Olshausen & Field, 2004; Willmore & Tolhurst, 2001). Kurtosis has also been used in blind source separation (LeBlanc & Len, 1998), extracting sources from mixtures of speech.

The high order statistics of speech are quite different from those of gaussian noise. Robust voice activity detection algorithms were developed based on this property (Li *et al.*, 2005; Nemer & Goubran, 2001), and optimal filter design (Nemer *et al.*, 1999). Here we use the kurtosis to measure the sparseness of sounds. It is assumed that the kurtosis of speech is higher than gaussian noise.

The experiment described in this chapter investigates the relationship between sparseness and glimpsing areas of speech, where sparseness is measured by kurtosis. The assumption is that mixtures of speech and noise with large kurtosis (more sparse) should have greater opportunities for glimpsing and will be easier to understand than the mixtures with kurtosis closer to zero (less sparse). In general terms, kurtosis would indicate the effective signal-to-noise ratio (SNR) and so the chances of glimpsing. Cooke (2006) showed that the glimpsing area is highly correlated with speech recognition in noise using (Vowel-Consonant-Vowel) VCV words. This chapter calculates the kurtosis for the same speech materials

and compares these values with the corresponding glimpsing area [1] calculated by Cooke. A high correlation between glimpsing area and kurtosis therefore implies that the kurtosis is a good quantification parameter for glimpsing and it can be used as a good indicator for speech recognition in noise.

2.2 Methods

Sparse signals such as speech necessarily have distributions with more extreme peaks than gaussian signals, due to the intermittency of production. A standard method to quantify sparseness is to use kurtosis, the 4th moment of the signal, as defined in Eq. (2.1).

$$K = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \mu)^4}{\sigma^4} - 3 \tag{2.1}$$

with x is the amplitude of signal, μ is the mean and σ is the standard deviation. For a normalized gaussian (non-sparse) distribution with $\mu = 0$ and $\sigma = 1$ the kurtosis is (by definition) $K = 0$, for other signals the kurtosis may be super-gaussian ($K > 0$) or sub-gaussian ($K < 0$).

Fig. 2.1 shows two examples of using kurtosis to quantify the sparseness of two signals. The top right panel shows the time course of random gaussian noise. Underneath, a histogram of the signal is illustrated; on the left is the waveform of speech sound /aga/ and its histogram. The kurtosis of the speech sound is higher (K=7.9) and the distribution is more peaky than that of gaussian noise (K=0.14). Note the different scale on the y-axis and that it is not the spread of the signal in the histograms, which would be represented by the second order moment (standard deviation), but the sharpness of the peak that is represented by kurtosis.

[1] The glimpsing area can be calculated based on local SNR. It can be defined as an area with a higher SNR. Listeners are assumed to be able to understand speech by taking advantages of these glimpsing areas in noisy enviroment

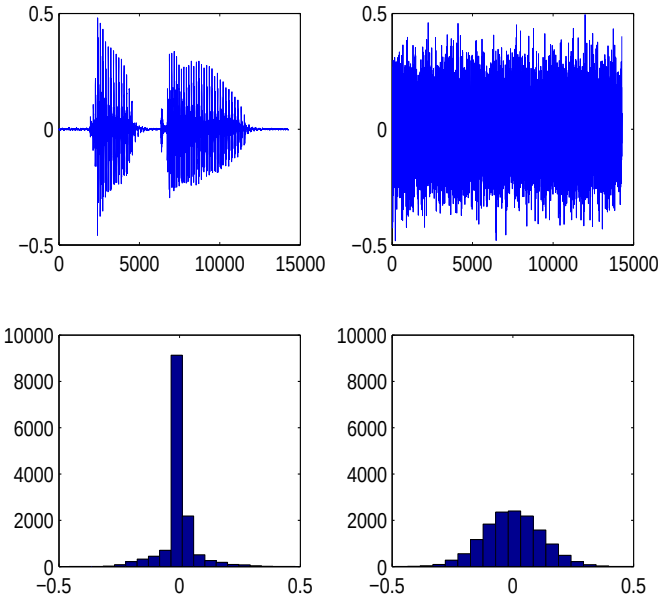


Figure 2.1: *Speech has higher kurtosis than gaussian noise. The waveform (top left panel) and its corresponding histogram (lower left) are shown for the speech sound /ada/ with kurtosis=7.9 and gaussian noise (right, kurtosis=0.14).*

2.3 Speech materials

VCV words were used (Shannon *et al.*, 1999). They comprised the sixteen consonants (b, d, g, p, t, k, m, n, l, r, f, v, s, z, sh, tch) in the context of vowel /a/. The total test set included 160 items from five male talkers and two examples of each talker were used. The babble-modulated noise conditions were created by multiplying speech shaped noise with the long-term magnitude spectrum of the The Texas Instruments/Massachusetts Institute of Technology (TIMIT) corpus (Garofolo *et al.*, 1993) for twelve different numbers of talkers (N = 1, 2, 3, 4, 6, 8, 16, 32, 64, 128, 512, ∞). The final noisy speech tokens were obtained by summing the speech items and the 12 noise conditions at a global SNR of −6 dB. Note that the SNR in terms of signal power was the same in all conditions.

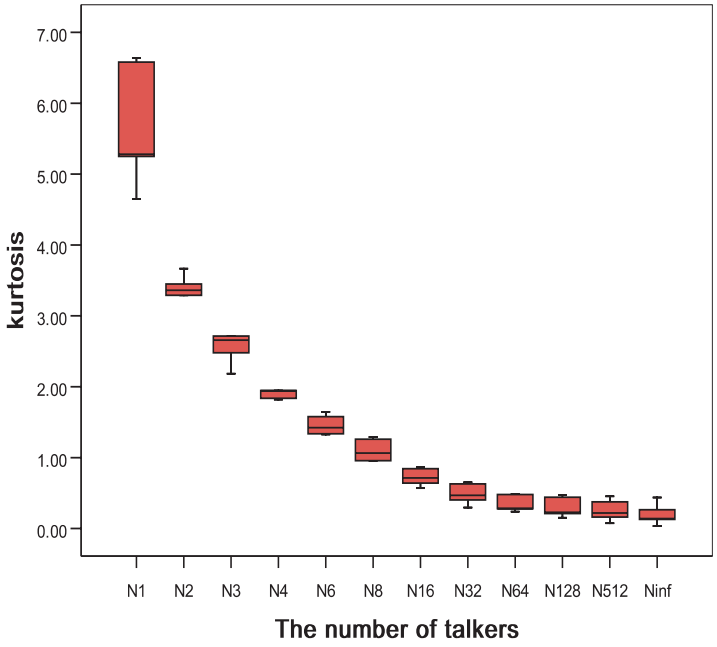


Figure 2.2: *Distribution of kurtosis values for the noisy speech tokens used in experiment 1 as a function of the number of talkers. The boxes include 95% of the distribution, the error bars include the maximum and the minimum. The median is indicated by the line inside the box. Kurtosis value decreases as the number of talkers increases.*

2.4 Results of quantitative analysis of kurtosis

The kurtosis of each noisy speech token was calculated as a function of number of talkers. The kurtosis was calculated in the time domain for each stimulus using Eq. (2.1). Fig. 2.2 demonstrates that the kurtosis of the noisy speech tokens decreases continuously with increase in the number of talkers. Note that the range of kurtosis among samples is greater when there are fewer talkers. For babble-modulated noise, conditions $N = 1, 2$ differed significantly from each other and all other conditions ($P < 0.05$). No condition with $N > 16$ is significantly different from speech-shaped noise. Pairs $N = 3, 4$, $N = 4, 6$, $N = 8, 6$ are not significantly different from each other. With increasing N, the distribution approaches a gaussian distribution, as indicated by the decrease of kurtosis and

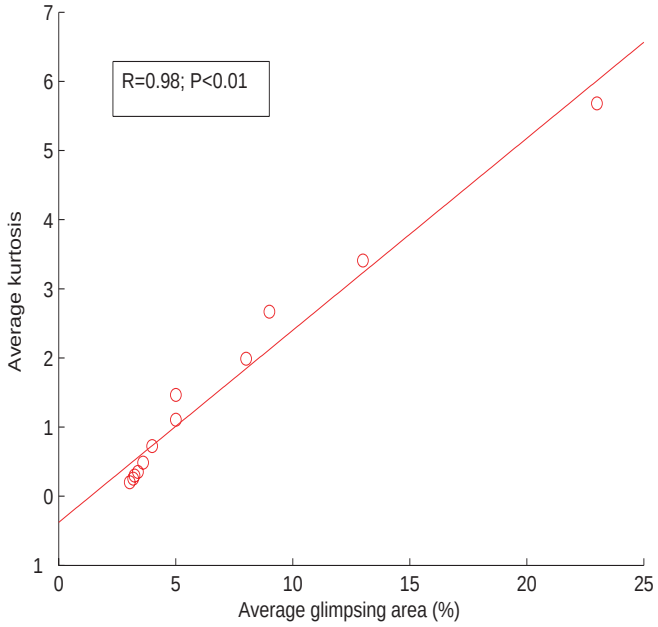


Figure 2.3: *Correlation between kurtosis and glimpsing area. Kurtosis and glimpse area are highly correlated (averaged across all tokens and all subjects), showing that the kurtosis can work as a good indicator for glimpsing areas.*

the kurtosis nears zero.

Fig. 2.3 shows the correlation between the percentage mean glimpse area and the corresponding kurtosis on average. The correlation between average kurtosis and glimpsing area is very high ($r = 0.98$; $p < 0.01$). This shows not only that kurtosis can be used as a good indicator for glimpsing areas but also provide a new method to investigate psychoacoustic perception. It also provides evidence that sparseness of speech data is important for speech perception. Cooke suggested to use the glimpse area as a predictor for speech recognition in babble modulated noise and he demonstrated a high correlation between glimpse area and speech recognition score of normal hearing subjects ($r = 0.955$) (Cooke, 2006). Therefore it follows that there should be a high correlation between the kurtosis and the recognition score results of Cooke (Cooke, 2006), which is shown in Fig. 2.4.

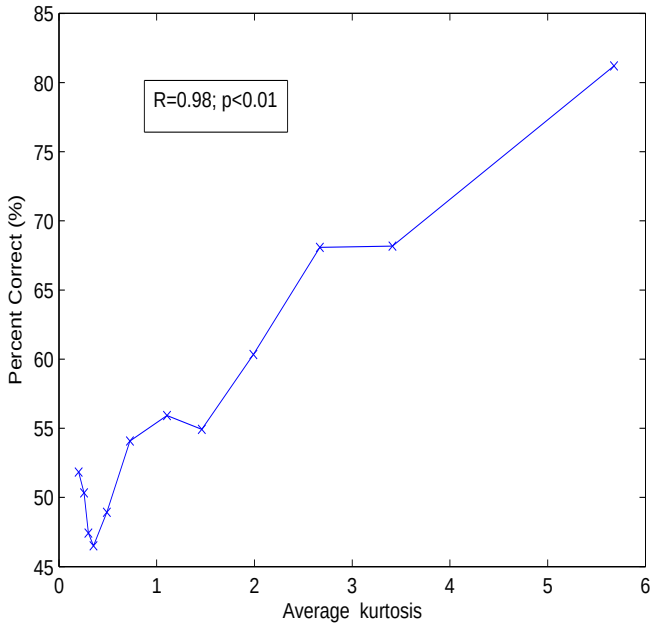


Figure 2.4: *The correlation between kurtosis and speech recognition score with modulated babble noise.*

2.5 Discussion

The present experiment has shown that the 4th order signal statistic, kurtosis, is a valid method of measuring sparseness and that it is a good predictor for consonant perception in babble noise.

Determining kurtosis as a measure of sparseness with Eq. (2.1) has an advantage over other methods such as glimpsing area: as the calculation is simple and computationally efficient.

However, several issues must be considered. The kurtosis based calculation of sparseness is very sensitive to outliers. A small number of outliers might create significant changes in the overall kurtosis. Secondly, kurtosis measures how near the signals' distribution is to the gaussian distribution. The glimpsing areas used by Cooke is controlled by adding different numbers of talkers. As the central limit theorem states, if more sources sum up, they become more like a gaussian signal. And the kurtosis will be deemed to be smaller (close to zero). So the decline of

kurtosis could be explained by a more gaussian like distribution.

However, such explanation cannot exclude the possibility that speech perception itself is affected by the higher order statistical distribution of the data. The experiment shows that a signal with more gaussian like distribution could be harder to be understood as it has many different sources. These sources may interact with each other and the glimpsing ability of normal hearing will be significantly decreased.

Thirdly, we calculated the kurtosis only in the time domain and we averaged the kurtosis over each whole VCV item. Even the simple time-domain kurtosis used here already shows high correlation between sparseness and speech recognition. In addition to the time domain, kurtosis could also be calculated in the frequency domain or as a combination in the time-frequency plane. Such calculation has a closer resemblance to the glimpsing area that is also calculated in the time-frequency plane. Areas with high spectro-temproal kurtosis would be correlated with high glimpsing areas. Kurtosis measurement in the time-frequency domain has previously been used as a SNR estimator for sub-bands (Nemer *et al.*, 1999). A modified computation of kurtosis, in which kurtosis is calculated separately in different temporal and spectral regions, could be used for more complex signals. This might provide a better predictor for speech recognition, because it takes into account the time-frequency domain overlap of competing signals. A simple improvement for the kurtosis measure as it is used here is to calculate a dynamic representation using a sliding window, analogous to a short time window Fourier Transformation. This dynamic representation would make it an attractive option for real-time applications that require an estimate of sparseness or prediction of speech recognition score at each moment in time.

The experiment has shown that glimpsing has a high correlation with sparseness. Glimpsing area is a good predictor for speech perception in babble modulated noise. The high correlation between sparseness and glimpsing also shows that sparseness itself is also a good predictor for speech perception performance.

The experiment is different from other psycho-acoustical experiment in that it analyses the speech as a whole and investigates its statistical properties. This approach can be thought of as a macroscopic approach and it considers the speech

item as a whole, rather then different cues or elements which affect speech perception in general such as formants or envelopes. Chapter 3 will give a brief introduction of this macroscopic approach for speech perception research.

Chapter 3

Exploring redundancy: a macroscopic approach for hearing research

3.1 Introduction

Our auditory system is able to explore the redundancy of speech. The description of our auditory system has been largely based on a microscopic approach by controlling different parameters of sound: onsets, loudness, frequency or amplitude modulation, formant transitions, resolution of auditory filters. Most stimuli are abstract forms of natural stimuli. Although these experiments using simple stimuli provided fundamental knowledge about hearing and speech perception, they fail to explain how we can deal with complex sound in daily life. Our ears are made to handle the situations in natural conditions, where redundancy is a key element of the environment. When we only present simple stimuli, the carrier of the information is quite different from the natural signal, which is highly structured and with redundancy. We might have over-simplified the characteristics of the auditory system from environments based on simple stimuli. When a system is complex itself, such as the auditory system, the systematic view is not only necessary but also critical to understand the behaviour of the system as a whole.

A macroscopic approach is needed to investigate the auditory communication system. With the systematic view, the stimuli used could be quite different

from the 'lab stimuli', which are pure and abstract. The investigation of the auditory system based on the macroscopic approach could reveal some important properties of the auditory system which are not seen otherwise. Such an approach also provides mathematical tools to describe and investigate the auditory system. Our auditory system can be thought as an efficient projector: it projects the input into a different parameter space, where different characteristics of audio signals become more obvious. Hidden underlying factors are revealed and at the same time the origins of the structured signal may be inferred. These projections can be simulated or approximated by modern signal processing techniques such as Principal Component Analysis (PCA) or Independent Component Analysis (ICA). This systematic view of the auditory system provides a new rationale to design psycho-acoustical experiments.

3.2 Microscopic approach for hearing research

3.2.1 Traditional approach for hearing research

The main object of hearing research is to have a better understanding of how sounds presented to the ear are encoded and understood by the hearing process. Such research requires a multi-disciplinary approach to understand the system, including Audiology, Neurophysiology, Psychology and Engineering. Here we focus on psycho-acoustical experiments. Different audio stimuli are designed to answer different research questions, such as what is the minimum level of pure tone humans can hear? What is the lowest (or highest) frequency humans can perceive? Different characteristics of stimuli also reflect different stages of hearing research according to Plomp (2002). Four stages were described by Plomp as listed below.

- The dominance of sinusoidal tones as stimuli

  This is mainly due to the legacy of the successes of physics in the 19th century. It provides critical information about auditory filters and valuable information about the frequency-resolving power and characteristic frequency of auditory neurons. The research based on pure tone stimuli

largely builds on the idea that the ear's frequency resolution mechanism could be compared with a series of band pass filters (Bekesy, 1948). However, such abstract sinusoidal stimuli are far from the real sounds to which we are exposed everyday. It, to some extent, provides some basis of our understanding of the perception of practical sounds.

- The predilection for a microscopic approach

  This is the key argument of Plomp (2002). The microscopic approach proposed to study a complex system by concentrating manageable small subsystems. This is contrary to the macroscopic approach, which take the system as a whole to study. Speech is divided into the smallest units, phonemes. Sounds are processed into different frequency elements: formants with high energy in certain frequency area and its transitions; also other basic sound elements such as onset, modulations depth and so on.

  These different subunits are organized within different stimuli. Some important principles were found based on this research. For example, vowel perception can mostly be explained through the first and second formants (Yang, 1998). Consonant perception is affected by the transitions of the formants (Assmann, 1995). Onset has an important impact on voicing (Liberman & Cooper, 1958).

  The drawback of this research is that important global phenomena, such as the cocktail party effect, remained unexplained. The conclusions based on this approach tends to be over extended.

- Emphasis on the psycho-acoustical aspects of hearing

  The main aim of this part of research focuses on bottom-up processes. The top-down process was not considered in psycho-acoustical experiments. Although cognitive behaviour is very important for speech perception, the top-down processes were not considered fully in this stage.

- Focus on stimuli abstracted from the 'dirty' acoustical conditions of everyday listening.

The sounds used in psychoacoustics experiments are far from natural stimuli. And the experiments with these sounds will not give insights how we perceive sounds in everyday life. The traditional approach described above can be in general seen as a microscopic approach. It equipped us with the basic knowledge of how we understand speech and pure-tones. But it is not enough for us to have a global view of hearing processing system. Lacking of this global view could prevent us from understanding the whole system. For example, the focus should be much broader than the understanding of pure tone perception or masking. And now we are approaching a new stage, where more natural and more realistic sounds are used to evaluate and investigate the auditory system both in psychoacoustic experiments and computational modelling of the auditory system (Lewicki, 2000, 2002).

Interactions or inter-dependency among elements of a signal (which require higher order statistics to represent them) are more important than the individual elements themselves. Therefore, psycho-acoustical experiments using simple stimuli cannot explore these interactions and cannot characterize the most important features of the system. To characterize the auditory system meaningfully the stimuli must be complex to explore the interactions.

3.2.2 Traditional approach to speech perception

Speech perception research has been carried out for more than half a century. The main methods for speech perception research today are still largely based on psychoacoustic experiments. It normally involves subjective responses to different stimuli, whose parameters are controlled through some methods, such as frequency or amplitude modulation. The aim is to find some basic principles or cues that listeners might use to recognize or discriminate different sounds. Based on this approach, we can understand many auditory phenomena such as localization, time-frequency resolution, auditory filtering, and masking. Speech perception research has been focusing on the mapping between physical acoustic properties and linguistic features such as phonemes (Pisoni, 1985). Since the idea of visualizing the speech spectrum and regarding it as a research platform the acoustical spectrum was thought to carry most important information about

speech, and it should be enough to allow correct recognition both for human and machine. However, the lack of acoustic-phonetic invariance has been the issue in automatic speech recognition research. The same phoneme can appear as many different forms of spectrum from different people, in different environments (e.g. noise, reverberation), but humans are able to recognize the speech easily in daily life.

The lack of a one-to-one relationship makes the mapping process become complex and the explanation of how we recognize speech like consonants and vowels also becomes complex. Different levels of features are included for the explanation of speech: formant, formant transitions, onset, voice onset time (VOT), co-articulation due to context. Different theories also derived from these puzzles (Diehl *et al.*, 2003). One example is the motor theory of speech perception, which states we can understand speech because we can articulate speech. It is the neuromotor commands we perceived that make us understand speech. Further computational models of speech perception have been pursued: Computational Auditory Scene Analysis (CASA) is an example.

3.2.3 Computational auditory scene analysis

Based on these experiments and with the good reasoning, computational implementation of perception can be realized. Such computational implementation can provide more detail of the internal processing of perception. However, this kind of research requires that these principles have to be mathematically described. Much of this work was termed computational auditory scene analysis. Unfortunately, many of perception experiments failed to provide such rigorous logical formula, which is necessary for the computational simulation. The psychoacoustic experiments done by Bregman (1990) explain many auditory perception phenomena and represent an important phase for speech perception research. Based on his various experiments, he explains how we group different frequency components into one source, and pointed out some useful principles to analyse the auditory scene. Many computational approaches have been used to simulate speech perception based on his findings (Cooke & Okuno, 1999; Cooke & Brown, 1993). But many of these principles are hard to implement as algorithms or codes, due

to their fuzzy description, such as common onset, or common destination. As Smaragdis (2001) pointed out in his thesis:

> *This has created a research bias in the auditory research community, which has been lagging compared to visual research mainly due to lack of formal definitions and robust formulations.*

One method to get a mathematical description of the perception process is to use information theory. The brain tries to decode the speech information encoded by the auditory neuron spikes. Through evolution, our auditory system has mastered an efficient way to encode audio signals from an informational point of view. And the speech perception research can be seen as the science to investigate the communication between acoustical space and brain interpretation. A systematic view of the auditory system will give us more insight into what our auditory system can achieve and how it can explore the redundancy of audio signals. A similar information processing approach was proposed in experimental psychology (Massaro, 1975).

3.3 Auditory communication system and information theory

We need a language that can describe the hearing process mathematically, which is more strict and can be easily simulated by computer. This computer audition will certainly play an important role for the understanding of speech perception, just like computer vision gives a strong input for the understanding of vision.

Currently, most psycho-acoustical experiments are designed according to the physical characteristic of sounds such as onset, modulation and pitch. Computational audition involves analysis and design of psychoacoustic experiments according to some quantified features of speech, which can be described by simple mathematical algorithms or equations. One possible language for this task is to use information theory, which is heavily based on mathematics. It was evolved in the 1940s and 1950s for electrical engineers to develop practical communication devices. It is a framework for fundamental issues such as efficiency of information

representation and communication channel limitation in communication. Speech perception processing can be understood as a process of communication, which involves coding and decoding processes. It also has communication channel capacity, which will decrease when a normal person loses hearing capacity.

By viewing the auditory system as an acoustical information processing chain, we can look into how information theory can help us to understand this communication process. Speech is encoded by auditory spikes, and these auditory spikes are codes for the speech, The brain will decode the spikes and interpret them as specific meaning. The auditory system is a highly complicated system, and for a complicated system with many factors, a macroscopic approach and stimuli are needed for investigation. As Plomp (2002) stated that:

> *it would be clear that the macroscopic point of view is essential and at least as important as the microscopic approach for explaining the perception of the complex sounds of everyday life.*

An acoustical signal, like a frog call, can be described by many parameters, such as fundamental frequency, amplitude, phase or shapes of envelopes. Most psycho-acoustical experiments are designed to test the effects of these parameters. For example, a speech recognition task can be used in these tests. We can then summarize how these parameters may affect recognition performance. But in the real world, these parameters are not isolated or independent of each other, the natural stimuli are varying continuously along all these dimensions. The disadvantages of such a microscopic approach might be that it never led us to understanding of more complex phenomena in real life. The German physicist Max Plank (1858-1947) pointed out that if we only focus on an isolated molecule as a subsystem and follow its movements, it would be very difficult to discover the rules of ensemble molecules, which are to maximize the entropy (complete disorder), illustrating the second law of thermodynamics. As in physics, both microscopic and macroscopic methods are needed to investigate the whole system. More and more people are coming to realize the complexity of the auditory system, even at the very peripheral level. Here we propose a macroscopic approach based on information theory, which might be able to provide more insight about the auditory system in terms of information processing.

A systematic view of the auditory communication can take speech communication as an example of an information communication system. The speech production system includes vocal folds, oral cavity, and nose. And speech production can be viewed as a stochastic process. The linear prediction model, for example, explains speech in terms of acoustical sources and filters. This production system certainly has some physical limit; for example, it is difficult for us to speak too fast or two slow. It can have different states and thus different audio tokens can be generated. These tokens altogether make up speech. Speech is then transmitted into the air, and our ears are able to pick up the pressure of the air movement, auditory neurons encode the signals into spikes and send them to the brain. Our brain, with certain knowledge of the language and context, then can interpret or estimate the meaning of the speech.

Information theory was mainly developed for electrical engineers to design practical communication devices. However, the theory has been widely used in many areas, such as psychology, neuroscience. The fundamental question that information theory investigated was to measure the efficiency of information representation and how reliable is communication.

The environment around us is a highly structured world. Effects of statistical regularities on the efficiency of information representation was especially picked up by psychology and neuron science. Similarly, the communication system described by Shannon can be a general model of how information should be encoded to make efficient communication. With this model, the information and redundancy can be quantified. Although it is a big concern that information theory might not be appropriate for biological systems because the information which is defined by Shannon may be quite different from the information the biological system is interested in. However, as Fig. 3.1 shows, Shannon's communication model provides a conceptual framework to analyze auditory communication, i.e, a systematic view of a communication system. As suggested by Plomp (2002), the auditory system needs such systematic view to investigate. Shannon's information theory defines two important characteristic for communication. One is information entropy and the other is channel capacity.

Entropy is a measure of the information available, or minimum bits required to encode a signal. Although it is different from our normal understanding of

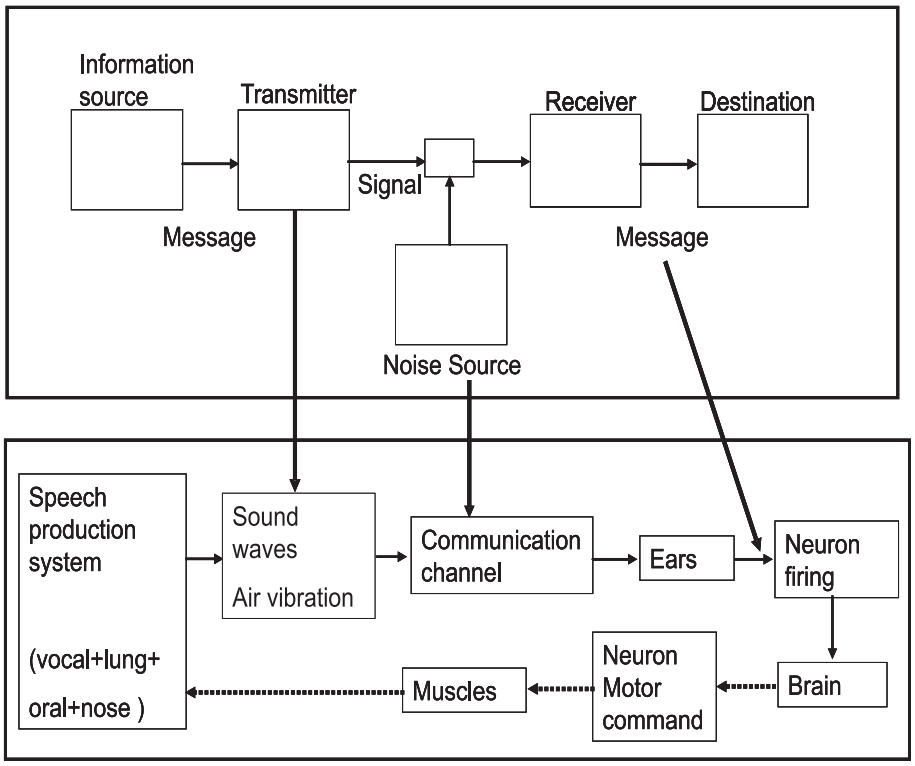


Figure 3.1: *Information theory for communication and auditory communication. The first block shows the communication system described by the information theory of Shannon (1948). Source information is encoded and transmitted through the communication channel, where noise could be added and the signal could be distorted. The receiver then decodes the signal and sends it to the receiver. The lower panel illustrates auditory communication. First, information is coded as speech and transmitted through the air as sound waves; the communication channel could be noisy and adding distortion to the sound waves. Our ears receive the sound signal and decoded as neuron firing patterns, and the brain will interpret the firing patterns as having meaning. The dashed line shows the possible efferent path from brain to motor neurons, preparing to produce speech again.*

information as some sort of meaning, it can be used at least to quantify the physical tokens that carry meaningful information.

$$H = -\sum p(x) log_2 p(x) \tag{3.1}$$

where H is the entropy of the random variables x in bits. $p(x)$ is the probability distribution of x. Entropy describes the uncertainty of an event and how compactly an object can be represented. As can be seen from Eq. 3.1, the entropy of an event is 0 bits when the probability of an event is either 0 or 1. Also the entropy is maximized when all outcomes are equally likely, which means there is great uncertainty, and more bits are needed to code this event.

One interesting finding is that the entropy has a direct link with the distribution of the data. If the data is more gaussian, the signal is more unpredictable. So when noise are added to a clean signal, the entropy will increase since there would be more uncertainty, and the mixed signal would be more gaussian according to the central limit theorem. On the contrary, if a signal is sparse with only a few possible states, the entropy of the signal will be smaller and the uncertainty is much reduced.

Entropy is believed to be always increasing, and a system tends towards disorder if without external input. Instead of asking philosophical questions, we can understand some practical problems with the concepts of entropy. It is known that signal transmission in noisy channels will be difficult. For sound perception, when noise is added to speech, masking is used to explain why it is difficult to hear properly in such conditions. But according to information theory, the effect of noise on our perception can be understood as an entropy increasing process, where the information left for neuron encoding is more difficult due to the increasing uncertainty. To make us hear better, we should reduce the uncertainty and make entropy smaller or make the distribution more non-gaussian.

Another application of information theory is redundancy. The *redundancy* of the signal can also be measured based on information theory. The redundancy is defined by Shannon as the difference in bits between the channel capacity and the source information (Barlow, 2001; Shannon, 1948). Normally the channel capacity is greater than the information. And the channel bandwidth/capacity

can thus be represented as the sum of redundancy and information of the signal (see Fig. 1.6). Thus when communication channel bandwidth is limited, the redundancy of the signal should be reduced so that most of the information can be transmitted within the capacity of the channel bandwidth.

Channel capacity is a big issue for hearing impaired listeners, for example, many signal processing methods for hearing aids and cochlear implant reduce this redundancy of sound in one way or another to suit the limited capacity of hearing impaired listeners. Cochlear implants, for example, only choose a few frequency channels to represent the original acoustical signals. Information theory has been applied in many different disciplines such as psychology and computational neuroscience.

Information theory could bring new techniques and new methods for both computational modelling and psycho-acoustical experiments. Also from a systematic point of view, it may be used to analyze the auditory system.

3.4 Statistics for perception

Information theory is based on the statistical analysis of the data. Our perception actually also can be seen as a statistical analysis of the environment. Our sensory organs collect data from the outside world. On the basis of the data, a model of the external world is constructed (Laming, 1997). The modelling process is an estimation process based on the firing patterns of the sensory neurons (Rieke *et al.*, 1999). The statistics of the environment have an important effect on how our neurons encode the environmental stimuli. Specially, the statistical regularities of the environment make efficient coding possible. If we can define the stimulus space S and perception space P, the perception process is a projection from the elements in S to P. Perception thus can be thought as a data processing and estimation process. The internal processing of perception can be a black box, as we are going to use systematic approach to investigate the system. From an information theory point of view, perception is to transform the information of the external stimuli to the internal perception space with minimum loss of information. The redundancy in the information should be reduced

so that the bandwidth of the communication can be efficiently used. In information theory, principal component analysis (PCA) is a statistical tool which can be used to reduce the redundancy of the data. At the same time, the coding patterns should be independent from each other so that the coding resources can be used efficiently. Independent component analysis (ICA) can be used to implement sparse coding. Both of these two statistical techniques can be understood as linear transformations, which rotate the axes of stimulus space to form a different parameter space. The structure of the stimulus data is much more explicit in the transformed space than in the original stimulus space.

.

3.4.1 State space and perception

Field (1994) proposed that state space can be a good method to describe the redundancy of the natural stimuli. Each element of a stimulus (e.g. pixel amplitude) is representing a dimension of the state space. The state space describes all possible states of the stimuli. For example, to describe the state space of an image set, one can use the pixel amplitude to represent the coordinate axes of the space: for a 256 $\times$ 256 pixel image, the state space of all possible images at this resolution requires a 65,536-dimensional space where the amplitude of each of the pixels represent an axis of the space (Field, 1994). Natural stimuli, such as a natural image, will only occupy a few dimensions in this space. And this limited dimensionality is a key property of natural stimuli, which means that the probability density of any natural scene is highly predictable. Field's state space idea was also used to show the relation between neuronal responses and the characteristics of stimuli (Olshausen & Field, 2004) (see Fig. 3.2). The natural stimuli of the sensory data lie along a continuous curved surface in the high dimensional space of the stimuli. The patterns of the stimuli can be represented by a vector in this state space. And neurons only fire when stimulated by a preferred specific pattern. Normally, the number of neurons is much bigger than the dimensions of the input (e.g pixels of images), which allows for a piecewise representation of the highly curved manifold. It is thus helpful for further analysis by the higher level of neurons.

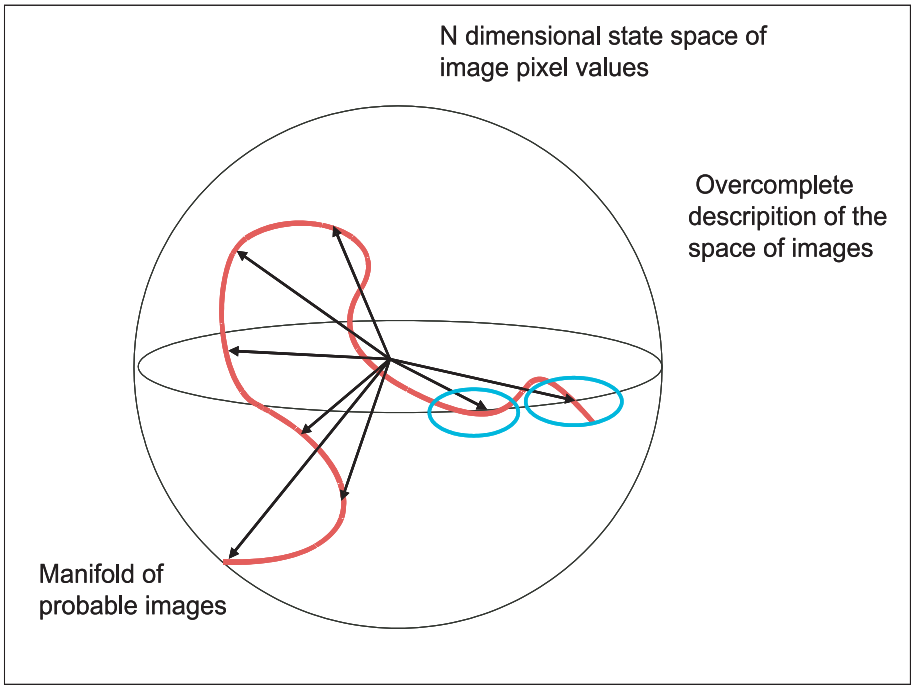


Figure 3.2: *Example of state space and sparse coding by neurons, redrawn from (Olshausen & Field, 2004). Here is an illustration how our neurons might encode the external stimuli in the stimulus state space. N pixels of image can be represented in a N dimensional state space, which covers the spaces of all possible images with N pixels. If an object moves across the pixel array of an image passing by different pixels, it will produce a series of different spatial patterns. Each spatial pattern can be represented by a point in the state space. The natural image is supposed to lie along a low dimensional manifold embedded in this space. The red curve represents a the hypothetical trajectory of an image feature (such as an edge) as it would appear in this space as a result of translating over the pixel array. The black arrow corresponds to the preferred feature of a neuron. The blue ellipses denote the response zone of the neuron.*

Field used the state space to show why sparse coding is needed to represent the natural environment. In the state space, there are many more factors to represent a particular stimulus, but they will not be active at the same time. Such representation will be helpful to represent the highly curved external stimuli in this state space. Here we propose that the perceptual system actually projects this stimulus state space into a new space, where the structure of the data is much more easily observed. For example, mixture of sounds can be seen as separated in the new transformed space. By analogy, in a football match, it requires many video cameras to broadcast the football match. One important reason is that it can show what is going on from different directions. There are usually one or two angles which can perfectly show, for example, who violates the rules. And many judges may not be able to see this from other angles.

Our perceptual system could have developed a powerful computational system that can project or observe what is happening around us. The perfect angle can help us trace back to the driving forces or physical causes of what happened. The redundancy and the structure of the data can be revealed by projection of this state space by appropriate rotation the axes of the space with appropriate scaling. Most importantly, as the natural stimulus state space is highly structured (Barlow, 1961; Olshausen & Field, 1996; Schwartz & Simoncelli, 2001; Smith & Lewicki, 2006), the structure of the data can thus be more appropriately observed in this new state space. Considering so much information is received everyday, our perceptual system must have developed to an efficient strategy to process this information. One method is to transform the data to a much more convenient space so that the structure of the data can be more explicitly observed. Such transformation could be relatively fixed, driven by evolution or experience, or it is possible that the transformation is continuously adjusting according to changing stimulus context. Or it could be a mixture of both.

Obviously, it is hard to prove that our perception is doing an exact rotation or scaling on the stimulus state space. But nerveless, such mathematical analysis is helpful for both understanding our perception and developing possible enhanced signal processing methods to make speech easier to recognise. Field argued that the goal of sensory neuron encoding can be revealed by analyzing the state space of the stimuli and the probabilistic distribution of the firing patterns. He found

that the visual system is near to optimal in representing natural scenes only if optimality is defined by 'sparse distributed' coding. This sparse coding is related to how the stimuli in the stimulus state space are distributed. An appropriate analysis of the stimuli via some statistical technique can reveal the important structure of the data.

If we take perception as a process of recovering the stimuli structure explicitly, the perception process can be seen as transforming state space of stimulus X, to our perception space P.

$$P = WX; \tag{3.2}$$

Where W is the transformation, X is the stimulus and P is perception.

The perception space is to trace the causes of the stimuli, revealing the causes of the structure in the state space.

Once our perceptual system figures out the causes of the stimulus structure, we can then identify the object and get a clear perception on what's going on. This is quite like the speech perception theory called Direct Relative Theory (DRT)

by Fowler (1986), stating that what we perceive are the gestures of the articulation, which are the causes of the structure of the data. Fowler (1996) summarized that

> *" Perceptual systems have a universal function. They constitute the sole means by which animals can know their niches. Moreover, they appear to serve this function in one way: They use structure in the media that has been lawfully caused by events in the environment as information for the events. Even though it is the structure in media (light for vision, skin for touch, air for hearing) that sense organs transduce, it is not the structure in those media that animals perceive. Rather, essentially for their survival, they perceive the components of their niche that caused the structure".*

According to DRT, the structure of the sound is the information medium and analysis of the structure will lead people to the gesture of articulation (e.g. the

closing and opening of the lips during the production of /da/). And it is the gesture that causes the structure of the sounds.

Once our perceptual system can identify the causes of the structure of the stimulus state space, we would then be able to identify and recognize the object. In the following sections, three techniques are introduced that can be used to transform the state space into a more useful space.

3.4.2 Principal component analysis

One efficient redundancy reduction method is PCA, or data whitening (Hyvarinen & Oja, 2000; Smaragdis, 2001). It can make a set of variables uncorrelated by applying linear transformations. The processing of PCA can be explained through state space linear transformation. The PCA recovers the causes of the structure of the data by taking advantage of the correlation of the data to reduce redundancy. For example, a two pixel image state space in the horizontal axis is the intensity of Pixel A and the vertical axis represents the intensity of Pixel B. Any two-pixel image is represented as a unique point in the two dimensional state space (Field, 1994) as seen in Fig. 3.3.

Suppose that the pixels are chosen from a normal distribution, the two pixel data are a correlated mixture of gaussian sources. In the state space, the second order correlation redundancy is quite explicit and PCA can rotate the state space with a new coordinate system. We can take advantage of the redundancy of the data (correlation) to re-represent the data as 'Á' and 'B́'. In the new co-ordinate system, or this new space, most of the variance in the data can be represented with only one single vector ('Á'). Thus the state space can be represented with only a subset of the vectors and minimal loss in the RMS error. As in Fig. 3.4, the hidden structure of the left panel can be explained by the right panel. In a sense, the compact representation on the right panel can be thought as the underlying causes of the data in the left panel. Here PCA explores the causes of the data structure through exploring the correlation.

Mathematically, PCA analysis needs to make the two or more random variables uncorrelated. Two random variables x_1 and x_2 are uncorrelated when:

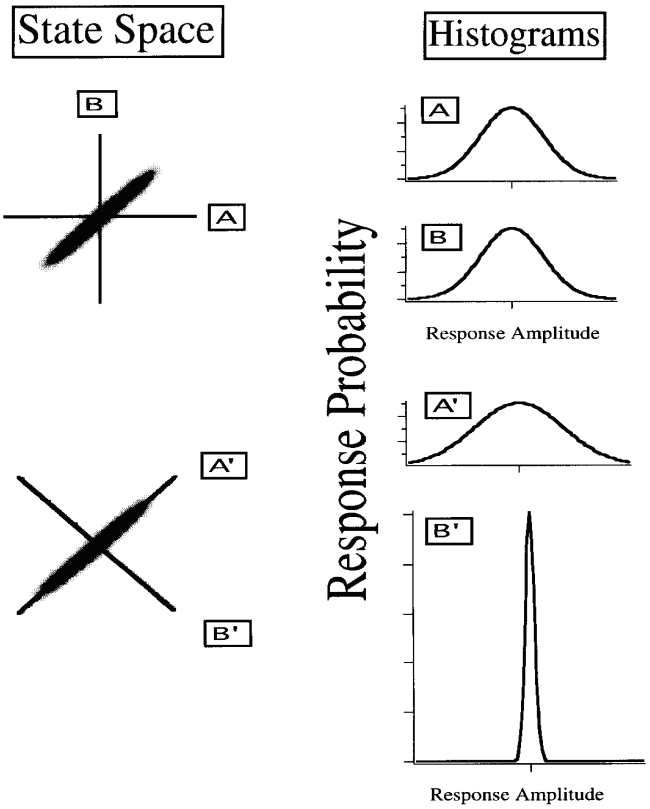


Figure 3.3: *Example of exploring stimulus state space of a two pixel picture from (Field, 1994). Field argues that when the two pixel image set is represented in the transformed space (Á-B́). The neural response probability will also change with the probability distribution of the stimuli. In the original space (A-B), the neuron firing response will be like gaussian as shown in the upper panel, like the stimuli distribution. Once the stimuli are transformed in the new space, where the data align with the principal component axis, the neurons will allocate more resources to the signal with bigger variance(Á). And B́ can even be discarded without affecting the main structure of the data due to its small variance. It can be seen that the neuron response amplitudes are probably zero for B̀. Although this is obvious in the transformed space (Á-B́), it is not explicit in the original space. This is why the state space transformation by PCA is useful.*

$$cov(x_1, x_2) = E(x_1 x_2) - E(x_1)E(x_2) = 0 \tag{3.3}$$

Where *E(x)* is the expected value of x. The measure $cov(x_1, x_2)$ is the covariance of x_1 and x_2. For random variables, the covariance matrix needs to be constructed to test whether the set of data are correlated or not. The covariance matrix contains the covariances between all possible variables pairs.

$$\mathbf{V} = \begin{pmatrix} cov(x_1, x_1) & \cdots & cov(x_1, x_N) \\ \vdots & \ddots & \vdots \\ cov(x_N, x1) & \cdots & cov(x_N, xN) \end{pmatrix}$$

PCA makes a set of variables uncorrelated by diagonalizing the covariance matrix. The off-diagonal elements are the covariances, which will be zeroes if variables x_i are uncorrelated. In mathematical terms, we are looking for a linear transform W

$$Y = W^T X \tag{3.4}$$

W can make the covariance matrix diagonalized, and the vectors in Y uncorrelated. One of methods to get W is to maximize the variance of Y, and keep the Euclidian norm of W equal to one. It is well known from linear algebra that the solution to the PCA problem is given by the unit-length eigenvectors of the covariance matrix of X. Mathematically, to get the solution of W for equation and Eq. 3.5

$$Max(E\{Y^2\}) = E(W^T X)^2 = W^T E\{XX^T\}W = W^T C_x W \tag{3.5}$$

$$||W|| = 1 \tag{3.6}$$

Principal component analysis can identify the direction of maximal variance, which is normally the interesting part of the signal. For example, the direction of eigenvectors of an ellipse will be its short axis and long axis. The steps to get the PCA transform are as follows.

1. We can make the mean of X equal to zero by $X = X - E(X)$;
2. Get the eigenvector of XX^T, W;

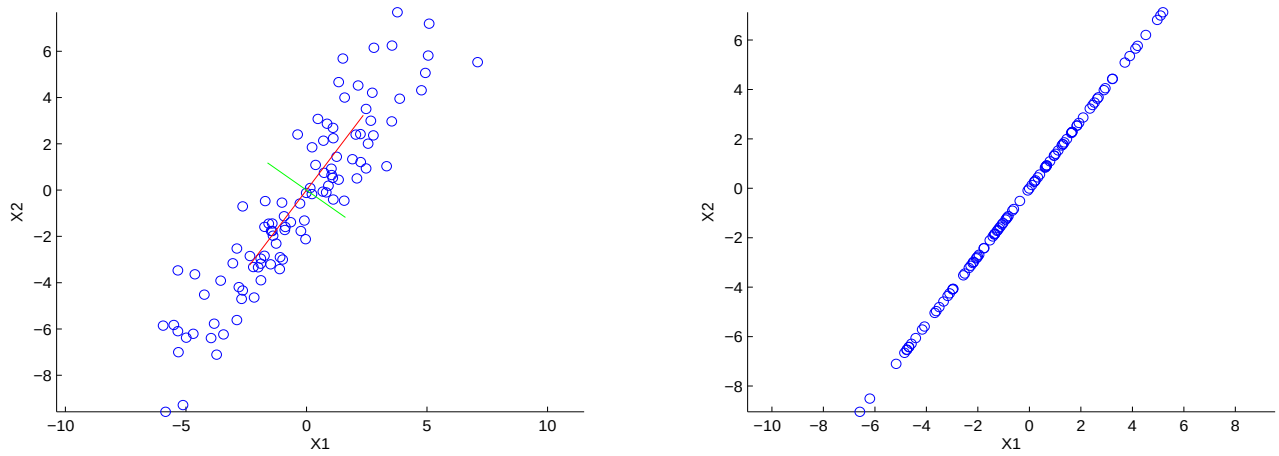


Figure 3.4: *PCA and dimension reduction. The dots are two random signals with correlation and lines are the direction of maximal variances of the data set derived by PCA. The right panel is the result of PCA dimension (redundancy) reduction by removing the smaller principal component.*

3. Transform X by $Y = W^T X$;

4. Reduce the dimensionality of the data according to the magnitude of vectors in Y, in other words, reduce to the largest few vectors;

5. Transform back to $\hat{X}$ by $\hat{X} = WX$ and add back the mean $E(X)W$.

The PCA can be understood by using a 'geometry' explanation. In the transformed space, the directions of signal's maximum variance are orthogonal. Fig. 3.4 shows that PCA can clearly identify the direction of maximal variance of two correlated signal. And the data after dimension reduction by PCA are also shown on the right panel.

3.4.3 Independent component analysis

Principal component analysis can take advantage of correlation based on second order statistics. It can rotate the data in the state space and make it align on the principal axes of the data. And the structure of the data becomes explicit as fewer dimensions are needed to represent the original stimulus state space. However, for

the data like in Fig. 3.5, the data is not correlated and the PCA rotation will not be able to take advantage of the redundancy by discarding one of the principal components, because all the principal components are equally important with the same variance after transformation (Field, 1994; Stone, 1993).

In order to recover the causes of certain data, simple PCA orthogonal rotation is not enough to make the data structure explicit. A much greater restriction on the rotation is needed: independence, as causes of the data structure can be thought of as independent physical sources or events. This pursuit of independence is called Independent Component Analysis (ICA). Detailed review can be found in Hyvarinen & Oja (2001); Smaragdis (2001); Stone (1993).

Independence requires non-correlation of the output in every order of statistics. For non-correlation:

$$E(x_1 x_2) = E(x_1)E(x_2) \tag{3.7}$$

Where *E(x)* is the expected value of x. The covariance of x_1 and x_2 is zero. For independence:

$$E(x_1^p x_2^q) = E(x_1^p)E(x_2^q) \tag{3.8}$$

For a gaussian signal, the output of PCA is independent because gaussian signals are determined by the their second moments. PCA does provide a set of independent signals, but only if these signals are gaussian. ICA can be thought as a method for extracting useful information from the data. The usefulness is defined by the specific application. But it is obvious the information extracted from the data should be the least redundant and this information should be complete to represent features of the original data. For example, give a linear mixture of two sounds, the causes of the mixture are the sources. If we can recover the sources, we can say we identified the hidden factors which cause the mixture. The physical sounds from two different people can be thought as independent since they are two different independent physical processes. Fig. 3.6 and Fig. 3.7 give examples of ICA, which separate mixtures of two sine waves into clean sine waves as shown in Fig. 3.7. Similarly, given some stock prices over one month from the London stock exchange, we would like to extract the

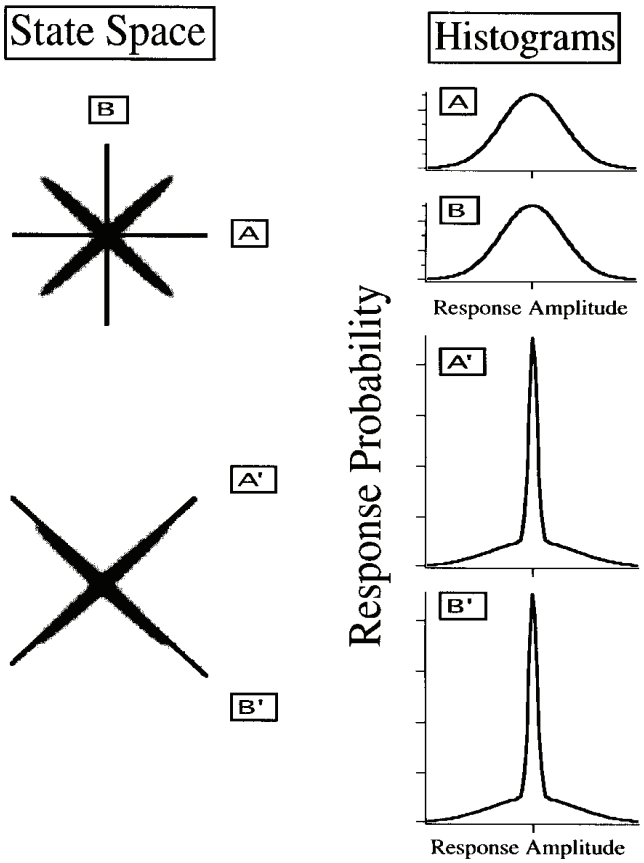


Figure 3.5: *An example of state space transform with PCA, unable to reduce the dimension by using PCA only. It shows that the transformed Á and B́ become sparse than the signal in the original space A and B. Compared to Fig. 3.3, dimension of data Á and B́ cannot be reduced by simply reducing the principal components when reconstructing the signal, as the variance of A' and B' are almost the same.*

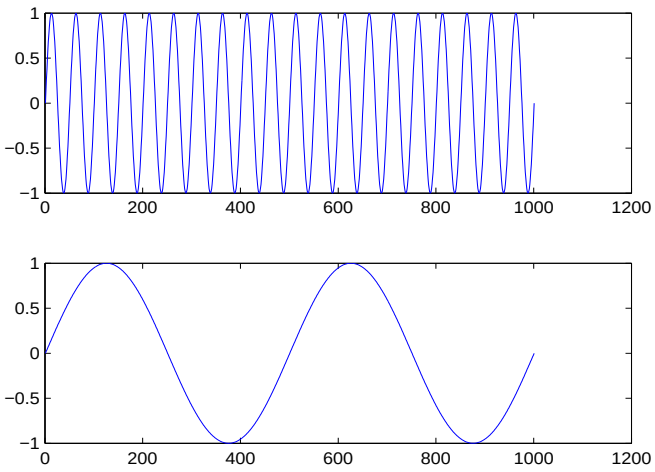


Figure 3.6: *Two signals before mixing.*

underlying factors which affect the prices. These factors could be unemployment rates or any political incident. Once extracted, these factors could even be used to predict further trends of those these stock prices. And similarly, the analysis of the output from neurons can give us some clues about the underlying factors, which would be important to understand the relation between stimulation and responses.

It is also important for perception: how to decompose perceptual inputs into their underlying physical causes (Stone, 1993). If we hear a sound then each auditory neuron has an output which is a function of several physical causes, including frequency, talker emotion, reverberation. Our perceptual system can extract these causes based on the mixed stimuli only, and trace them back to the causes. So ICA is not only a data analysis tool, but also can be thought of as an underlying principle of perception (Barlow, 1961).

The principle of ICA is to make the output of the transform as independent as possible. Independence is the critical idea of ICA. The measurement of independence has been extensively investigated. The independence can be measured based on the minimization of mutual information of the ICA output, which can be calculated by entropy. Hyvarinen and others suggested that to make the outputs independent through minimization of mutual information is the same as to

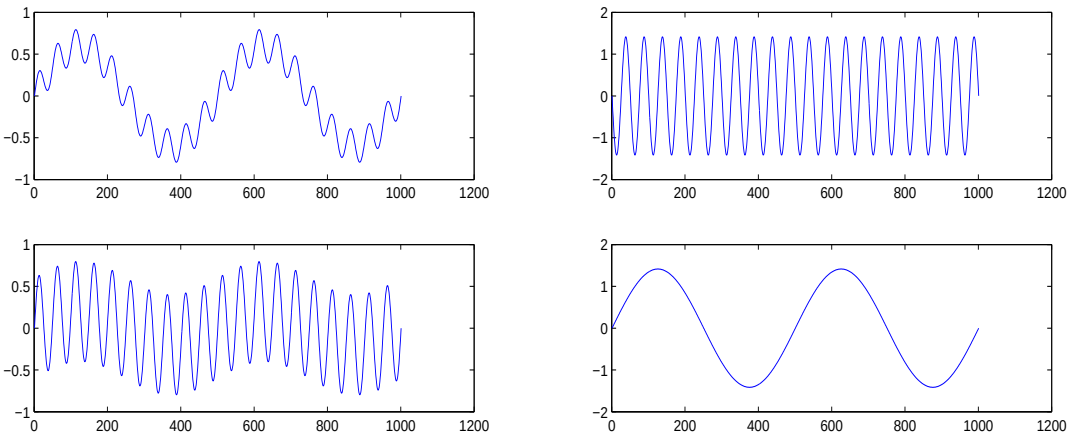


Figure 3.7: *Example of ICA separating two mixtures. The left panels are the two mixtures. And the right panel are the two separated signals after ICA processing. It is seen that the outputs are quite similar with the clean signal in Fig. 3.6.*

make the signal as non-gaussian as possible (Hyvarinen, 1999). A gaussian signal is the least interesting signal for ICA output. According to the central limit theorem, the distribution of a sum of independent random variables tends to a gaussian distribution. So if the extracted signal is as non-gaussian as possible, the extracted signal will be more far away from the mixture. And one extreme of this trend is that the signal will be either super-gaussian or completely sub-gaussian, which will lead to independent components, the sources. Mathematically,

$$X = AS \tag{3.9}$$

X is the mixture by a matrix A and S is the sources we are interested. To get S, if A^{-1} is known,

$$\hat{S} = A^{-1}X = WX \tag{3.10}$$

Since we normally do not know A, and so A^{-1}. we can only estimate the sources S by $\hat{S}$. W is the un-mixing matrix, estimated based on the observed data X. There are various ways to estimate W based on certain features of S. One of the principles to estimate matrix W is to make $\hat{S}$ as independent (or non-gaussian) as possible. Here we assume the sources are non-gaussian, which is true for

most speech signals. The measurement of non-gaussian properties is introduced in detail in the next section. Based on the non-gaussian optimization principle, the un-mixing matrix W is updated:

$$W_{new} = W_{old} + f(W,X) \tag{3.11}$$

where f is a nonlinearity function such as $\tanh(W, X)$ or $\exp(W, X)$, which are used to derive the cost function for W at each iteration to make $\hat{S}$ as non-gaussian as possible.

The non-gaussian property can be quantified in many different ways. One simple idea is to measure the kurtosis of the speech data.

There are several assumptions in this simplified model Eq. 3.9 (Hyvarinen & Oja, 2000; James & Hesse, 2005):

1. The independent components are assumed statistically independent.
2. The independent components must have non-gaussian distribution.
3. The mixing matrix A is square: the number of independent components is equal to the number of observed mixtures. The mixing process is linear.
4. The mixing matrix A stationary.

These assumptions make the ICA model simplified, although in practice, there are many other approaches which can solve the ICA problems without specifying these assumptions (Hyvarinen & Oja, 2000).

Besides these assumptions, there are also some ambiguities of ICA output:

1. The variances or the sign of the independent components cannot be determined.
2. The order of independent components cannot be determined. However, the columns of the mixing matrix A can reveal which detector recorded a particular independent components. And the columns of the mixing matrix are also called 'basis functions'.

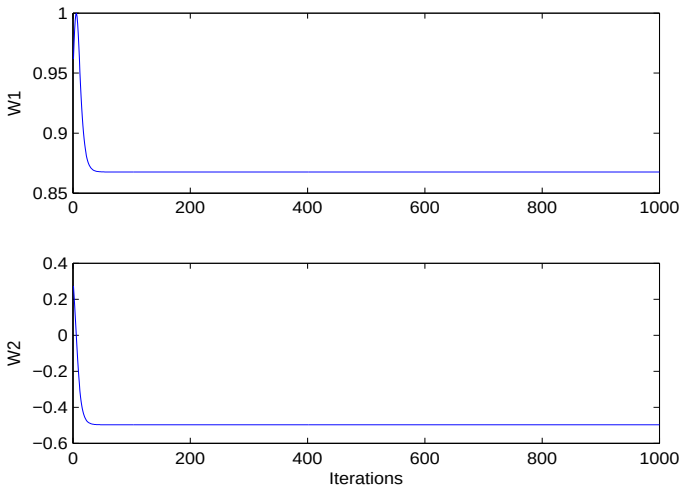


Figure 3.8: *Convergence of project pursuit algorithms. The unmixing matrix becomes stable after a few iterations (around 30). At the same time, the kurtosis reaches maximum.*

3.4.4 Projection pursuit

Projection pursuit is another idea to explore the structure of the data in the stimulus state space. It only get a single independent component one at a time. The basic concept is based on the central limit theorem: the mixture of the data is always supposed to be more gaussian than the individual sources. The causes are non-gaussian and can be measured using kurtosis. A detailed algorithm description can be found in Stone (1993).

The projection pursuit algorithm was introduced originally to separate two sounds by exploring sparseness of the signals, expressed by kurtosis. Here it is used to generate a series of signals with different kurtosis.

First the mixture of signals was preprocessed by Principal Component Analysis, which transforms the mixtures to a new set of mixtures X that are uncorrelated and with unit variance.

The equation for kurtosis of the new mixtures will be:

$$K = E[(W^T X)^4] - 3 \tag{3.12}$$

The gradient of kurtosis for an extracted signal $Y = W^T X$ is:

$$K'(W) = cE[X(W^T X)^3] \tag{3.13}$$

where X is the mixture of signals, W^T is the unmixing matrix, Y is the extracted signal, c is a constant set to 1.

The unmixing matrix is updated by the calculated gradient and the old matrix:

$$W_{new} = W_{old} + K\eta' \tag{3.14}$$

η is the step size of the iteration. The unmixing matrix is normalized before the next iteration:

$$W_{new} = W_{new}/|W_{new}| \tag{3.15}$$

And set:

$$W_{old} = W_{new} \tag{3.16}$$

before repeating the update in Eq. 3.14.

According to equation Eq. 3.14, the kurtosis of the extracted signal in each iteration will be higher than that of signals extracted in the last iteration. Fig. 3.8 shows that the unmixing matrix becomes stable after a few iterations.

3.5 Conclusion

The traditional approach to hearing research is an elementary rather than a systematic approach. Perceptual cognition can be viewed as an information processing scheme involving state space transformation. The structure of environmental stimuli has shaped our perceptual system, making sparse coding of natural scenes possible. Higher order statistics have been a major topic in signal processing in recent years, since they can describe many important features of signals. The PCA and ICA methods have been extended to psychology and computational perception, which can use explicit mathematical expression to simulate the perception process. The state space stimulus transformation approach is a systematic view of our perceptual system with high order statistics in support.

The perceptual system projects the stimuli into a different space, where the causes of the structure can be perceived. Analysis of the state space and viewing

the auditory system as an information processing unit could open new possibilities for hearing research. The concept could well be applied in signal processing for cochlear implants and improve the speech perception for cochlear implant (CI) users. Linear mathematical tools do not necessarily reflect how our perception works. But they provide a possible quantitative research framework to investigate the hearing system.

In next chapter, the relationship between speech perception and the state space transformation will be built, and we can see how the state space approach may influence our view of perception.

Chapter 4

Sparseness and speech perception

4.1 Introduction

Chapter 2 demonstrated that kurtosis as a measure of sparseness co-varies with speech recognition scores. This implies a connection, but is not proof of their interdependency. Therefore an experiment is needed to investigate if speech recognition performance can be improved by simply increasing the kurtosis of noisy speech signals to increase sparseness. Kurtosis has been used widely for source separation, but few psycho-acoustical experiments have investigated the relationship between kurtosis and speech recognition performance.

Direct manipulation of sparseness requires a method to change the kurtosis of a stimulus in a controlled manner. To achieve this, a mathematical algorithm was used to change the kurtosis of noisy signals in an iterative way. Many such algorithms have been developed to separate instantaneous mixtures of two sounds by increasing kurtosis or independence of the output signals. In principle, we would be able to get a series of signals with increased kurtosis by saving the output of such an algorithm after each iteration. We here use a algorithm called the projection pursuit algorithm (Stone, 1993). Projection pursuit refers to the notion that the algorithm extracts the independent components one by one. Fig. 4.1 shows examples of the sound /asa/ in noise with different kurtosis values. The kurtosis of the signals from S1 to S5 increases as a result of sparseness optimization based on the project pursuit algorithm. It is based on two noisy input signals. A series of signals such as S1, S2, and S5, with increasing kurtosis, can be generated. We

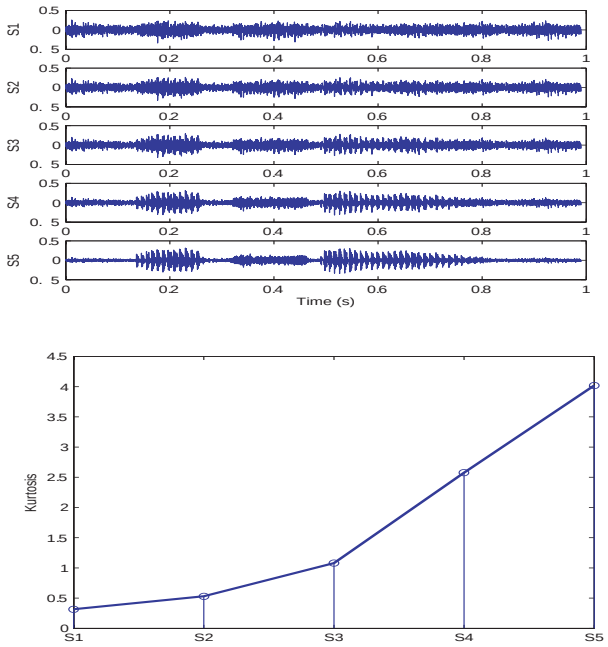


Figure 4.1: *Example of signals with increasing kurtosis. The upper panel shows the examples of signals, the corresponding kurtosis is shown in the lower panel. These signals are generated by project pursuit algorithms.*

can then analyze the kurtosis of these output signals and organized them into different folders with different kurtosis.

4.2 Generation of speech and noise material

As shown in Fig. 4.2, speech tokens (VCV words) were mixed with babble noise using standard Head-Related-Transfer-Functions (Gardner & Martin, 1994). The noise was simulated to come from 0 degrees (straight ahead of the listener) and the speech was simulated to come from 90 degrees. A set of 13 consonants (b, d, f, g, j, l, m, n, p, s, t, v, z) in the vowel context /a/ were used [1]. The babble speech noise was obtained from the Signal Processing Information Base (SPIB, 2007, Signal Processing Information Base (SPIB) retrieved from

[1] These items are selected in order to produce a range of kurtosis, which are 0.38, 0.5, 1, 2.5, 4.

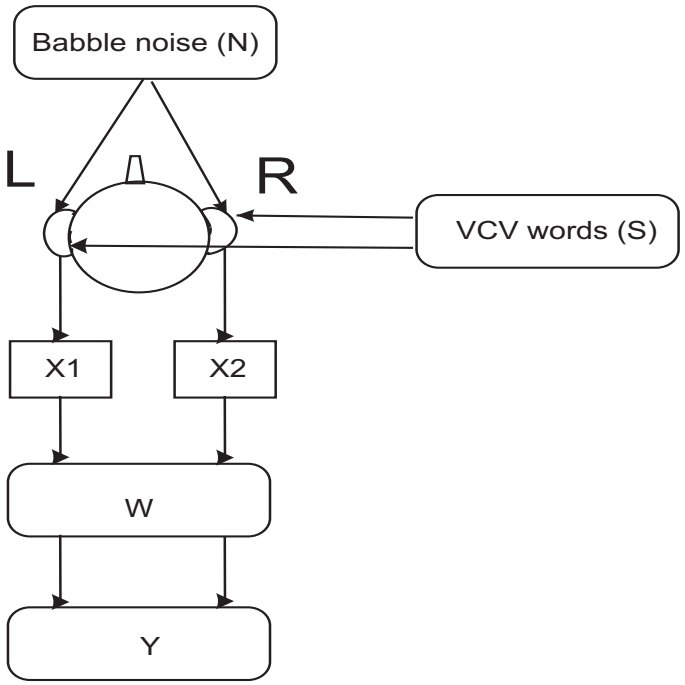


Figure 4.2: *Setup of the mixing process. In order to produce a series of signals (Y) with increasing kurtosis, a set of VCV tokens S was mixed with babble noise by Head Related Transfer Functions (HRTF). Two mixed signals X1, X2 were then fed into the Project Pursuit Algorithm and transformed by W. The output signal Y was then saved after each iteration. After analyzing its kurtosis, signals with similar kurtosis were then saved in the same folder, thus a group of signals with different kurtosis was generated.*

http://spib.rice.edu/spib/data/signals/noise) [1]. So in this experiment we use a speech shaped noise, which is different from the noise used in experiment of kurtosis and glimpsing, described in Chapter 2, in which babble modulated noise with different combinations of talkers are used. Noisy speech signals were grouped into 5 levels of sparseness (kurtosis) based on the project pursuit algorithm (Stone, 1993). The groups that were used in the experiment had mean kurtosis values of $k = 0.38, 0.5, 1, 2.5$ and 4, which were generated by the project pursuit algorithm.

[1] Voice babble acquired by recording samples from condensor microphone onto digital audio tape (DAT). The source of this babble is 100 people speaking in a canteen. The room radius is over two meters; therefore, individual voices are slightly audible.

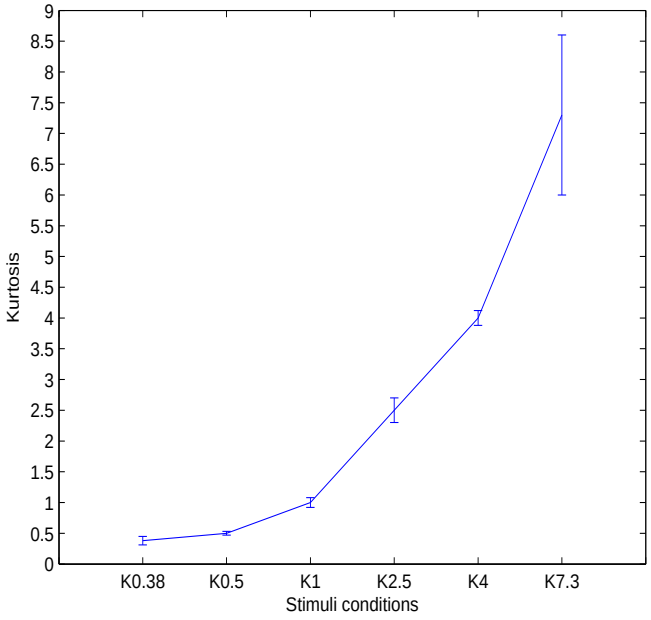


Figure 4.3: *Different level of kurtosis. Six stimulus conditions used for experiment 2. The error bars show the standard deviations for each condition.*

For clean signal, its kurtosis is about 7. Fig. 4.3 shows the standard deviation for each group. For the clean signal, the standard deviation is highest ($SD = 1.3$). Speech recognition performance was measured with normal hearing subjects in a sound proof booth to obtain speech recognition scores as a function of kurtosis.

4.3 Subjects and procedure

Seven normal hearing listeners (3 male, 4 female) participated in the experiment. The experiments got ethical approval from the ethic committee of Institute of Sound and Vibration, University of Southampton. Stimuli were presented monaurally through TDH-39 earphones. Each subject performed the test under all five kurtosis conditions. Each condition was repeated four times. The first time presentation of each condition was used for practice and was not scored. Order of conditions and tokens were both randomized.All listeners passed a pre-test on VCV word recognition in quiet by demonstrating a correct recognition rate of

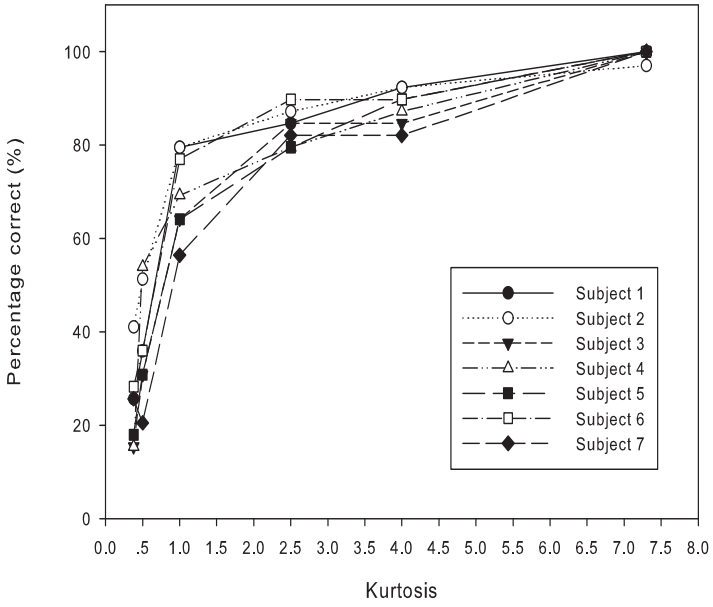


Figure 4.4: *Increased kurtosis predicts improved speech recognition score. Speech recognition score of seven normal hearing subjects increases with the increases of the kurtosis. The ceiling performance was reached when kurtosis is around* 2.5.

consonants at least 97%.

4.4 Results

Fig. 4.4 plots the speech recognition score as a function of kurtosis, including clean speech (kurtosis=7.3). It can be seen that speech recognition score increases with increase of kurtosis ($r = 0.8$, $p < 0.01$). It also shows that the recognition score increases very steeply with increasing kurtosis when the kurtosis value is low. The slope gets shallower when the kurtosis reaches 2.5, indicating a ceiling effect. The kurtosis of clean speech was on average 7.3. One-way analysis of variance (ANOVA) with post hoc comparisons (Bonferroni correction) of paired differences showed that all speech recognition scores (except for the pairs $K = 4, 2.5$ and $k = 1, 2.5$) grew significantly with increasing kurtosis.

In order to further investigate the relation between the speech recognition

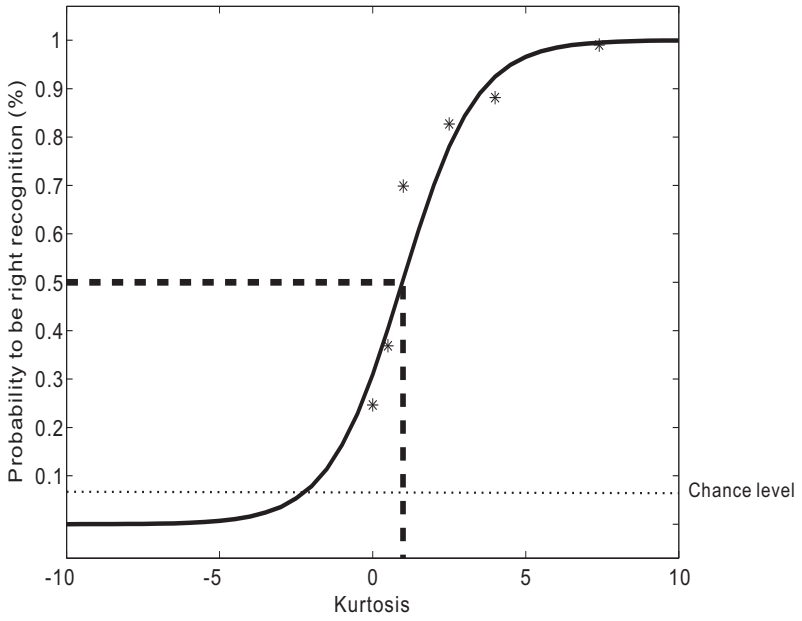


Figure 4.5: *Logistic regression to show the relationship between speech perception and kurtosis based on the result of 7 normal hearing subjects shown in Fig. 4.4. The observed mean recognition score from the second experiment is plotted as 'asterisks'. The point of* 50% *correct occurs when kurtosis equals* 1*. The chancel level is indicated by the dashed line on the bottom, chances=1/13*=0.08*.*

score and kurtosis a logistic regression curve was calculated, as shown in Fig. 4.5. The regression curve shows that the speech recognition can be effectively predicted by the value of kurtosis, higher kurtosis predicting higher score. To get a 50% correct score, kurtosis of such signal should be no less than approximately 1.0.

4.5 Discussion

4.5.1 Implications for auditory neuroscience

The notion of sparseness explored in the present study may be helpful to further understanding of auditory perception. The proposed approach differs from most

previous approaches by adopting a purely statistical rather than a deterministic view. The Projection Pursuit method used to manipulate sparseness can be considered as a process of pattern recognition, by searching for orthogonal dimensions in the incoming data stream (Stone, 1993). Note that this approach makes no a *priori* assumption about the dimensionality of the data; hence the approach should generalize to a wide variety of situations. Moreover, as the nature of the incoming data stream changes, the structure of extracted dimensions will change. This suggests that a system based on these principles will automatically demonstrate adaptation, a common feature of auditory perception that is not included in simple deterministic models. Sparseness is a key factor for neural representation of a natural environment (Bell & Sejnowski, 1996, 1997). A signal with sparse representation can be considered to be biologically efficient. However, it remains to be seen whether the principles of sparse coding can be used as the basis of a general-purpose model of auditory perception including auditory scene analysis.

Neurons in the auditory pathway adapt to stimuli that differ only in higher order statistics including kurtosis (Kvale & Schreiner, 2004). Although it is unclear at the present state of knowledge how this responsiveness arises from neuronal computation, it suggests that neurons could indeed adapt to higher order signal statistics. From here it is only a small step to suggest that this information is used for sharpening sensory responses. This could for example happen by targeted inhibition of neurons that indicate a smaller kurtosis, leaving neurons with higher kurtosis to code a better signal-to-noise ratio. Such a targeted inhibition would decrease the number of active neurons and therefore increase the sparseness.

4.5.2 Further research

The present study is limited to VCV words and could be usefully extended to other speech materials, such as sentences. According to the central limit theorem, a mixture of signals is usually more gaussian than each individual signal. Accordingly, the kurtosis of the mixture is usually smaller than that of individual sources. A sparser representation of the mixture, that is one with higher kurtosis, would be more similar to the representation of an individual signal and hence

it should be easier to recognize. This should also be true for whole sentences, although the added complication of redundancy due to syntax and meaning will make it a more complex problem. We therefore aim to perform further psychoacoustic experiments to examine sentence recognition in a similar fashion to the present study. We expect that kurtosis will also be a good predictor in more complex speech situations.

4.5.3 Implications for hearing aids and cochlear implants

We have shown that speech recognition in noise in normal listeners may be described as a statistical optimization process. By contrast, subjects with impaired hearing get less benefit from the glimpsing areas in modulated noise (Bronkhorst & Plomp, 1992). Impaired temporal resolution has been considered to be one of main reasons for this finding. This implies that the ability to explore high order statistics is worse in the impaired auditory system. Improving the speech recognition and quality is the holy grail of hearing aid research. Many digital noise suppression algorithms have been suggested, and some are in use in hearing aids. However, attempts to reduce environmental noise specifically are only just starting to emerge (Bentler & Chiou, 2006) and current technologies do indeed improve listening comfort, but not speech recognition (Dahlquist *et al.*, 2005; Ricketts & Hornsby, 2005). Summarising, there is a big demand for digital enhancement of speech quality in noise. We suggest here that kurtosis maximization could form the basis of algorithms for enhancement of speech in noise, such as used in modern digital signal-processing hearing aids or cochlear implants. Repeated calculation of kurtosis would be computationally more efficient than computationally more expensive methods such as glimpsing. As an additional benefit, the results of the present study imply that the proposed algorithm not only reduces noise but also improves the sparseness of the speech signal itself, leading to potentially even better speech recognition.

Specially, cochlear implant users might benefit from the kurtosis based speech processing algorithm considering the electrical-neuron interface in cochlear implants: Based on sparse coding theory, neurons fire sparsely under stimuli (Lewicki, 2002; Olshausen & Field, 2004). Hartmann *et al.* (1984) and Kiang & Moxon

(1972) found that the neurons firing patterns are highly synchronized with electrical stimuli, with almost similar phase as electrical stimuli. In order to get sparse neuron firing patterns for CI users, the electrical stimuli with high kurtosis should be helpful and they are supposed to be more biological efficient to code external stimuli (Olshausen & Field, 2004).

Chapter 5

Speech perception in different spaces

5.1 Introduction

Perception is a process that transforms the physical external world to an internal representation. This internal representation provides more clear and structured information to the cognitive level for further analysis. In order to understand how we perceive the external environment, an appropriate analysis of this internal representation is the key to understanding many perception problems. However, the internal representation is always hard to analyze, even harder to localize because we can hardly describe this internal process. For example, we can easily tell people whether we hear a sound or not, but we cannot describe how we hear. That knowledge always comes from the external world, an audiologist or a hearing scientist, who may explain the anatomy and physiology of the ear. But this is indeed not a surprise, it is a very old philosophical question of human being: we barely know ourselves.

This inability to describe the internal representation is a key barrier for many perceptual problems. One way to tackle this problem is to find underlying principles of perception, which could be used for our perceptual system. These principles guide us towards how this internal process works.

Macroscopic approaches can be used to investigate the perceptual system. The speech perception process can be thought as a transformation in linear equa-

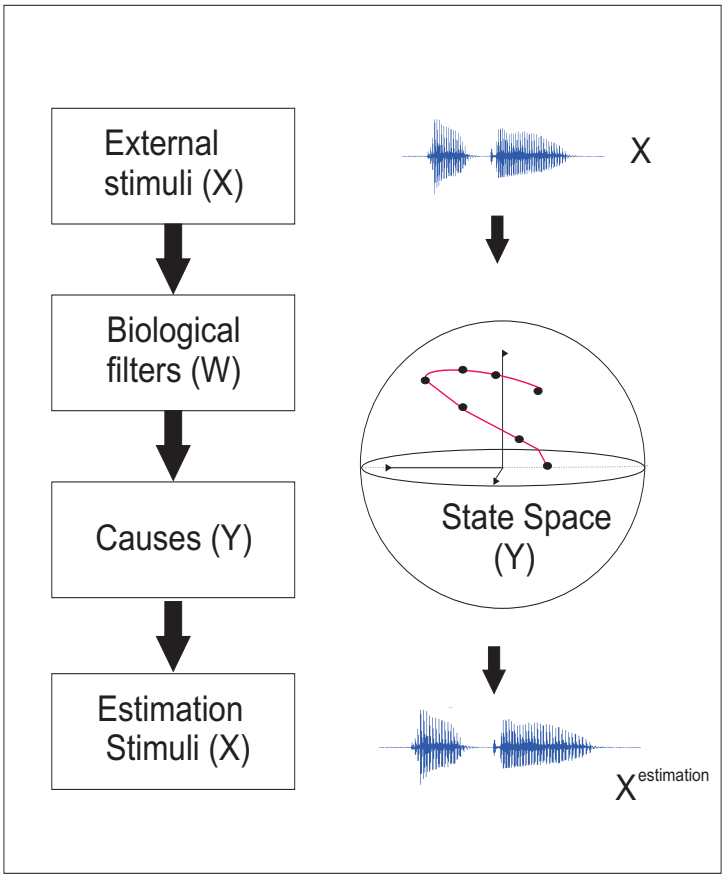


Figure 5.1: *Perception in the state space. Perception process can be described based on macroscopic approach as a linear filtering process. The external stimuli are transformed by biological filters and projected into a state space, where the causes of the data structure is more explicit. The right panel shows the imaginary state space, where the black dots are the corresponding vectors of the external stimuli in the state space. And the smooth line connecting these dots is the representation of the whole stimulus in the state space.*

tions (see Eq. 5.1), which makes the external signal explicitly coded as an internal representation in the state space Y. Such an internal space is the abstract representation of the information in external signals, which can approximate the true state of the environment by an inverse or a pseduo-inverse operation.

$$Causes(Y) = WX \tag{5.1}$$

Where W is biological filters [1] and X are external stimuli.

Fig. 5.1 shows the transformation and the imaginary state space. The structures of external stimuli are much more easily observed in the transformed state space. In the transformed state space, only few vectors are active and the representation is much more smoothed than the representation in the higher dimension, where each element of a stimulus can be thought of as one dimension (e.g. one pixel of an image) (Field, 1994; Olshausen & Field, 2004). In the state space, a stimulus can be transformed to few active vectors. The smooth line can be thought as high order representation of the stimuli, which connects the active vectors as a smooth line. The causes of the structure of the stimuli are thus clearly identified in the state space. Any other stimuli, whose state space transformation falls in the same line, would be thought by the perceptual system as similar external stimuli. The perceptual system recognizes and achieves perception by estimating the stimuli based on the components in this state space.

It is assumed that our biological filters have evolved efficiently to reveal the causes of the external stimuli. There are several computational methods modelling the biological filters. One is Fourier transformation, W in the peripheral auditory system, normal known as filter banks or Short Time Fourier Transforms (STFT). Y is the amplitude of different frequency components of stimulus X. For PCA analysis, the filters would be transformations which make causes Y uncorrelated. For ICA analysis, the filters would be transformations which make causes Y as independent as possible.

The columns of W^{-1} also called basis functions, which can show how cause explains individual external stimuli X. These basis function can be seen as basic

[1] Here W is a simplified as a linear filter.

calculation elements of biological systems. Once these basis functions, or filters, are known, we are able to better understand the processing of perception. There have been many modelling processes on what these biological filters should be (Field, 1994; Lewicki, 2002; Olshausen & Field, 1996). The detailed modelling process is certainly not trivial as the detailed processing principles of the perceptual process is not completely clear. However, These macroscopic approaches, based on some unified principle, can provide new insights on how these biological filters work (Olshausen & O'Connor, 2002).

The following sections will give a brief review of the three different basis functions and their relation to speech perception: Fourier transform, PCA and ICA. The investigation of these different options of the transformation will certainly help us understand how closely the computational methods help us understand the perception process and it may even provide new ideas for the design of cochlear implant speech processing. We can select or combine some of these transformations which are most close to the principle of how the human ears work. Once these possible transformations can be identified, we can then apply these transformations into enhanced speech processing algorithms. At the same time, these enhanced speech processing algorithms could shed new light on the study of speech perception.

For simplicity, we use signal Vowel-Consonant-Vowel word /aga/ as an example of external stimuli and it can be transformed either by Short time Fourier transform (STFT), PCA or ICA. The stimulus X can be processed in buffer and processed as a matrix:

$$\mathbf{S} = \begin{pmatrix} s_1 & s_{N-M+1} & \dots \\ s_2 & \dots & \dots \\ \dots & \ddots & \dots \\ s_N & \dots & \dots \end{pmatrix}$$

where N is the window length, and M is the length of overlap. The corresponding transformation functions will also be in a form of matrix (Smaragdis, 2001).

Just like in the short time Fourier transformation, the signal can be segmented by a time window of $N = 128$ (128 samples at 16000 Hz sampling rates) and

distances of overlap of $M = 32$, (32 samples at 16000 Hz sampling rate). In the following sections, the basis functions of three different transformations and their relation to speech perception will be discussed.

5.2 Time frequency domain

The auditory system has been traditionally seen as a time-frequency analyzer (Ohm 1843; Helmholtz 1863). Although with a limited precision, the spectrum provides a faithful representation of the spectro-temporal properties of the acoustic waveform. Using this representation, all speech sounds can be described in terms of an energy distribution across frequency and time (Greenberg *et al.*, 2004). Each spoken word can be decomposed into constituent sounds, known as phones, each with its own distinctive spectral signature. So the auditory system need only encode the spectrum, time frame by time frame. In this way, it can provide a complete representation of the speech signals for higher level processing. The spectrogram is thought to be a good display for speech, since it captures the dynamics of the sounds as well as the structure.

The basis functions of FFT for signal /aga/ are plotted in Fig. 5.2. Each subfigure is the real part of the column of the inverse FFT transformation matrix. The transformation matrix is a set of sines and cosines.

The fundamental link between spectrogram and speech perception was built by Liberman (1996) and his colleague Frank Cooper. They invented pattern playback (See Fig. 5.3) to investigate what are the most important cues for speech recognition. Pattern playback can reproduce sound based on the hand drawn spectrogram. Their research enhanced the belief that the spectrogram is a good tool for speech perception research. The visual approach of studying hearing through the spectrogram did make great contributions to speech perception research. For example, the role of formant transitions and its effect on speech perception is now more clearly understood based on their study of the spectrogram. For example, formants, transitions, common onset and modulation all can be traced in the spectrogram of sounds. These cues could be picked up by the auditory system and facilitate auditory processing.

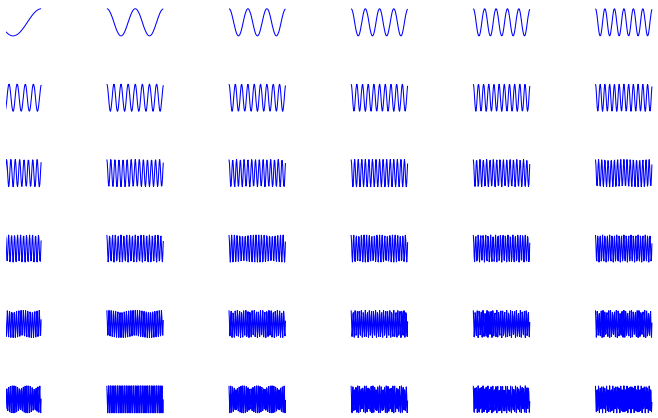


Figure 5.2: *Basis function of FFT. The FFT transformation is based on a set of fixed sine and cosine functions.*

Spectrum analysis or filter design inevitably involves trade-off between the precision of frequency tuning and temporal tuning. A low frequency tone includes cyclical fluctuation of air pressure and it needs longer time window to observe these cycles. But a longer time window means a decrease in the temporal accuracy. The discrimination of real-world sounds often requires accuracy both in time and frequency.

It is a challenge for us to understand how our ears's spectrum analysis function, which can be accurate both in time and frequency. But it does not necessarily mean that humans perceive sounds using the spectrogram only. The unified sine and cosine functions alone may not be best representations to describe the dynamic speech information, considering the trade off between time and frequency resolution of Fourier transformation. In fact we can certainly say speech perception must be more than the spectrogram analysis, considering the complex task which ears have to face every day. For example, in noisy environments a truly faithful representation of the spectrum could actually hinder the ability to understand due to competing speech or background noise (Greenberg *et al.*, 2004). It is very likely that the auditory system uses other strategies besides spectrum analysis. Otherwise the hearing system would be hardly able to

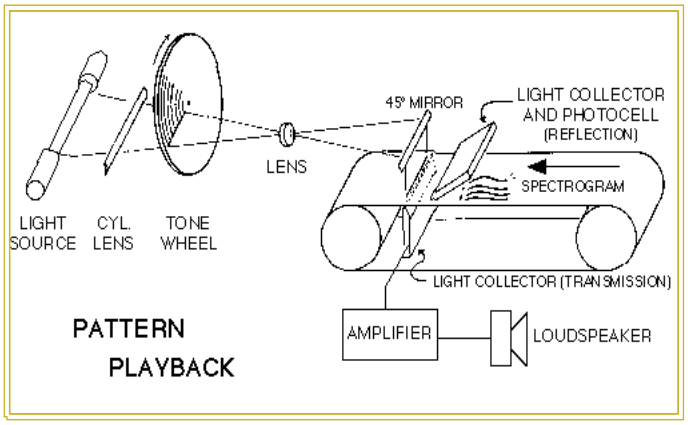


Figure 5.3: *Patten playback of the spectrum. It scans a spectrogram, using a line of light modulated by a tone wheel at some fifty harmonically-related frequencies which match approximately the frequency scale of the spectrogram. The selected light by the spectrogram then can be changed to sounds through a phototube. The intelligibility of the playback speech has been found to be 95 percent (Liberman, 1996).*

deal with normal daily life, where the spectrum of sounds is always a mixture of different sounds.

5.3 PCA and ICA transformation

Principal component analysis has been used for a long time in data analysis. The main advantage of PCA is that it can explore the data in an orthogonal space. It reduces the dimensions of the data easily by throwing away parts of the signal contributing less to the main signal. Reducing dimensions is especially important when the communication bandwidth becomes narrower or the channel capacity reduces. The auditory system of hearing impaired listeners can be thought of as a model of reduced channel capacity. The redundancy reduction is especially important to enhance speech in such conditions. The PCA analysis of noisy speech signals may help hearing impaired listeners to overcome some of the effects of reduced capacity in their impaired auditory system.

The auditory system could use a PCA-like strategy to discard certain gaussian

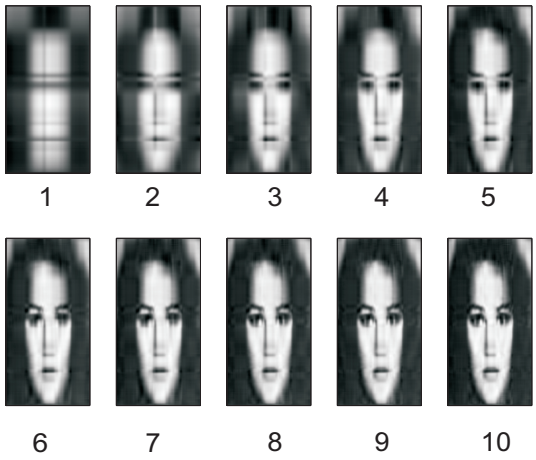


Figure 5.4: Redundancy reduction of image by PCA. The picture shown are reconstructed by PCA with different principal components. The number of principal components used is shown under each picture. With 4 or 5 principal components, the face can be easily recognized

noise, whose variance is smaller than the signal. Also it reduces the complexity of further processing. The complexity of the ear was less known until the research question of the cocktail party effect was proposed, where different sounds are presented simultaneously. The hearing system probably needs a faithful representation in the first stages (receiving most of the information), but further processing by exploring the redundancy or structure of the information could be also important for speech perception.

An example of a clean face image processed by PCA is shown in Fig. 5.4 in clean condition. To understand PCA from a filter point of view, PCA is doing a low pass filtering. The picture with only one principal components is blurred, which is low pass filtered. With the increase of principal components, the image is getting clearer and clearer. At the same time, it is evident that the image can be recognized clearly with only as few as 4 or 5 components. Using a similar approach, Fig. 5.5 shows the spectrogram of speech with and without PCA processing. With only 3 to 7 principal components, the spectrum of the processed signal is virtually the same as the original signal as shown in the bottom of the figure.

It is not clear though when this strategy would be used. But it seems that the ears need to explore the redundancy of information and only focus on the key parts of the signal. PCA can be used to transform the audio data to a more structured space.

Since the hearing system has the 'key' to the transfer function or so called basis of the transform, the signal can be easily transformed back to the original space after some processing such as de-noising or dimension reduction. The idea of such a strategy has been used often in engineering science. For example, the signal can be transferred to the frequency domain to reduce certain noise and then transformed back to the time domain. Similarly, perception could happen in this transformed space after certain processing.

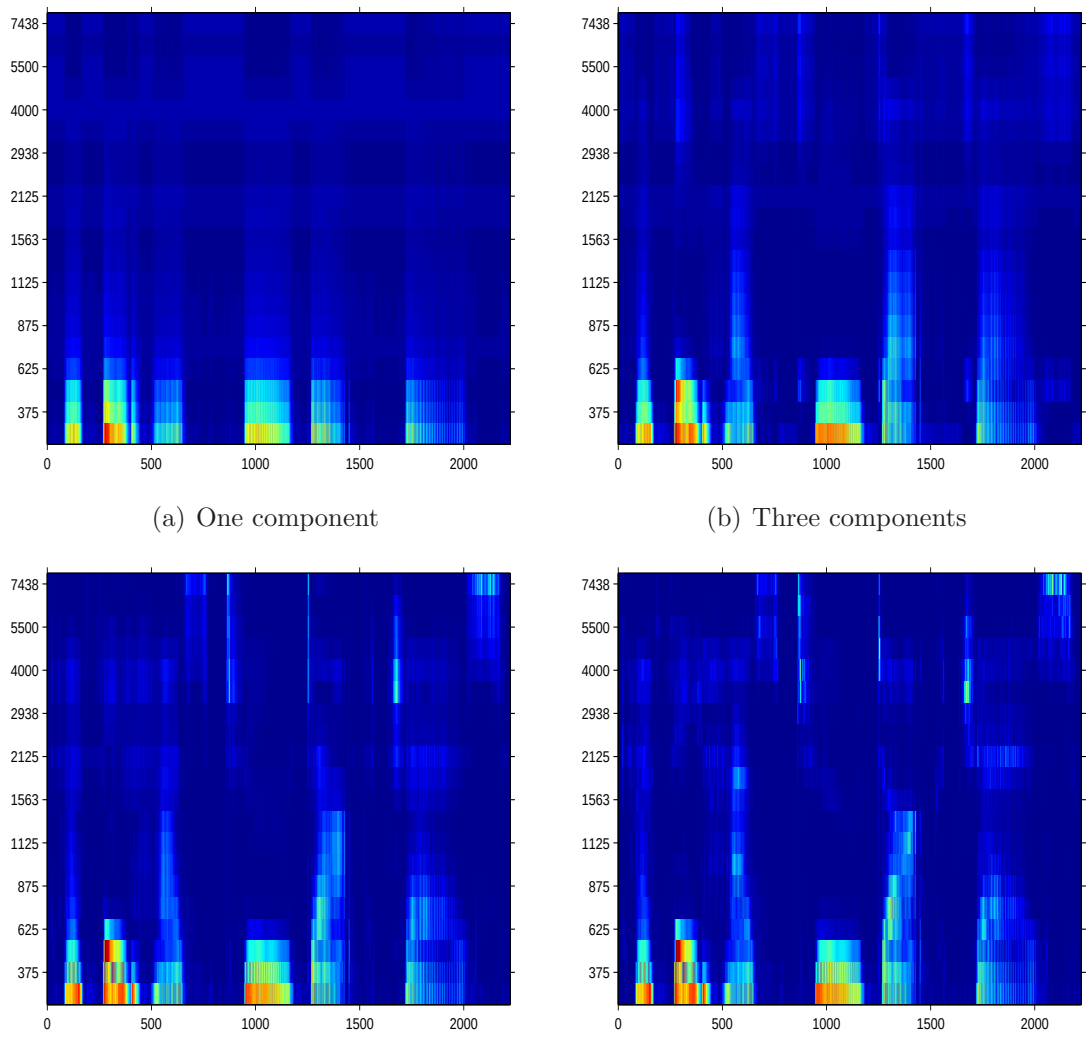
Both PCA [1]and Fourier transform can be thought as a transformation or filter, which changes the signal to another space where the structure of signal is more easily observed. The difference is the characteristics of the transform. As mentioned earlier, FFT uses fixed basis (sines and cosines, in Fig. 5.2), or so called data independent basis, while PCA uses data dependent basis, which are derived from the audio data itself.

Fig. 5.6 shows the basis functions of PCA. These basis functions are more centralized with a wide time spread. Speech data normally includes rapid changes both in time and frequency, such as transients; the optimal set of basis functions should have the property of time localized sinusoids which are able to capture rapid changes as in ICA basis in Fig. 5.7.

5.4 Sparse domain transform

Sparseness and redundancy are the key features of speech. The auditory system must use these characteristics to compensate for the distortion of communication channels and errors. Speech signals can be transformed by ICA to a more sparse space, where most of elements are zeroes. In the sparse space, only few non-zero points in this space are important to represent the whole signal.

[1]PCA can be thought as low pass filter



(a) One component

(b) Three components

(c) Seven components

(d) Original Spectrum

Figure 5.5: *This graph shows the usefulness of PCA for CI speech processing. d is the speech spectral envelope used for CI stimuli. The X-axis is the time in 'ms' and Y-axis is the character frequency of each electrode in 'Hz'. a, b, c are the results of PCA processing based on the original spectral of d: a is reconstructed by 1, b by 3 and c by 7 principal components. The reconstructed spectral envelope c is almost the same as that of d, the original spectrogram without PCA processing, suggesting PCA can be helpful to reduce redundant information for CI stimuli.*

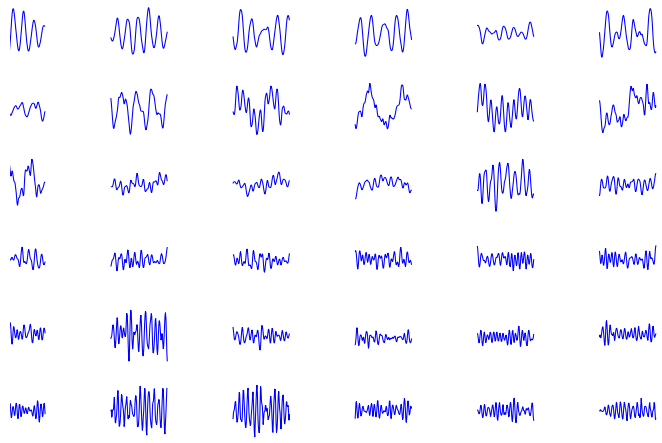


Figure 5.6: *Basis functions of PCA. The basis function of PCA are centralized with a wide time spread. PCA basis also include some sine or cosine waves.*

Correct recognition of these non-zero points might be critical to identify the whole signal correctly. The key characteristics of the speech can be preserved by these points in the sparse space with higher amplitude. As long as these key points exists, the identity of the signal will not change. This also could be the reason why we can recognize the same words in different environments, even from different people.

The neurons should fire sparsely, and only respond to specific features patterns based on sparse coding theory. ICA can be used to transform the signal to a sparse space, where only few components are important at the same time. In that case, the output should be as independent as possible.

The basis function of ICA for one word is derived from the ICA transformation and is shown in Fig. 5.7. The basis functions of ICA are much more localized in time than those of PCA. Such time localized filters should be much more efficient in representing speech signals.

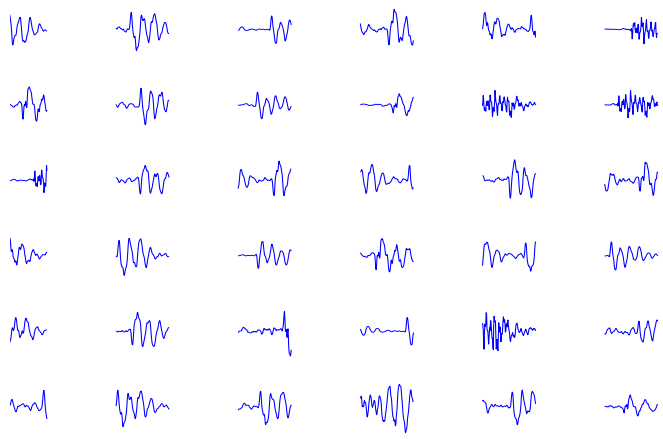


Figure 5.7: *Basis functions of ICA. Basis function derived from ICA are localized bursts through time.*

5.5 Discussion

To view our auditory perceptual system as an efficient transform is useful in understanding the behaviour of the system as a whole. The auditory peripheral system is an efficient information processing system. The modelling process can be viewed as a processing of looking for optimized filters. And the filter has to be biologically motivated. The characteristics of ICA based transformation have many advantages over Fourier transform and PCA, with a much more localized features in the basis functions for example. Both PCA and ICA basis functions are learned from the environmental data itself. And the statistics of input data are important for deriving such transformations.

Our hypothesis is that the PCA and ICA based transformation will be helpful for hearing impaired listeners, as speech communication can be thought as an information communication system. Hearing impaired listeners have a narrow dynamic range of communication channels. The consideration of reducing redundancy and making the stimuli more sparse should be helpful for efficient electrical stimulation of auditory neurons. Also the transformation provides alternative methods to enhance speech signals in noisy conditions.

In the following, a novel speech processing algorithm, SPARSE, based on PCA and ICA will be introduced in Chapter 6 and results of subjective experiments will be introduced in Chapter 7.

Chapter 6

Sparse stimuli for cochlear implants

6.1 Introduction

A cochlear implant is an electrical device that helps to restore partial hearing to the profoundly deaf. The main principle of cochlear implants is to use electrodes, inserted in the inner ear, stimulating the auditory nerves. Electrodes at different places correspond to different frequencies. Cochlear implants transfer acoustical information to the auditory perceptual system via electrical pulses representing modulation of the speech spectrum. Although the speech information sent through cochlear implants is quite crude, the performance of CI users has seen a big increase with new speech processors and algorithms. The majority of implant users have benefited from this device. Many of them can talk through the telephone without difficulty. Some top CI users even get similar performance in quiet as normal hearing subjects using clinical speech recognition test sentences (Wilson & Dorman, 2007).

However, the average performance of most cochlear implant users still falls below normal hearing, especially in a noisy environment. Normal hearing people understand speech well in a moderately noisy environment, but it is a very challenging situation for cochlear implant users to cope with. Normal hearing subjects are able to get masking release by exploring the characteristics of noise, for example, it is much easier to recognize speech when a different talker's voice

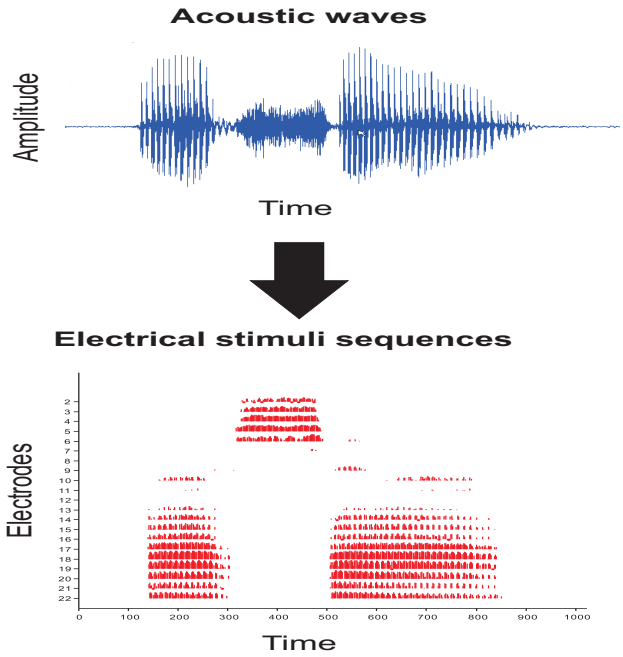


Figure 6.1: *Illustration of the bottleneck problem between acoustical information and electrical stimulation. The acoustical waves have to be transformed to a much more compact electrical stimuli space, in order to stimulate the auditory neurons with limited dynamic range. The upper panel shows an example of acoustic waveform and the lower panel shows the corresponding electrical stimulation. The x-axis is the time and y-axis is the electrodes. It is also called an electrodogram of cochlear implant stimuli. The dynamic range of the electrical stimuli is much smaller than the acoustic wave, compression is always needed to transform the stimuli from acoustic space to electrical stimuli.*

is used as masking noise . CI users are unable to take advantage of masking characteristic and no masking release was observed when a different masking voice were used (Stickney *et al.*, 2004). Friesen *et al.* (2001) found that the CI users on average were unable to further explore the benefit of more channels (up to 7), while normal hearing users were able to explore the advantages of more channels.

One of the important differences between normal hearing and cochlear implant user is the dynamic range they use to analyze sound. Normal hearing listeners are generally capable of detecting sounds as low as −10 dB SPL and as high as 110 dB without pain. Thus, the human ear is able to transduce about 120 dB

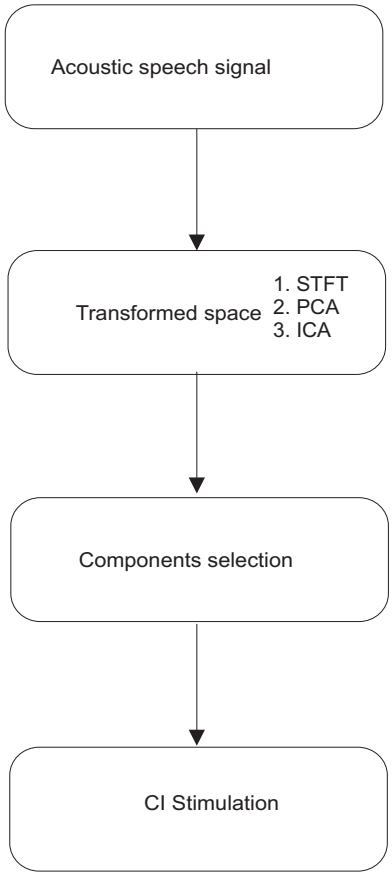


Figure 6.2: *Diagram of proposed processing scheme. The acoustical signal can be transformed to a space, where the causes of the data structure are more obvious. Examples of such transformations can be the Fourier transform, PCA or ICA transform, or a combination of these three transforms. The selection of the important information for cochlear implants can be processed in the transformed domain. Currently for many cochlear implants, this transformed domain is based on the Short Time Fourier Transform (STFT) only. Few channels are selected for stimulating the auditory neurons. Here we propose that PCA and ICA may help choose the right components of information for cochlear implant stimulation*

(1:1,000,000) dynamic range of sound pressure under normal hearing conditions. A normal hearing listener can discriminate up to 200 intensity differences within this 120 dB dynamic range in acoustic hearing (Bacon *et al.*, 2004).

For cochlear implant users, the dynamic range for electrical stimulation ranges from their threshold (T level) to their most comfortable level(zC level).

These electrical levels must be measured for every electrode channel. The dynamic range of the impaired ear is much smaller than the normal ear (Greenberg *et al.*, 2004). Thus the electrical stimulation has a severe bottleneck which only allows limited acoustic information to be transmitted to the auditory neurons.

In order to achieve higher speech recognition performance, a cochlear implant has to transfer the most essential information into the limited dynamic range of impaired auditory neurons. The limited dynamic range problem is essentially an information transmission problem. The CI processor has to find a way to optimally transfer most relevant speech information to CI users.

Fig. 6.1 illustrates that a cochlear implant processor needs to narrow down the acoustical space to a much narrower electrical stimuli space. Only limited acoustic information can be selected and used by cochlear implant users due to the limited dynamic range of cochlear implant users. The encoded information actually competes with redundant information for the limited encoding resources. In order to efficiently use the limited encoding resources, a redundancy reduction strategy can be used to select the most necessary information and discard the redundant or noisy speech. Searching for this necessary or important information for speech understanding has been the key question since the beginning of cochlear implant and speech perception research. Many features of the speech signal, according to psychoacoustic experiments, are important for speech perception, such as formants[1], envelopes, fine structure and frequency modulation (Diehl *et al.*, 2003). However, this selection was based on a microscopic approach, not from an information processing point of view or a macroscopic approach by viewing the auditory system as an information transformation encoding process (Plomp, 2002).

It is possible that we recognize what is the important information for speech recognition by systematic information analysis, without pre-assuming what are

[1]Peaks in the spectrum of speech.

or are not important cues related to speech recognition. Also if appropriate transformation is applied on the acoustical signal, the underlying causes of the data structure can be more explicit. Once we are able to reveal the causes of the data structure we then should be able to perceive or understand the signal properly. And this transformation has long been pursued based on information theory.

Here we propose that PCA and ICA can work as an information analysis tool, exploring the data structure of speech signals, choosing the important information for CI users. The principle of the present work is to use such techniques to overcome the bottleneck problem, with a view to optimize information transmission through the cochlear implant. We can only transfer limited information from the acoustic domain to electrical pulses due to the information bottleneck. However, the understanding of speech, after all does not require every piece of the acoustical information.

The information selection may actually also provide a good opportunity to enhance the signal as the selection process may well reduce the redundant parts or even noisy parts of speech signals. Thus it can help to improve speech-to-noise ratio. The advantage of PCA/ICA is that it is purely data driven. The calculation is purely based on the signal itself, rather than based on any prior theories or models of speech features. PCA can be used to reduce the dimensions of the data by identifying only the components with larger variance. ICA can recover the causes of the data structure by extracting factors, which are independent of each other and representing the most abstract information of speech signal.

PCA or ICA transforms the signal into a space, in which the signal structure is much more explicit as the causes for independent factors of the data can be extracted. Signals can be transformed from the original acoustic space to a more compact (PCA) or sparse space (by ICA). After reducing non-important components in the transformed space, the signal can then be transformed back to the original time-frequency space for auditory stimulation. Processing such as reducing redundant components or de-noising, can be done in the transformed space, where the distinct features of signals are more obvious. This can help to filter out the key information, which can represent the key structure of the data.

Fig. 6.2 illustrates that an acoustic signal can be transformed into a different space, and in this space we can select and enhance the important information or suppress the non-important information. In Fourier transformation, we can reduce noise or enhance speech if we know the specific frequency characteristic of noise or speech. In PCA/ICA transformation, we can do similar filtering by looking into the basis of the transform or the amplitude of the independent factors.

Also the sparse coding theory has explicitly suggested that only few sensory neurons fire at the same time Olshausen & Field (2004) . Here we assume that the stimulus patterns should be sparse in order to get a sparse firing pattern for the cochlear implant as the firing of neurons is highly synchronized under electrical stimulation. Once the sparse coding principle can be implemented in the cochlear implant, the auditory neurons should be firing sparsely. And the performance of CI users should also improve.

Seeing that sparseness optimization of a signal can be an efficient strategy for speech perception, the application of ICA or PCA will be helpful for cochlear implant stimulation, where information transmission is restricted (see Fig. 6.1).

Preliminary investigation with subjective experiments has shown that sparse coding principles are a promising approach to improve the presentation of speech in noise via cochlear implants.

6.2 Introduction to cochlear implants

6.2.1 The principle of cochlear implants

It was believed that the original idea of using electrical stimuli to get a sense of hearing was from Italian scientist Alessandro Volta (Volta, 1800). While he was studying the effect of electrical stimulation from a 50-V Battery, he found that the electrical stimuli to his ears caused a sense of hearing (Bacon *et al.*, 2004).

> *"at the moment when the circuit was completed, I received a shock in the head, some moments after I began to hear a sound, or rather noise in the ears, which I cannot well define: It was kind of crackling with*

> *shocks, as if some paste or tenacious matter had been boiling...The disagreeable sensation, which I believed might be dangerous because of the shock in the brain, prevent me from repeating this experiment...”*

Based on the idea that senses can be triggered by electrically stimulating the sensory neurons, the modern cochlear implant transfers the acoustical speech signal to electrical pulse stimuli for auditory neurons. A schematic overview of a cochlear implant is shown in Fig. 6.3. Although commercially there are a number of different cochlear implant systems available, all of them consist of a microphone, a signal processor, a signal coupler (transmitter and receiver) or a plug, an internal decoder chip and an electrode array, which is inserted in the cochlea. The microphone and signal processor are worn outside the body (external part). The microphone senses sound pressure variations of acoustical signals and converts them into electrical signals. Based on the algorithm in the processor, the electrical signals are processed to produce electrical stimuli for the electrodes that are implanted surgically inside the cochlea of the deaf ear (internal part). The external and internal parts are connected by a transmitter or a plug. Nowadays, all clinical devices use a transmitter and receiver as a link between processor and internal decoder chip, which decodes and extracts the current stimulus amplitudes. These decoded current signals are sent to the electrode array and stimulate the auditory nerve to elicit action potentials, and this neural activity is sent to the brain. Bidirectional transmission is also available now. It allows transmission from the internal to the external part (reverse telemetry), which can supply very useful information on assessing the status of auditory nerves and the fitting for the speech processor.

6.2.2 The task of speech processor

The task of the speech processor of CI is to transfer useful information of speech or environmental sounds, which the auditory nerves can interpret and send it to the brain for further analysis. The development of the speech processor can be seen as a process looking for such critical information which can be preserved and transferred to auditory nerves of deaf people. The core part of speech processor

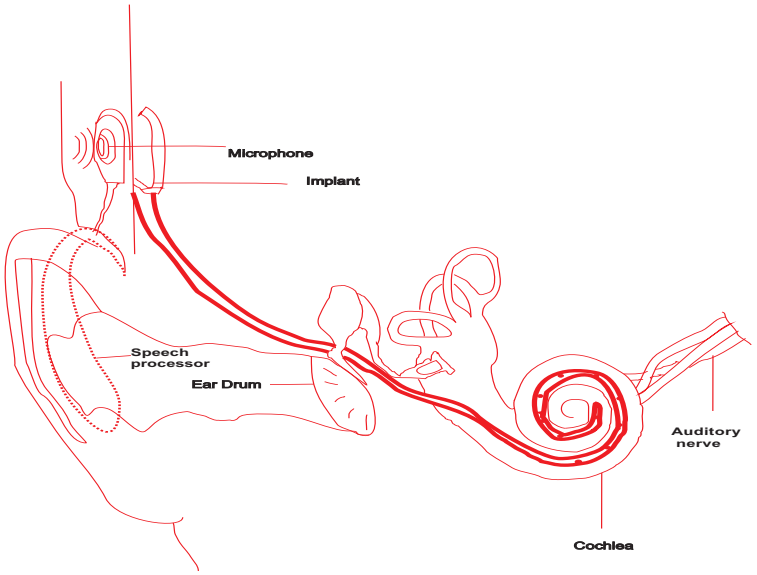


Figure 6.3: *A schematic overview of cochlear implant working principle. The cochlear implant bypasses the outer ear, middle ear and inner ear. The microphone first picks us the acoustical signal and then the speech processor processes the acoustic waves. Electrical stimulus sequences are then generated and sent to auditory neurons through electrodes of CI, which are implanted in the inner ear.*

is the speech processing algorithm, which decides what kind of acoustical speech information to use and how to deliver it to the impaired auditory system.

An investigation of such a process needs to look into how our ears analyze the acoustical signals. Most of the modern speech processing algorithms of CI today try to mimic to some extent the signal processing in the cochlea of the normal hearing person.

Fig. 6.4 shows some common principles between normal hearing and the signal processing in the speech processor of a cochlear implant. For normal hearing, information of acoustical signals is sent to the cochlea of the inner ear, after passing through the outer ear and middle ear. The basilar membrane in the inner ear plays a role of frequency analysis for the acoustical signals: different frequency components cause maximum vibration amplitude at different points along the basilar membrane (Helmholtz, 1925; Moore, 2003a). Different places on the basilar membrane correspond to different frequencies, which is referred as

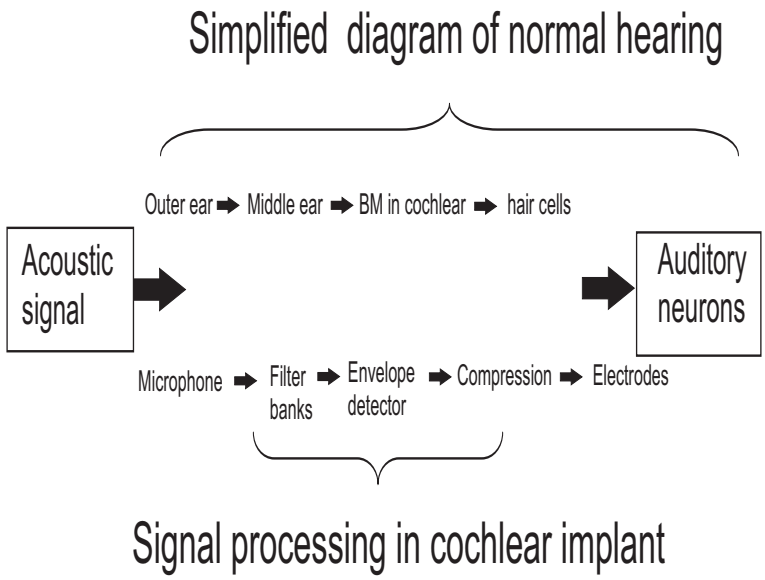


Figure 6.4: *CI speech processing mimic the function of normal hearing. Acoustical signal is firstly picked up by microphone and then envelopes are extracted across different frequency channels. Compression is done to fit the narrow dynamic range of auditory neurons.*

the tonotopic organization of the cochlea. Complex sounds are decomposed into different frequency components. The relation between frequency and position has been modelled by Greenwood (1990). According to his formula, those frequencies (the characteristic frequencies) are approximately linearly distributed along the length of the basilar membrane for frequencies below 1 kHz, and logarithmically for higher frequencies. Movements of the basilar membrane are sensed by the inner hair cells (IHCs),

and then the chemical transmitter substance is released. Thus the information of acoustical signals can be coded and sent to the central nervous system later.

The cochlear implant bypasses the outer ear, middle ear and hair cells, sending certain acoustical information directly to the auditory nerve. In the speech-processing algorithm, filter banks are used to decompose a signal into several frequency components. After envelope detection (rectification and low pass filtering), compression and modulation of electrical pulse trains, the electrical stimuli are sent to electrodes, stimulating the nerves at different tonotopical places, where

only specific frequency components of the signal can get maximum response.

6.2.3 Review of speech processing algorithms

What speech information should be used and how to stimulate the neurons were the key questions from the very beginning cochlear implant development. Also only limited acoustical information can be presented to the nervous system via current cochlear implants. This is also called the bottleneck problem of cochlear implants (Clark, 2003).

Speech processing algorithms replicate part of the function of the auditory system. This normally includes spectral analysis and compression to suit the wider dynamic range of acoustical sound (Greenberg *et al.*, 2004). The main differences between these algorithms are the process after spectral analysis. There are two kinds of speech processing algorithms. One is focusing on the spectral cues and the other focus on the temporal cues.

Psycho-acoustical experiments show that one of the main cues of speech perception for the normal hearing is spectral information, such as formants (Assmann, 1995). They are the spectral peaks of the speech. These spectral speaks reflect how the speech is physically produced and resonates in the vocal tract during the speech production process. Formants were implemented in the earlier cochlear implant speech processing algorithms: F0/F2 and later F0/F1/F2 based strategy (Greenberg *et al.*, 2004).

Besides exploring the spectral cues, the other idea of speech processing strategies is to use the temporal cues of speech for cochlear implant stimulation. There are three different signal processing strategies using this concept. They are Compressed Analog or Simultaneous Analog Stimulation (CA/SAS) (Eddington & Dobelle, 1978), Continuous Interleaved Sampling (CIS) (Wilson *et al.*, 1991), Spectral Peak (SPEAK) and Advanced Combined Encoder (ACE).

The CA strategy is to deliver the narrow band analogue waveform to the appropriate electrodes and directly stimulate the auditory neurons. The interference between channels was thought to be problematic for the CA strategy. The advantages of multi-electrodes are shown by CA. This strategy stimulates the auditory neurons using analogue or continuous waveforms for stimuli, instead of biphasic

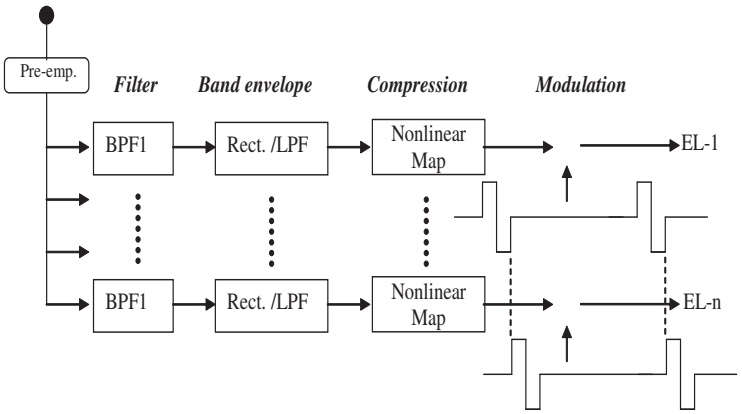


Figure 6.5: *Standard CIS processor. Abbreviations include 'Pre-emp' for 'Pre-emphasis'; 'BPF' for 'Band Pass Filter,' 'Rect.' for 'Rectifier,' 'LPF' for 'Low Pass Filter' and 'EL' for 'Electrode'. (Adapted from Wilson et al., 1991, NIH project N01-DC-8-2105, 2002.).*

pulses. The speech signal is first fed into an amplification gain control, and then separate into different channels, after some gain compensation, the stimuli are sent simultaneously to different electrodes.

Later the Continuous Interleaved Sampling (CIS) algorithm was developed (Wilson *et al.*, 1991). Channel stimulation rate was thought to be also important for speech perception based on electrical stimuli. The CIS strategy is designed to avoid interferences between channels. The speech envelopes are extracted and then modulated by biphasic pulses. These pulses are used for stimulation of auditory neurons.

Fig. 6.5 shows the signal-processing scheme of the standard CIS. The pre-emphasis filter is to compensate the -6 dB/octave natural slope starting from 500 Hz in the long-term speech spectrum. Spectrum analysis is based on filter banks. Rectification and low pass filters are used in order to obtain a measure for the speech envelope. The electrical pulse trains to the stimulation channels are modulated by these envelopes of signals in the corresponding band pass filters. In

addition, the pulse trains are separated in time and interleaved in order to avoid the interaction among the electrodes.

In order to increase the stimulation rates of cochlear implants, only a few channels are selected based on the spectrum amplitude to stimulate the auditory neurons. One strategy is called Spectral Peak or Spectral Maxima Sound Processor (SPEAK). The main feature of the SPEAK strategy is to select few channels to stimulate in order to achieve a higher stimulation rates, also partially it could potentially reduce noise by selecting the channels with higher amplitude, presumably containing speech.

Later the advantage of CIS and SPEAK were combined by ACE (Advanced combined Encoder). It can optimize the patients' response to rate and channel numbers. The ACE strategy now is regarded as a default strategy for cochlear implant subjects for the Cochlear CI24M device. For review on the speech processing algorithms in detail, please see (Bacon *et al.*, 2004; Greenberg *et al.*, 2004).

6.2.4 Sparse stimuli for cochlear implants: SPARSE

The history of speech processing strategy can be seen as moving in a 'sparse' direction, although it is true that CIS and SPEAK were not motivated by sparse coding principles. CIS stimulate the auditory nerves non-simultaneously in order to reduce the interaction between channels. From a sparse theory point of view, it stimulate the auditory neurons sparsely. At each moment, only a single group of neurons will be the main stimulation target. And thus only few neurons will fire at the same time, which is indeed the key concept of sparse theory of neurons coding. CIS stimulation is temporally sparse.

The SPEAK strategy was designed to select only a few channels to stimulate auditory neurons and thus can have a higher stimulation rates for each channel given a constant stimulation rate on average (channel stimulation rate = total stimulation rate/number of channels). From a sparse theory point of view, such a stimulation strategy will only stimulate a single group of neurons with similar characteristic frequency. This will give spectrally sparse stimulation. This sparse spectrum can be seen as a result of redundancy reduction on the spectrum of

speech. By reducing the dimensions of the spectrum, the information can be encoded more efficiently by the electrically stimulated neurons with narrow dynamic range.

The improvement of speech recognition performance achieved with the modern speech processing algorithms can be seen as a result of implementing the theory of sparse coding unconsciously. Here we propose a novel speech processing algorithm, SPARSE, based explicitly on sparse coding theory.

6.3 Sparse stimuli for cochlear implants

6.3.1 Combined compact coding (PCA)and sparse coding (ICA)

In order to deal with the limited dynamic range of CI users, the redundancy reduction techniques such as PCA can be used for CI speech processing. PCA can help to reduce redundancy of information based on second order statistics. But the output of PCA is too compact to be used by neuron systems (Barlow, 2001). PCA can make the data suit the limited dynamic range but cannot provide a sparse structure which neuronal systems might be able to explore. In order to implement sparse coding theory systematically, ICA can be used to transform the output of PCA into a sparse domain, in which speech enhancement processing can be done. So here we propose the idea of reducing the redundancy of the stimuli by PCA and then making the electrical stimuli sparse by ICA.

Both PCA and ICA analysis have different algorithms to be implemented. There are also online versions of PCA and ICA calculation (Hyvarinen & Oja, 2001), which make the real time implementation possible. This thesis will only focus on the off-line algorithm and investigates whether such combination will help to improve the speech recognition performance of cochlear implant users.

The application of PCA and ICA can be either in the time domain or in the spectral domain. Here we choose to apply PCA and ICA on the speech spectrum envelope. One reason to work on the spectrum envelope is to make the algorithm as an extension of the current algorithm and so that it can be implemented in the current commercial algorithms. The other reason is that spectral analysis is an

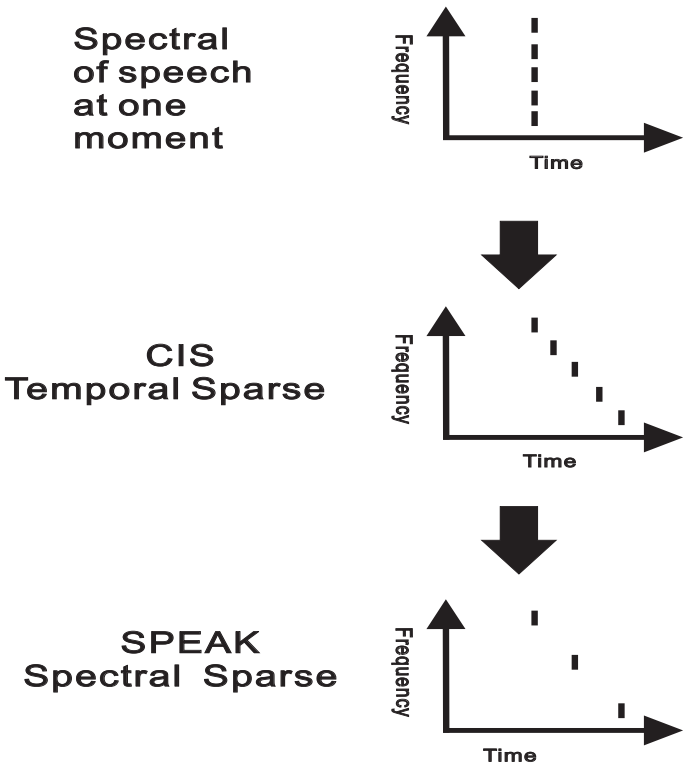


Figure 6.6: *The 'sparse' direction of cochlear implant developments. Given a speech spectrum at one moment, the CIS strategy stimulates auditory neurons non-simultaneously. This actually make the neurons fire not at the same time, which would reduce channel interaction and make the representation of the spectrum sparse along the time dimension. Further with SPEAK strategy, it only select a few channels out of 20 or 22 channels. This will introduce spectral sparseness. Accordingly, only a few group of auditory neurons will fire. Both ACE and SPEAK reflect the key concepts of sparse coding theory in that only few neurons fire at the same time.*

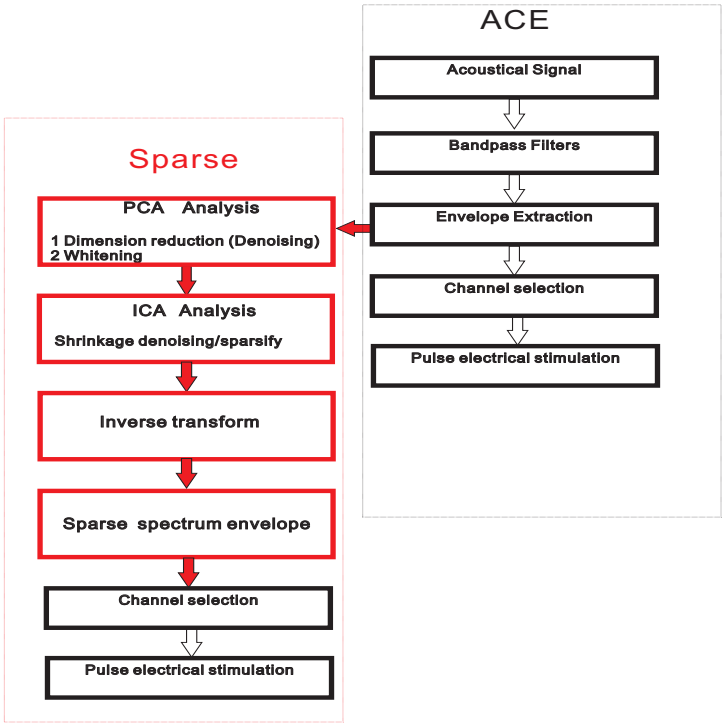


Figure 6.7: *Flowchart of SPARSE algorithm. The acoustical signal first is processed by filter banks and then envelopes are extracted. ACE algorithm will select a few channels out of 20 to 22 channels and stimulate the auditory neurons non-simultaneously. The SPARSE algorithm uses PCA and ICA analysis on the envelope signals, reducing redundancy and making the spectrum envelope representation sparse.*

essential function of the auditory system. The speech processing algorithm has to perform certain spectral analysis to achieve the place coding of frequency via the cochlear implant: different electrodes correspond to different frequencies.

Fig. 6.7 shows the idea of combination of PCA and ICA for CI stimulation. The spectral information of speech thus can be further analyzed by PCA and the redundancy of speech information can be reduced by reconstructing the speech signal using the larger eigenvectors from PCA. A whitening process is then followed, which can be done through eigenvectors and the diagonal matrix of the eigenvalues (Hyvarinen & Oja, 2001). After PCA redundancy reduction, ICA can be used and transform the whitened data into independent space. In the

independent space, the redundancy of the data can be further reduced by making independent vectors zero, if their values are smaller than a certain threshold, as the important causes of the speech signal are expected to be larger components in the independent space. This de-noising process in the independent space is also called sparse code shrinkage and the threshold is determined through maximum likelihood estimation (Hyvarinen, 1999).

In order to produce tonotopic stimuli for cochlear implants based on the spectral analysis, an inverse transform of ICA is then needed to transform the stimuli back to the spectro-temporal domain and appropriate electrodes can then be selected and stimulate the auditory neurons.

The statistical property of the input is important for perception, and these statistical methods (PCA and ICA) are used to explore the state space of the input stimuli and transfer them to a more sparse space. We can then do speech enhancement in this transformed space. Both PCA and ICA can work as preprocessing for CI speech processing. It can also be applied in channel selection (dimension reduction), de-noising and reducing the channel correlation.

6.3.1.1 PCA processing for speech envelopes

One of most popular speech processing strategies is ACE. It extracts the speech envelopes by filter banks and then selects a few channels with higher spectral energy. The selection is based on the magnitude of envelopes in each channel. Channel selection actually can be seen as a process of dimension reduction. For the ACE strategy, the default algorithm is to choose the most active 12 out of 22 channels at any time. A classic method for dimension reduction is PCA. After transforming the envelope matrix by PCA, only the channels corresponding to the major principal components are preserved, discarding the channels with smaller amplitude. And then the inverse PCA can transform the matrix back to the original time frequency domain.

Fig. 6.8 shows how PCA can be used to reduce the redundancy of the spectrum envelope. The spectrum envelope is first transformed by PCA and only few principal components are select to represent the original spectrum envelope. The main features of the spectrum can be preserved. As PCA is based on second order

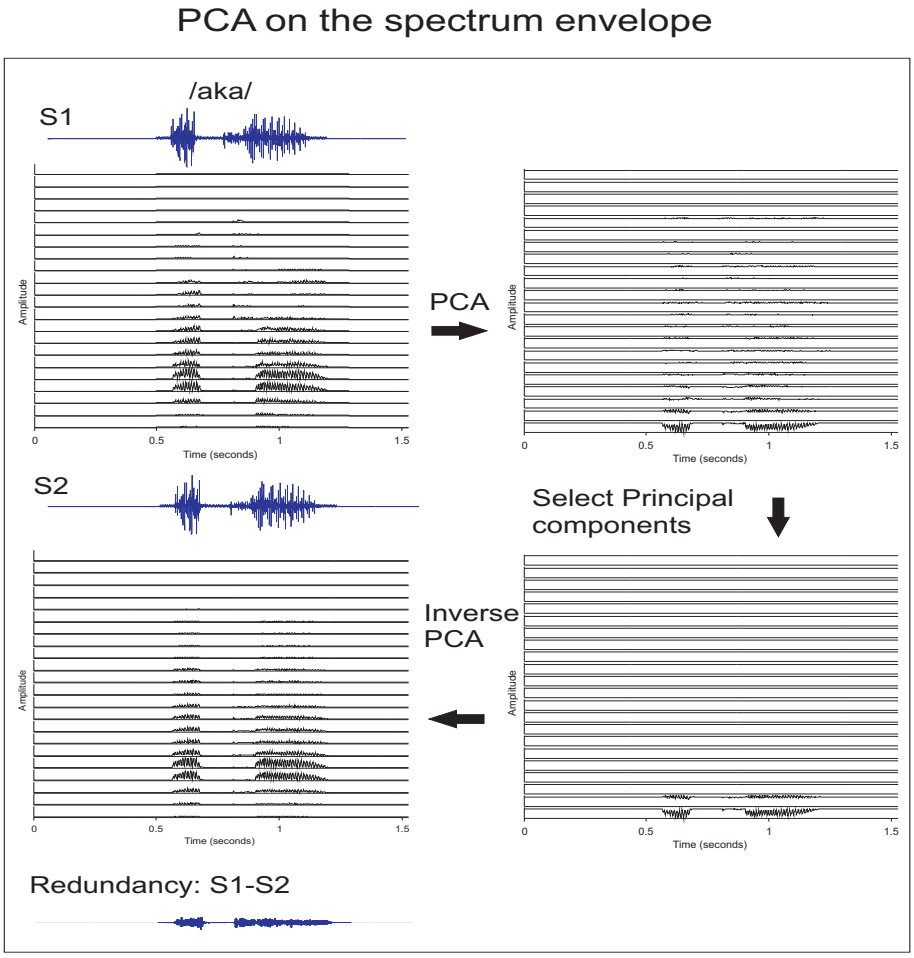


Figure 6.8: *Example of spectrum analysis by PCA. The spectrum envelopes can be further analyzed by PCA and redundancy can be reduced by using only few principal components to reconstruct the envelope matrix. Two principal components are used to reconstruct the speech signal /aka/. It can be seen from the reconstructed waveforms that the main envelope information can be observed in both the original signal S1 and signal through PCA, S2. The reduced redundancy is the difference between S1 and S2.*

statistics, it chooses important components based on the variances or energy. If too few principal components are selected, PCA redundancy reduction could reduce the component of speech which are small in amplitude but important for speech perception, such as some plosive consonants.

So there is a balance of how many principal components are needed to reconstruct the signal while at the same time keeping maximal the essential information of the original signal. If all the principal components are used, the signal will be the same as the original signal. If too few principal components are used, the signal will be reconstructed with loss of important information such as complete loss of weak consonants. The difference between ACE and the PCA approach, is that PCA works on orthogonalized components to select important information, which are uncorrelated, whereas ACE works on original channel components that are likely to be correlated with one another.

One possible principle to select the number of principal components is to detect the changes of eigenvalues of the covariance of the speech spectrum envelopes (Hyvarinen & Oja, 2001). When the change of eigenvalues become constant, this is taken as the threshold value. The values smaller than this eigenvalue can be seen as noise or very redundant parts. PCA based de-noising has been used in cochlear implant speech processing by reducing the noise in the subspace (Loizou *et al.*, 2005). For the signal /aka/, the eigenvalues becomes almost constant after 8. So the signal can be almost perfectly represented by using only 8 principal components as shown in Fig. 6.9.

Cochlear implant stimulation requires a compact representation for speech information given limited dynamic range. The representation needs to include the key information for speech recognition and at the same time, to reduce unnecessary data. PCA processing could enhance the formants, which are the areas with peak energy. which are important for vowel perception. As seen in Fig. 6.11 and Fig. 6.10, the PCA produced signal can enhance the formants. At the same time, it also produces sparse representations of acoustical signals.

This sparseness in representation of acoustical information has been applied in the stimulation of cochlear implant for a long time and achieved great success. But the idea of sparse representation was not properly quantified or formally

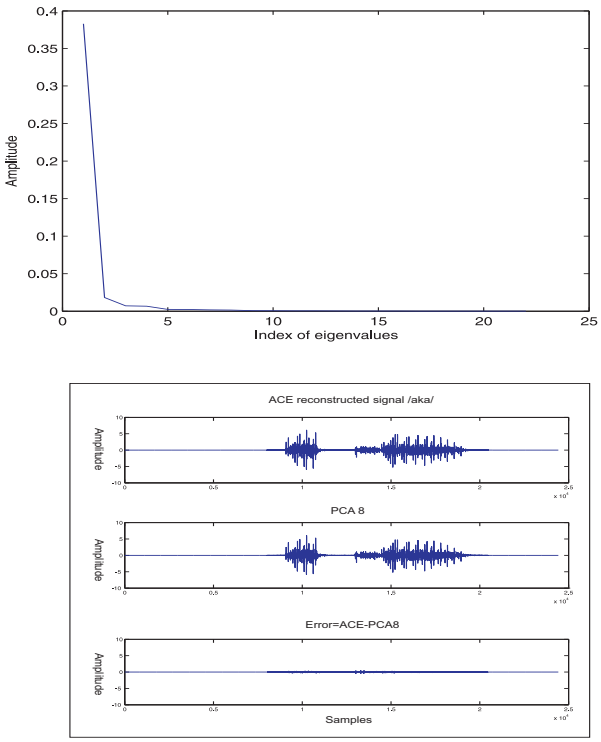


Figure 6.9: *Eigenvalues and reconstruction of a signal by PCA. The eigenvalues of the covariance matrix become constant after index 8 as shown in the upper panel. The dimensionality of 22 can then be reduced to 8. Thelower panel shows the reconstructed signal is quite similar as the original signal and the difference is quite small.*

proposed for cochlear implants. One of parameters to quantify sparseness is kurtosis.

Although simple PCA can make the speech sparse, it could lose the weak consonants if extremely small numbers of principal components are selected. ICA is needed to implement sparse coding as it decides the importance components based on higher order statistics.

6.3.2 ICA for cochlear implants: SPARSE algorithm

PCA only uses second order statistics, by means of the covariance matrix. Given a set of vectors, it can transform them to uncorrelated vectors by multiplying

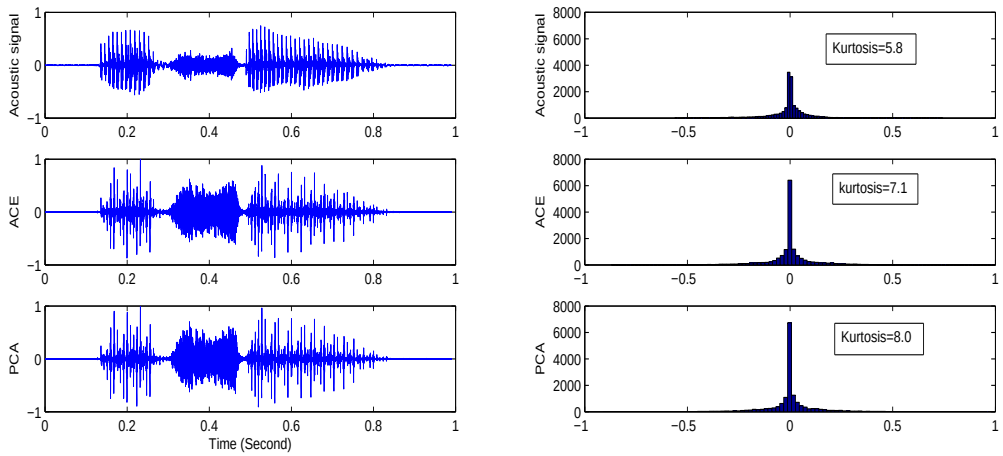


Figure 6.10: *Kurtosis and reconstruction of ACE output by PCA. The left panel shows the waveforms of acoustical signals before processing and reconstructed signals processed by ACE and PCA. The right panel shows the corresponding histograms. Kurtosis is indicated in the textbox, showing the sparseness of the PCA based signal is the highest.*

eigenvectors of the covariance matrix. ICA uses higher order statistics, making the vectors as independent (sparse) as possible. A sparse representation of the signal could be helpful for speech perception in noise. And it also can be useful for electrical stimulation for the auditory system, reducing channel interaction and making the auditory neurons fire sparsely. PCA normally work as pre-processing for ICA to reduce the data dimensionality. Then it is common to use ICA to do further processing to make the output of ICA as independent as possible. In the following section we assume that the signal has been pre-processed by PCA including dimension reduction and whitening.

ICA can transform signals into a space where mutual information between channels is minimum and each output signal in this space is independent. Distinct features can be observed in the this sparse space, as it can indicate the causes of the speech signal. Speech enhancement can be done in this domain by thresholding. The assumption is that the important causes should have larger amplitude in this independent space.

Fig. 6.12 shows the idea of implementing ICA for cochlear implant speech pro-

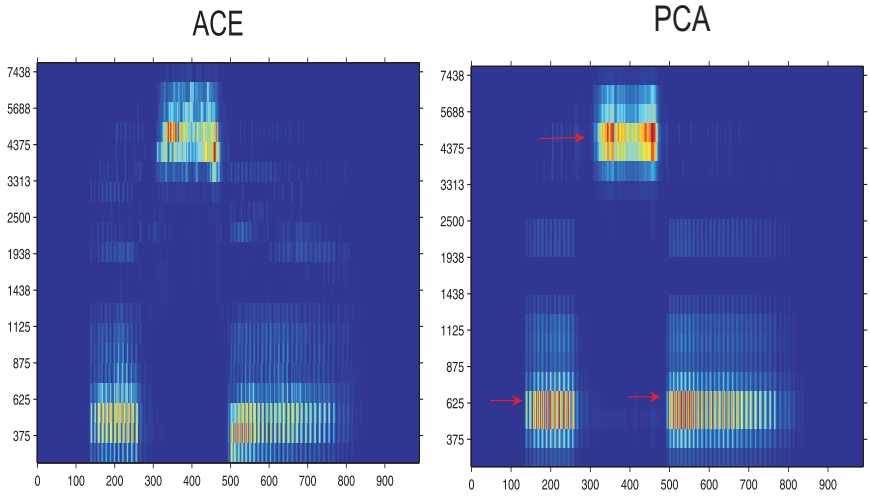


Figure 6.11: *The spectrogram of ACE and PCA output. The left panel is the ACE algorithm set to choose 12 channels out of 20 channels, as used in the cochlear device. The right panel is the signal after PCA processing and set to choose 12 maxim unit spectra for stimulation, based on the magnitude of the components (eigenvalues). The arrows on the right panel show that the formants, which are peaks in the frequency spectrum, are enhanced by PCA based selection*

cessing, the SPARSE algorithm. Speech is first fed into filter banks and envelopes are extracted. The speech envelopes in different channels are then processed by PCA. PCA can work as pre-processing for ICA to get uncorrelated channels (whitening) and the dimensionality of the signal can also be reduced. ICA then transforms the data into independent space where each channel becomes independent. The causes of the data are thus disclosed through higher order statistics. A threshold can be applied to these independent channels. The reconstruction of the speech spectrum envelope with the inverse ICA transform can then be used for cochlear implant stimulation. The output of the spectrum envelope is supposed to be more sparse than the ACE output. As ICA extracts independent channels out of the speech spectrum, the information between these independent channels is minimal. The FastICA package (Hyvarinen & Oja, 2000) is used in the algorithm, but there are many other methods which could produce similar results. FastICA is a fixed point algorithm, it converges fast and can be potentially implemented as an online algorithm.

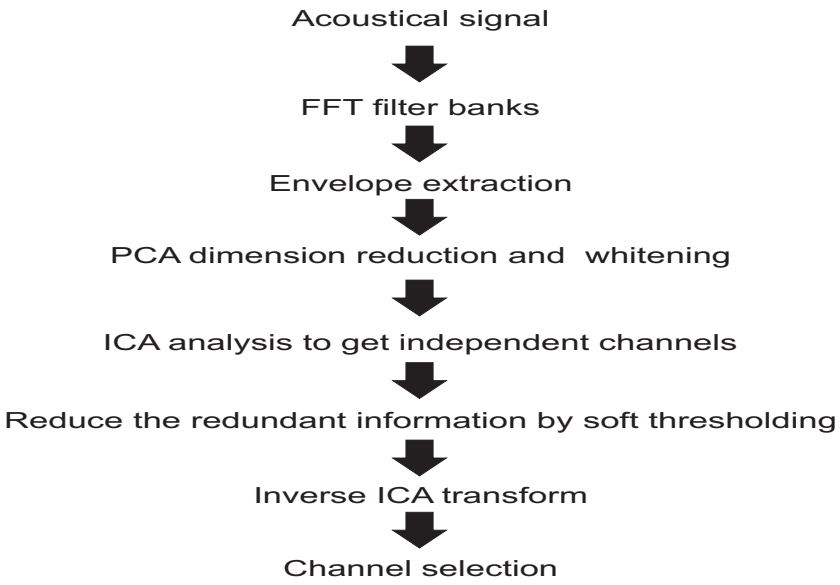


Figure 6.12: *The application of ICA in the SPARSE algorithm. The ICA is used in the SPARSE algorithm to derive independent channels and select the most essential information from the a sparse space. Inverse ICA produces a more sparse spectrogram and it can be used for stimulation of cochlear implants.*

The stimuli produced by ICA may have the following advantages:

1. Reduced channel interaction. At each moment, only few of channels are active, which is very much like the auditory neurons. One channel stimuli will hardly affect the other stimuli.

2. Saving energy. The stimuli produced by ICA are much less intense than stimuli produced by ACE, reducing unnecessary stimuli. This can save the energy of electric stimulation, which is one of the main drains on the battery of a cochlear implant.

3. The stimuli may reduce gaussian like noise. A faithful representation of the acoustical signal may make the speech recognition task harder in noisy environments. A selection of independent channels with thresholding would mean less chance of transferring unwanted noise to the auditory nerves.

Fig. 6.13 shows an example of the proposed speech processing strategy for a Vowel-Consonant-Vowel word /aga/ and the results of each intermediate step.

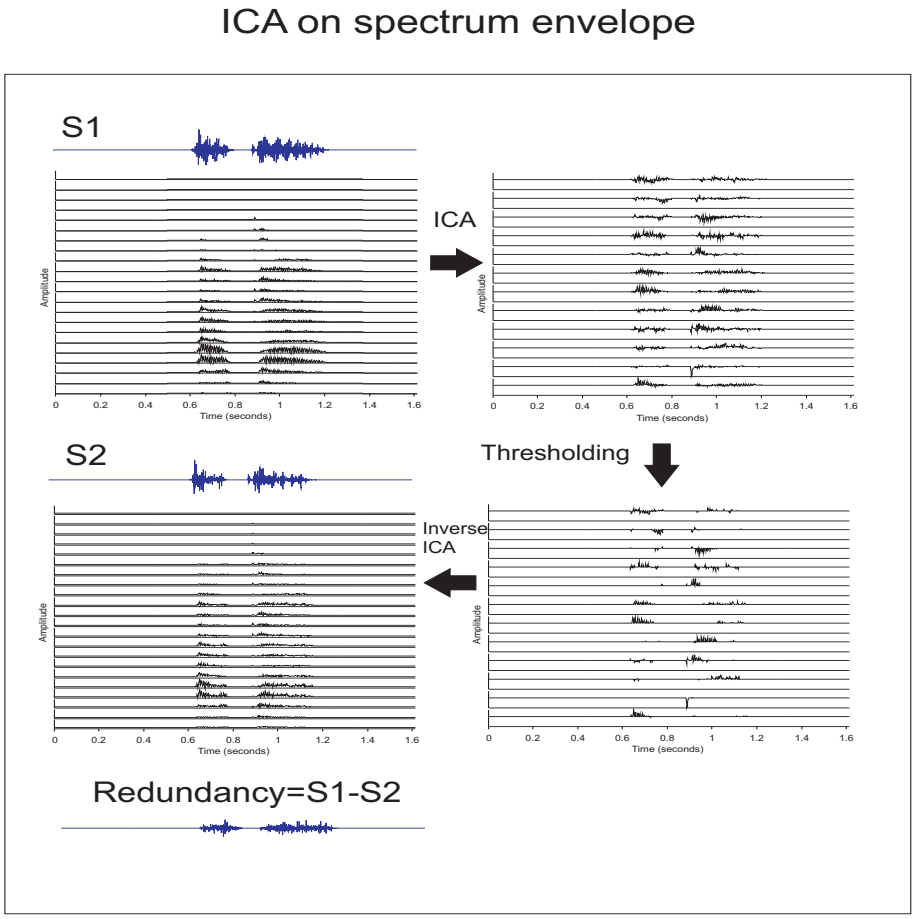


Figure 6.13: *Example of spectrum analysis by ICA. The /aga/ spectrum envelope signal is first analyzed by PCA and dimension reduction can be from 22 to 12. The whitened spectrum is then transformed into 12 independent vectors. Thresholding can then be done on the independent vectors and then transformed back to the time-frequency domain. The reconstructed spectrum has a much enhanced part for the consonant /g/. The redundancy can be defined by the difference between S1, vocoder signal of ACE, and ICA processed spectrum, S2. It mainly includes the constant vowel parts.*

The envelope spectrum is derived based on the ACE strategy. The spectrum is first processed by PCA and dimensionality has been reduced from 20 to 12. ICA can then process the whitened spectrum envelope and transform it to the independent space, where the causes of the data can be disclosed. Thresholding can then be done on the independent channels, reducing the components less than certain threshold. Reconstruction can then be done by the inverse transform of ICA. The reconstructed signal has a character of enhanced consonants, as reduced redundant parts are mainly vowels. Fig. 6.14 shows that the spectrograms of the ACE and sparse representation of the spectrum by ICA thresholding. The sustained vowels have been reduced. Onsets and consonants are enhanced by the thresholding. This is important for efficient coding. As limited resources only focus on the non-constant parts, the coding resources can then be most efficiently used. Fig. 6.15 shows that the result of the sparse transform is sparse and it can be measured by kurtosis.

6.3.3 Derivation of the SPARSE algorithm

A signal $\hat{X}$ can be seen as two parts, the most essential part X' to represent the original signal, and the redundant or noisy part N.

$$X1 = \hat{X} + N \tag{6.1}$$

First, we use PCA to reduce the dimensions of the input signal $X1$, creating a transformed signal (X), containing only the main principal components (eg 8 out of 20).

$$X = PCA(X1) \tag{6.2}$$

We then perform a sparse transformation (W) using ICA on the relatively clean signal X:

$$S = WX \tag{6.3}$$

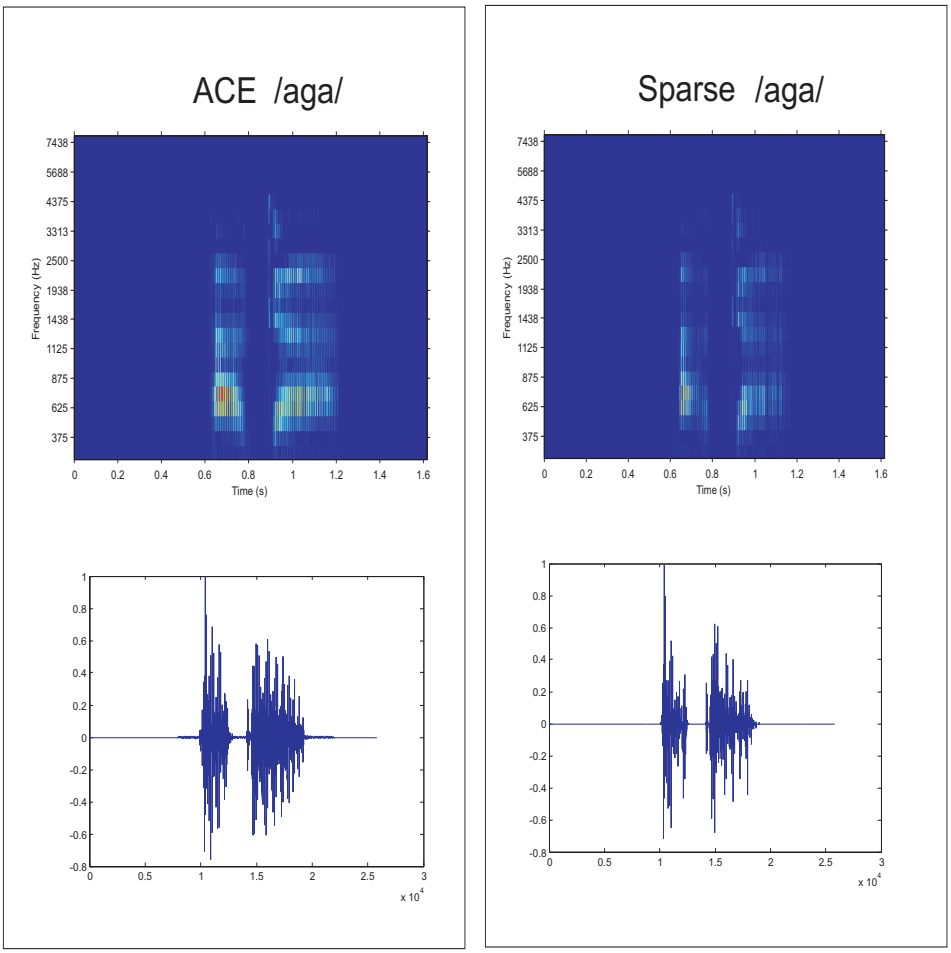


Figure 6.14: *Example of sparse spectrum by ACE and SPARSE . The left panel shows an example of ACE output for /aga/ and the right panel shows the output after sparse processing and thresholding. The waveforms are the simulation of spectrum output of ACE and output of SPARSE algorithms. The spectrum envelope based on SPARSE strategy is quite sparse and it keeps the essential components of consonants and vowels, as can be seen through the simulated waveforms above.*

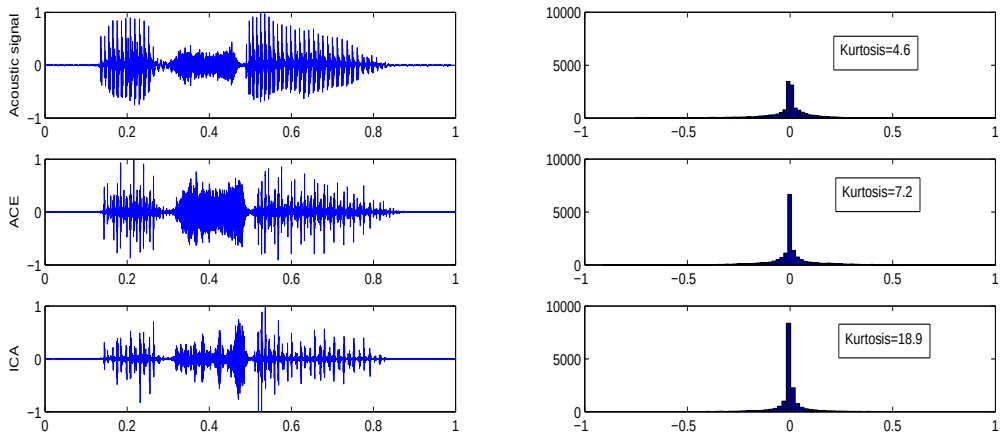


Figure 6.15: *Kurtosis and reconstruction of ACE output by SPARSE. In the left panel, the top is the original acoustical signal, middle is the ACE produced signal, and the bottom is the signal based on ICA. The reconstruction only used 8 independent channels. The right panel is the corresponding histograms. The value of kurtosis is also indicated, a measure of sparseness.*

The ICA transformation must reduce or remove redundancy in S, because the independent components are calculated to be as independent as possible. The independence can be achieved by making the output vectors as super-gaussian as possible. And thus S is sparse. The independent components of S can be expressed in order of reducing magnitudes of independent components as :

$$S = S1 + S2 + S3 + ...Sn \tag{6.4}$$

We assert that the larger independent components of S are more important that the smaller independent components, as the meaningful events are expected to have bigger values in the independent space. We next remove the smaller components by applying a threshold. Components below the threshold are set to zero.

$$S - S_{threshold} = S^{'} \tag{6.5}$$

The maximum likelihood method can be used to derive the $S_{threshold}$ (Hyvarinen, 1999). If we assume the distribution of essential components S is super-gaussian, and the distribution of the redundant part and noise is gaussian, we are then able to estimate essential components given S. The maximum likelihood methods gives the following estimator for essential parts (Hyvarinen,1999):

$$S' = S_{Essential} = \frac{sign(S)max(0, |S| - a\sigma^2)}{(1 + b\sigma^2)} \tag{6.6}$$

Assuming that the essential parts follow the distribution as shown in Eq. 6.7:

$$P(S_{Essential}) = Cexp(-as^2 - b|s|) \tag{6.7}$$

Where $a, b > 0$ are parameters to be estimated, and C is a scaling constant. We choose $a = 3$ and $b = 8$ from experiment trials of listening the processed sounds. σ^2 is the variance of noise, which needs to be estimated. Here for simplicity, we use the variance of S to approximate the variance of redundant or noisy parts, and our results shows it is a reasonable approximation. The nonlinear threshold function is plotted in Fig. 6.16.

The neuron firing pattern will be simplified as S', representing the most essential parts of the signal X. We can then perform an inverse ICA transform W^{-1} on S' to obtain an estimation of the external stimuli $\hat{X}$.

$$\hat{X} = W^{-1}S' \tag{6.8}$$

$\hat{X}$ then contains the most essential parts of X, which should be used to generate the electrical pulse sequences for cochlear implant stimulation.

The SPARSE algorithm can thus be achieved by the following steps:

1. Get the sparse representation of external stimuli X by $S = WX$. W can be determined by many sparse coding or independent component analysis methods, such as FastICA approach;

2. Apply the shrinkage function as defined in Eq. 6.6 on S;

3. Invert transform based on $\hat{X} = W^{-1}S'$;

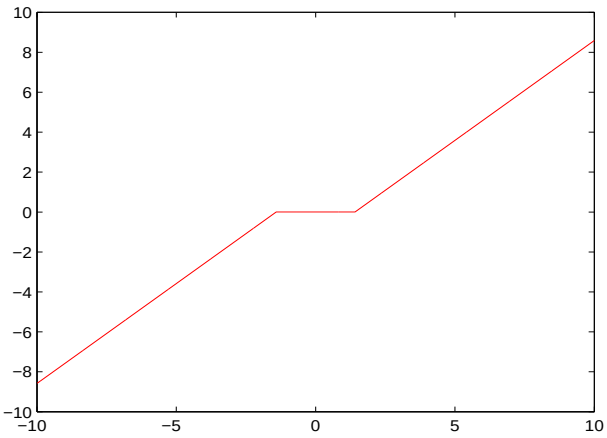


Figure 6.16: *Shrinkage function, a soft thresholding method. The shrinkage function is based on maximum likelihood method, with assumption that the distribution of the essential information is super-gaussian and the redundant or noisy parts is gaussian. Details can be found in (Hyvarinen, 1999)*

4. Apply $\hat{X}$ for CI electrical stimulation.

Fig. 6.17 shows an example of output of the SPARSE algorithm for a cochlear implant. The signal was first processed by PCA and only 12 principal components were selected for ICA processing. Eight independent component vectors were used to reconstruct the signal. Clearly, the signal is more sparse and SNR is higher than that of ACE produced signals.

6.4 Discussion and Conclusions

Cochlear implants have limited dynamic range and need to stimulate the auditory neurons sparsely. PCA can help reduce dimensions of the data to suit the limited dynamic range and ICA can help to produce a sparse representation of the spectrogram envelope for CI electrical stimulation. The proposed algorithm is named SPARSE. The SPARSE algorithm is motivated purely based on sparse coding theory.

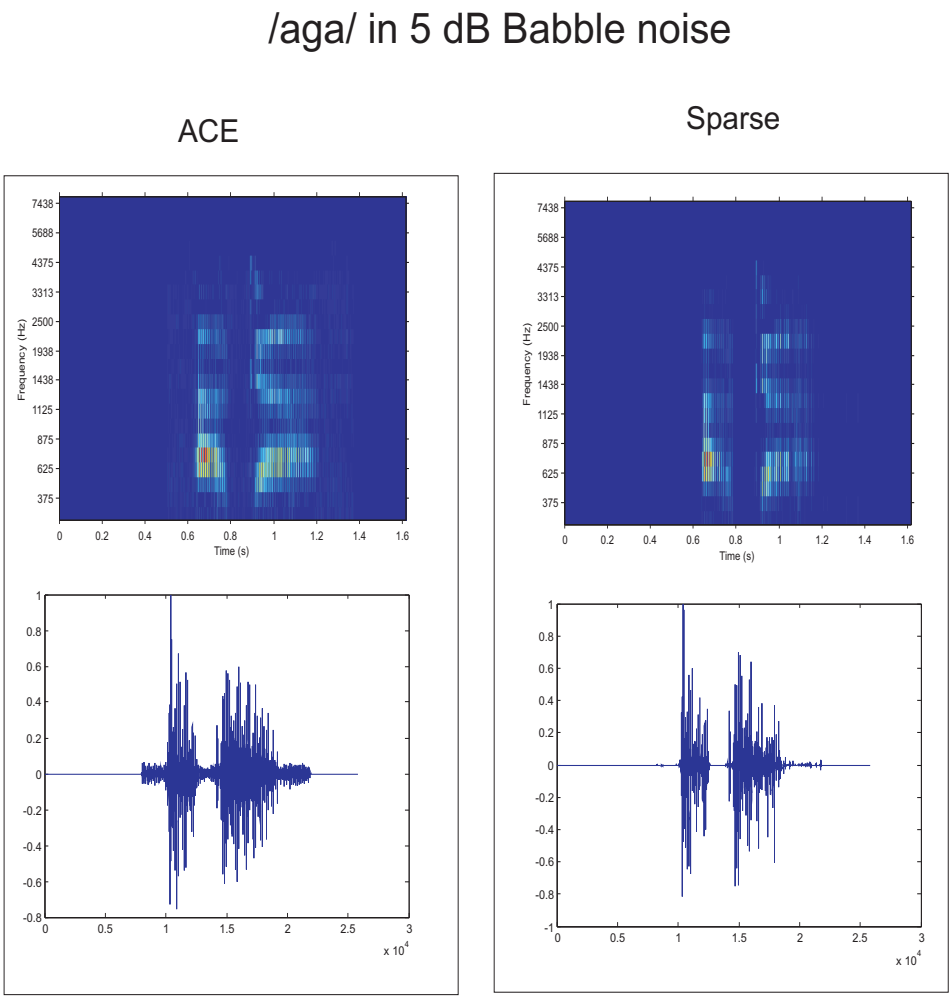


Figure 6.17: *Example of the SPARSE algorithm in 5 dB noisy condition. The left panel is the output of the ACE algorithms, which chooses 12 out of 22 channels. The right panel is the signal after ICA de-noising. First, PCA did dimension reduction to 12 channels and then 12 independent channels were used to reconstruct the stimulation. The representation based on ICA is much cleaner than the output of ACE.*

The soft thresholding in the independent space could help cochlear implant users to get better performance. The next chapter will focus on the objective and subjective evaluation of SPARSE and to see whether sparse stimuli would help to improve the speech recognition performance of both CI users with electrical stimuli and normal hearing listeners under simulation of CI. The simulation is a reconstruction of the acoustical signal based on the spectrum envelope only, also called the vocoder method.

.

Chapter 7

Experimental evaluation of SPARSE algorithm for cochlear implants

7.1 Introduction

The aim of the present research is to explore the application of sparse coding principles to the processing within a cochlear implant. These principles would determine what information in noisy speech should be extracted and used to excite the electrode array within the cochlea. We hypothesized that reducing redundancy in the signal, making it more sparse, would improve speech recognition scores.

The proposed sparse coding strategy, SPARSE, was based on a combination of ICA and PCA analysis, both operating on the spectrotemporal envelope of the speech signal. The strategy first reduces the redundancy in the spectrum by using PCA and then applies a nonlinear threshold to the output of the subsequent ICA. The reconstruction of the spectrum is realized by the inverse transform of ICA. Thus the spectrum for stimulating auditory neurons becomes more sparse and certain features of speech are also enhanced.

Possible reasons for the improvement could be: (1) the new strategy can reduce interaction between channels; (2) it selects primarily the essential information in speech for simulating auditory neurons; (3) as only the most essential

information is selected, the limited dynamic range of cochlear implant users can be used efficiently; (4) it might force neurons to fire more sparsely, and hence more physiologically, compared to neurons stimulated by the present commercial algorithms.

In order to test sparse coding for cochlear implant speech processing, we developed an algorithm and implemented in a cochlear implant. Subjective experiments have been done to compare the performance in both normal hearing listeners and cochlear implant users with Advanced Combined Encoder (ACE) and the proposed SPARSE algorithms.

7.2 SPARSE algorithm parameters

The input to the proposed algorithm is the spectrum of speech. The envelopes extracted from M channels are processed as a matrix by PCA and then by ICA. The PCA processing will select only K dimensions out of M analysis frequency channels. Here we take $K = 12$ and M=22, as 22 frequency analysis channels are normally used in the traditional ACE speech processing algorithm, and only 12 channels are selected for stimulation by default. And 12 channels are enough to represent the speech signals.

We also choose parameters $a = 3$ and $b = 8$ in the experiment. These parameters are derived based on subjective listening trials from the output of the SPARSE algorithm. These parameters can be further optimized by individual subjects. For simplicity, we keep these parameters same for all the subjects.

The output of PCA is then fed into the ICA analysis, K independent components are selected as the independent channels. Here the speech spectrum envelope is transformed into an independent space. In this independent space, soft thresholding can then be done based on the methods of Hyvarinen *et al.* (1998). The output of thresholding is then processed by an inverse ICA transform. A sparse representation of the spectrum envelope is thus produced and can be used for the channel selection process and cochlear implant stimulation.

7.3 SPARSE evaluation

In order to test the efficiency of the proposed speech processing algorithm both in quiet and in noise conditions, objective evaluations are need to test the output of the SPARSE algorithm. The direct output of the cochlear implant speech processing algorithm is the spectrum envelope. The spectrum envelope can be reconstructed to waveforms based on vocoder algorithms [1].

One of the important factors considered in the proposed algorithm is the sparseness of the reconstructed signal, which could potentially make the neurons fire sparsely and implement the sparse coding theory for cochlear implants. Sparseness can be quantified by kurtosis of the signal (Field, 1994).

The sparseness can be calculated through kurtosis. Kurtosis can be measured when $\hat{s}$ is normalized (mean is zero and variance is 1).

$$kurtosis = \frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i)^4 - 3 \tag{7.1}$$

Taking the output simulated waveform as a whole, kurtosis is calculated from the entire time series signal. If the kurtosis increases then the sparseness of stimuli is enhanced.

7.3.1 Speech materials

Speech tokens were drawn from 9 VCV (Vowel Consonant Vowel) words: (/aba/, /ada/,/aga/,/aka/,/ala/, /ama/, /ana/, /apa/, /ata/)[2]. Two different noises are used in three different noisy conditions (+15, +10, +5 dB). One noise is 8-talker babble modulated noise, which was used in Cooke (2006). The babble modulated noise is produced by modulating the speech-shaped noise (based on the VCV words) with the 8-talker envelopes. The 8-talker envelope is generated based on the TIMIT corpus (Cooke, 2006). The other is babble noise which is the sound

[1] Envelopes in different channels can be modulated either by filtered noise or sinusoids with the same centre frequency as the frequency channels. A simulation can thus be derived by summing the modulated envelopes together (Shannon *et al.*, 1995)

[2] To make the experiment short which is around one and half an hour, only 9 VCV were used.

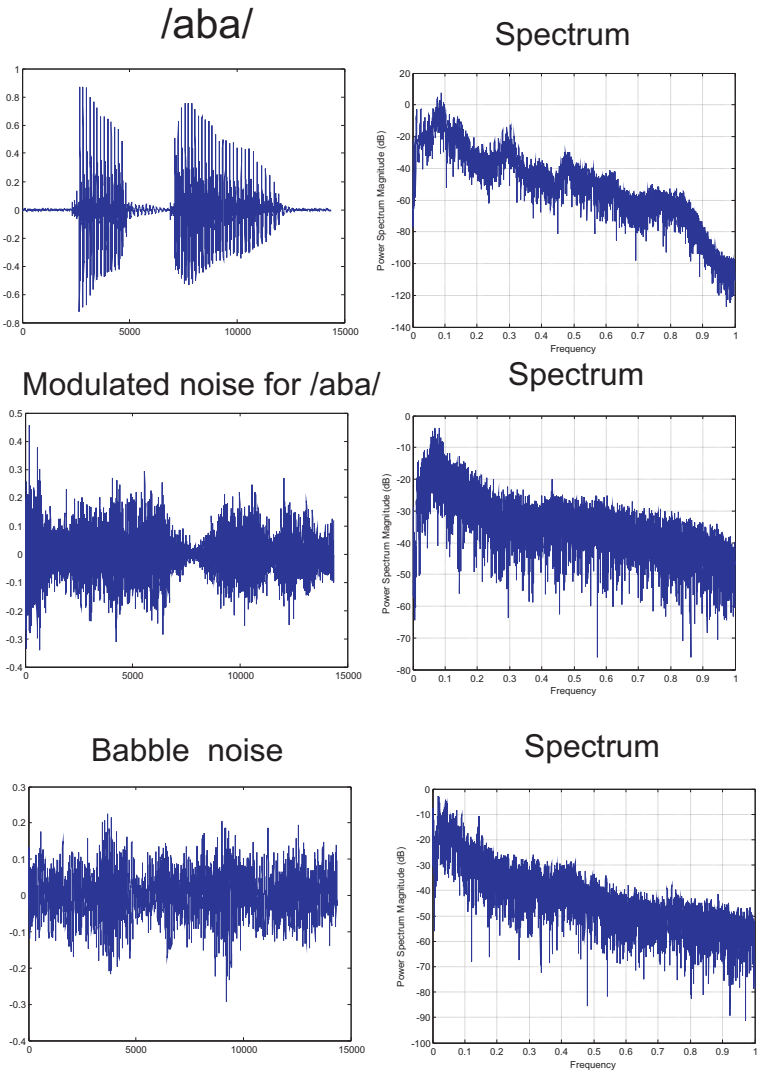


Figure 7.1: *Example of VCV word and the corresponding masking noises. The modulated noise has a similar spectrum energy distribution as the VCV sound. The energy masking from modulated noise is greater than that that of the babble noise.*

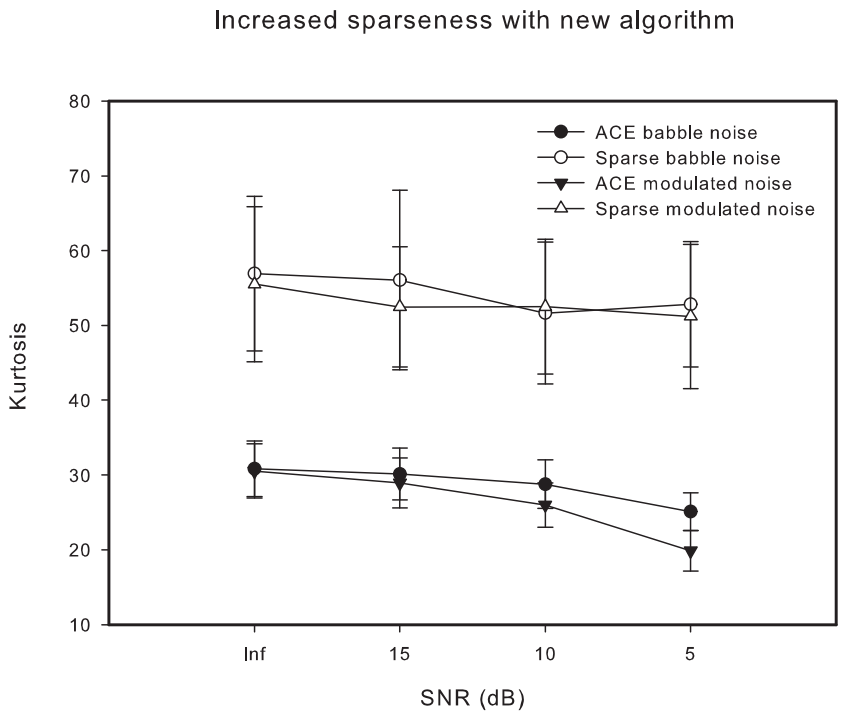


Figure 7.2: *Increased sparseness with SPARSE. The spectrum output of ACE and SPARSE algorithms can be simulated as an acoustical signal using noise vocoder technique The sparseness of these simulated waveforms can thus be measured by kurtosis. The output with the SPARSE algorithm is more sparse than the ACE output. The aim of making the spectrum more sparse is achieved.*

of 100 people talking in a canteen, with radius approximately two metres (SPIB, 2000).

Fig. 7.1 shows the waveform and corresponding noise. The modulated babble noise is different for each individual VCV word. The energy masking from modulated noise is supposed to be greater than that of stationary noise.

7.3.2 Increased sparseness

The sparseness of the spectrum from the SPARSE algorithm is expected to be higher than that of ACE output. Fig. 7.2 shows that the average kurtosis of the output of SPARSE is much higher. The enhanced signal after SPARSE processing is more sparse than the output of the ACE algorithm.

The kurtosis shown in Fig. 7.2 in average is higher than the kurtosis value of original acoustic signal. This is because that the cochlear implant speech processing only uses limited envelope information of acoustical signal to stimulate auditory neurons. The stimuli used by ACE is already a sparse signal compared to original acoustical signal. But due to the bottleneck existed in the cochlear implant, further sparse processing is needed in order to deliver maximal information with limited bandwidth.

7.4 Subjective experiments

Although the increased kurtosis of the enhanced noisy signals could indicate the intelligibility could be improved (Li & Mark, 2006), subjective experiments are needed to test the intelligibility of the sparse speech. The distortion introduced by the SPARSE algorithm can be indicated by the speech recognition score of listeners. The SPARSE algorithm is expected to be especially helpful for cochlear implant users, as the sparse stimuli are more useful for CI subjects, if sparse coding principles can be implemented in the electrical stimulation of CIs.

Normal hearing subjects also participated in the listening experiment to see the effect of the algorithms on normal hearing listeners. Ethics was also applied for the experiment and permissions were granted and all of the subjects signed agreement forms.

The speech materials used are the same as the speech materials used in the objective evaluation: nine VCV words, two different noise, four different noise conditions (Quiet, +15 dB, +10 dB, +5dB). Each item was presented four times to each subject and only the last three of them were calculated in the score.

7.4.1 Experiments I: Normal hearing subjects and sparse stimuli

Normal hearing people can listen to the simulated sound of the output of cochlear implant processing. The sounds are the vocoder output of the spectrum. Seven normal hearing subjects participated in the experiments.

7.4.1.1 Results and Discussion

Fig. 7.3 shows that the SPARSE algorithm is helpful for normal hearing listeners in 5 dB SNR. The difference between SPARSE and ACE is statistically significant ($P = 0.021$) for $SNR = 5\ dB$ in the babble noise condition. The difference between SPARSE and ACE is not statistically significant for other conditions [1].

Individual results are plotted and shown in tables in the Appendix A. Fig. 7.4 plots the differences between the score of subjects with ACE and the score with SPARSE. The positive difference is the improvement achieved by using SPARSE. It further shows that the speech recognition score has been improved when the baseline performance of subjects was poor, say when the speech recognition score was below 90%.

In the next section, the same algorithms are used for cochlear implant subjects to see the effect of the new algorithm on their speech recognition score. The cochlear implant users will listen through the VCV words using their own CIs. The electrical sequences will be saved in the computer and sent to the auditory neurons through electrodes.

7.4.2 Experiment II: Cochlear implant users and sparse stimuli

7.4.2.1 Speech materials and subjects

The same speech materials are used as in the experiment for normal hearing subjects; nine VCV (Vowel Consonant Vowel) words: (/aba/, /ada/, /aga/, /aka/, /ala/, /ama/, /ana/, /apa/, /ata/) and two different noises (babble noise and modulated noise); four noise conditions (Quiet, 15 dB, 10 dB, 5 dB); three times per item are presented to each subject.

The stimuli are presented to the subjects through the NIC streaming software (See Appendix B), which can deliver the electrical pulses sequence to the cochlear implant of subjects. The sequences are produced based on the CI mapping of the subjects used daily. The sequences were saved on computer before streaming.

[1] There is possibly a ceiling effect for normal hearing subjects in other conditions as shown in the Appendix Fig. 9.1. And no transformation was done on the data

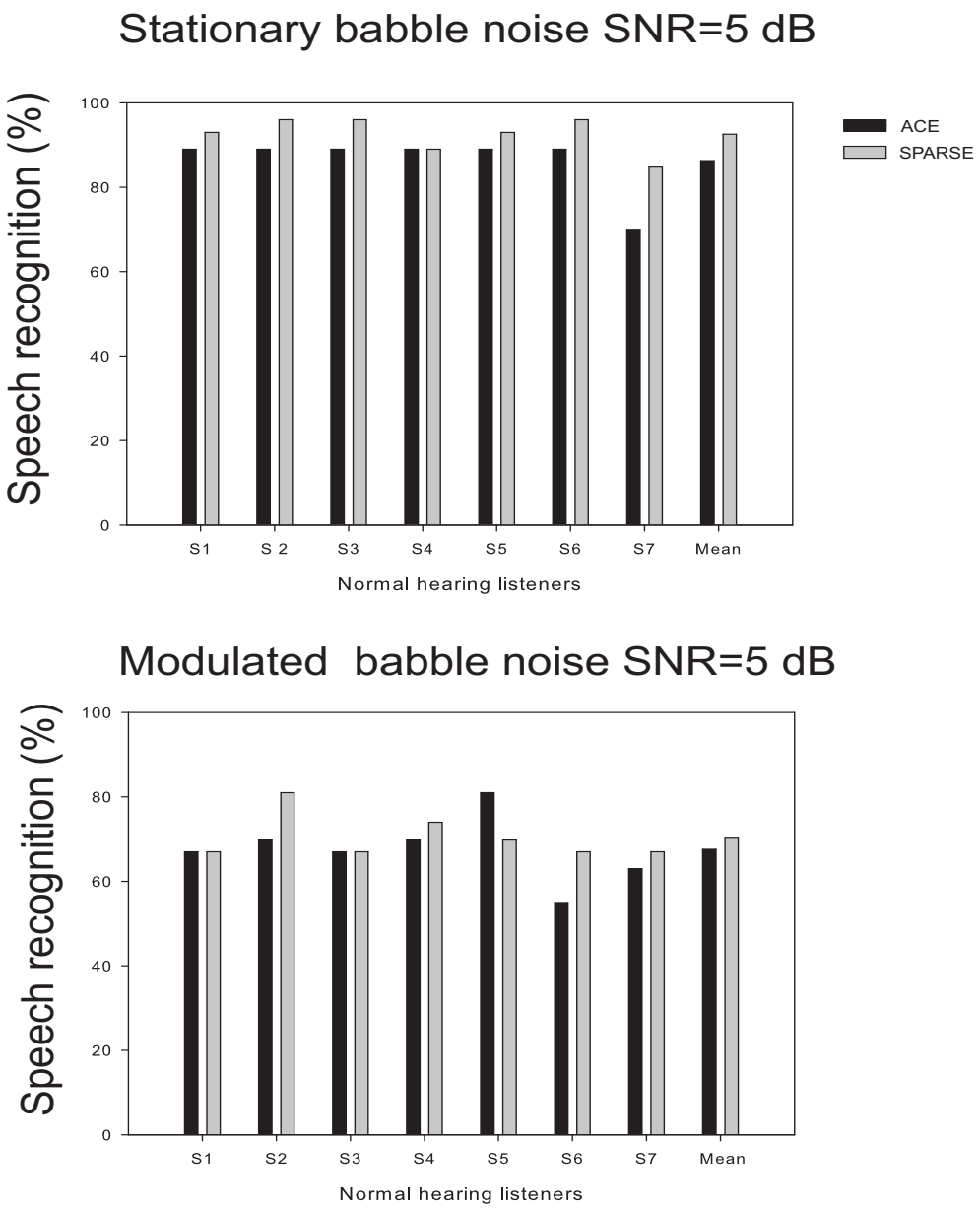


Figure 7.3: *Normal hearing subjects with SNR = 5 dB. The SPARSE algorithm is found to be especially helpful to listener whose speech recognition score is relatively poor or when SNR is lower. The difference between SPARSE and ACE is statistically different ($P = 0.021$) in babble noise condition (SNR = 5 dB).*

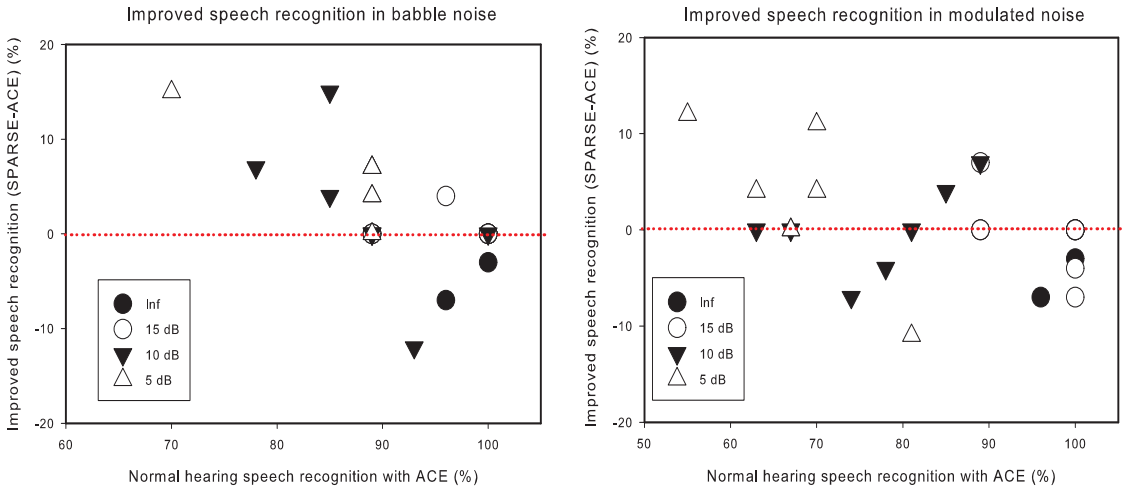


Figure 7.4: *The improved speech recognition score by the SPARSE algorithm for normal hearing listeners. The improved speech recognition score is defined by the positive difference between the speech recognition scores with SPARSE and ACE. The left panel is the babble noise condition and right is the modulated noise condition. X-axis is the speech recognition score for normal hearing with ACE algorithm in four different signal noise ratio conditions. Y-axis is the score difference between scores with ACE and score with SPARSE. There is less improvement when the speech recognition performance is high. Improvement is mainly for the subjects with poor performance in noisy conditions.*

The experiments also got ethical approval from the ethic committee of Institute of Sound and Vibration, University of Southampton.

7.4.3 Results and Discussion

Individual results are plotted and shown in tables in Appendix A (Table 9.4 to Table 9.6). Fig. 7.5 plots the difference between the scores of CI subjects with ACE and SPARSE in stationary and modulated babble noise. It shows that the speech recognition score has been improved when the baseline performance of subjects is poor, say when the speech recognition score is below 40 percent with ACE. There is no statistical difference across all the conditions, mostly due to the big variances across individual CI subjects.

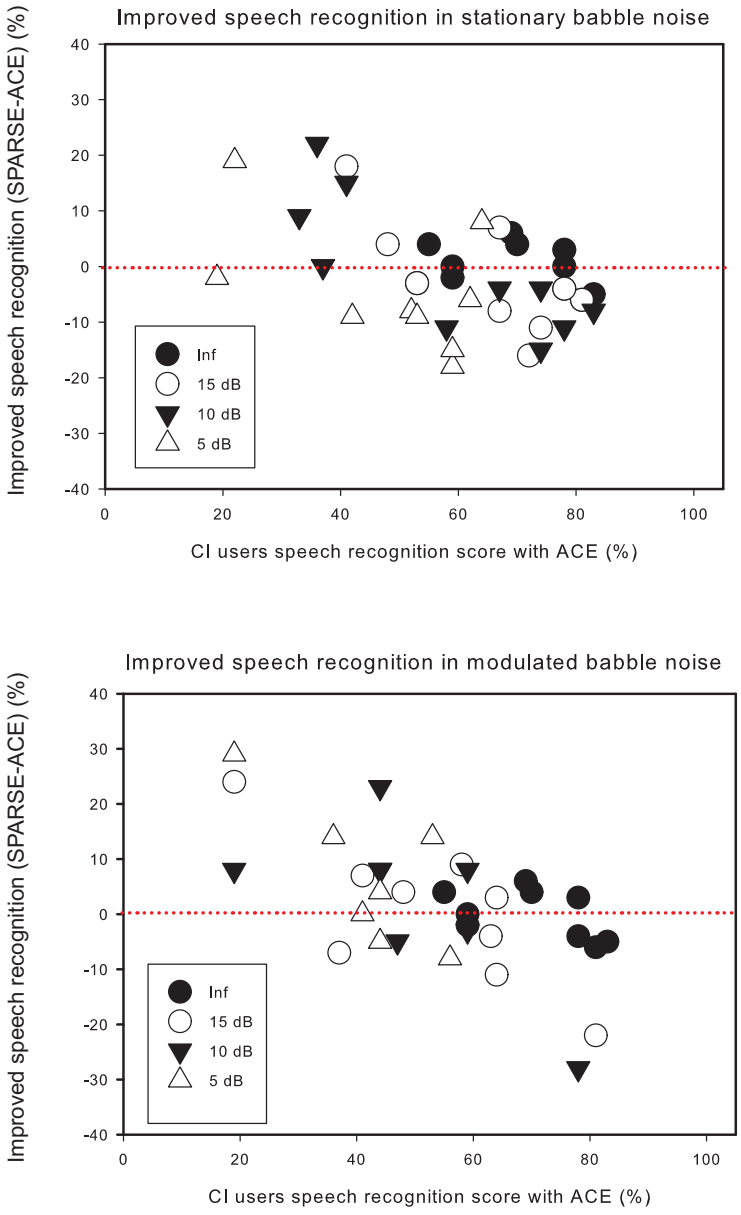


Figure 7.5: *The improvement speech recognition by the SPARSE algorithm for CI users. They are defined by the difference between the speech recognition scores with SPARSE algorithm and ACE. The upper panel is the babble noise condition and the lower panel is the modulated noise condition. X-axis is the speech recognition score for cochlear implant users with ACE algorithm in four different SNRs. Y-axis is the recognition score difference between ACE and SPARSE. The improvement can be seen in the upper left corner of each figure, indicating that SPARSE is useful for CI users when SNR is low.*

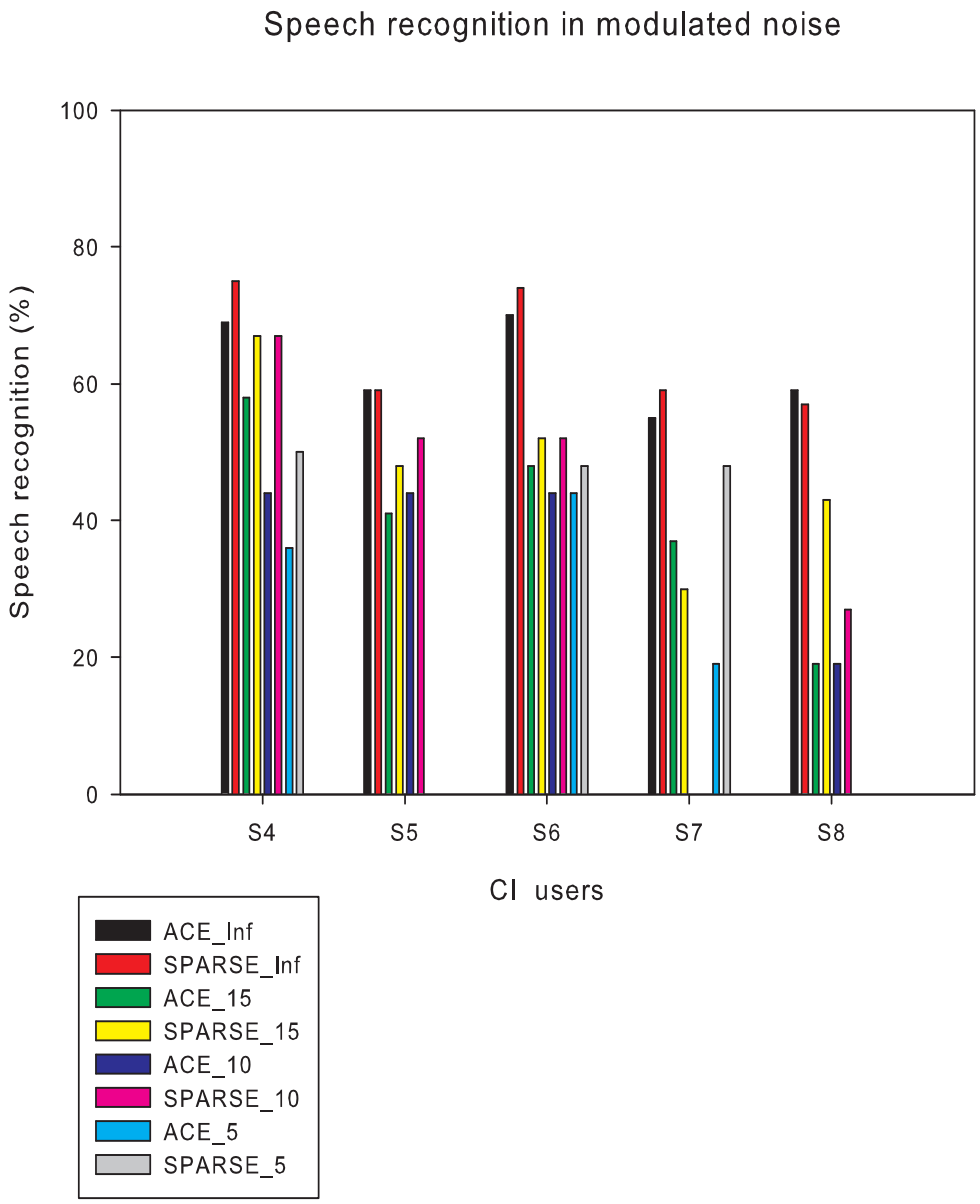


Figure 7.6: *Speech recognition score for CI users whose baseline performance is lower than 70%. The X-axis is the subjects whose speech recognition score is lower than 70% in quiet. Y-axis is the speech recognition score with different conditions. The SPARSE algorithm in average helps improve the speech recognition score across different conditions for these subjects. Some of these subjects reported the quality of SPARSE output is more clear.*

The average performance in CI users is much worse than that of normal hearing listeners in noise. It shows that the noise can have a much bigger effect on the speech recognition performance of CI users. The SPARSE algorithm is more helpful to improve the speech recognition of CI users when their baseline speech recognition score performance is low as shown in Fig. 7.6.

Both the results in normal hearing subjects and CI users show that the SPARSE algorithm can help improve speech recognition when the SNR is low and baseline speech recognition score is low. One possible reason for this is that when the speech recognition score is high, the listeners are able to resolve most of information presented and reach ceiling performance. Reducing redundant information will reduce speech recognition score.

In noisy conditions, when speech recognition performance is low, listeners may not be able to explore all the information presented, as the information presented is a heavy mixture of speech and noise. Under this condition, reducing the information which could not be used by the listeners will improve the speech recognition performance, because the encoding resources of the auditory system can focus on the most useful information left afterwards.

Also in quiet or greater SNR condition, the selection of thresholding could have reduced too much speech information. When the SNR is lower, more noise is reduced than the speech itself, the speech recognition performance is thus improved. It remains to be seen how to design a better thresholding which can reduce less speech information in quiet and more noise in noisy situation. The results in normal hearing subjects show that the SPARSE algorithm helps when SNR is low and speech recognition performance is lower with ACE. Fig. 7.4 shows that when the speech recognition score is below 70% in noise conditions, the SPARSE algorithm can help improve performance.

Another possibility is simply the statistical artefact of regression to the mean, where a listener who scores low by chance on one test will tend by chance to show better performance on the other.

As shown in Fig. 7.5 and Fig. 7.4, the maximal percentage of improvement (30%, in modulated noise) for CI users is higher than the improvement for normal hearing listeners (15%, in babble noise), suggesting SPARSE is could be particularly helpful for CI users.

This is interesting as improvement for the lower end performance users is quite a challenging task for current cochlear implant research. The SPARSE algorithm could help these subjects to get a relative high speech recognition score, which is comparable with the high end performance CI users. Such improvement will help the poor end performance users in noisy condition. The SPARSE algorithm does not help in quiet conditions and it seems that the sparse stimulation does not benefit cochlear implant users in the quiet conditions. It could be that the current stimuli are already sparse in quiet condition anyway.

Limitations on improvement were undoubtedly caused by lack of familiarity of users with the novel processing. CI users typically require days or weeks of familiarity to benefit from speech processor improvement. This algorithm can potentially be adapted for hearing aids users as well. Further research will focus on the fine adjustment of the parameters and its real time implementation.

Chapter 8

Conclusions

8.1 Overview

This thesis firstly starts from a key concept among the principles of perception: environmental statistics have an important effect on perception. This is mainly because our environment is highly structured and redundant. One efficient way to explore the redundancy is through higher order statistics. In terms of speech perception, similar principles should be used. One background aim of this thesis is to introduce higher order statistics into the field of speech perception research.

Sparseness is especially one of the important features of speech and it can be quantified by kurtosis. By comparing and calculating the kurtosis of speech data, both objective and subjective experiments have shown that speech recognition indeed is affected by the high order statistics of the speech.

Manipulating of higher order statistics can help to improve speech recognition as it has been used in signal processing such as blind source separation and project pursuit algorithms. The advantage of exploring higher order statistics can be further explored for the signal processing for hearing aid devices.

This thesis applies the key concept of sparse coding theory in the cochlear implant signal processing. Sparse coding theory suggests that sparse firing patterns will be efficient to encode the external stimuli. Such sparse firing could

be achieved by using sparse electrical stimuli, which also reduce redundancy of stimuli. Based on PCA and ICA analysis, a soft thresholding technique is applied directly in the independent domain, where the speech spectrum is transformed to independent channels. The SPARSE algorithm has successfully produced a sparse version of the envelope spectrum with increasing sparseness. Simulation of sound waveforms has a much higher kurtosis than the current stimulus strategy. Subjective experiments also show improvement of speech recognition score for those whose baseline performance is low. The SPARSE algorithm could help reduce channel interaction, using less energy to stimulate the auditory neurons.

8.2 Discussion

8.2.1 Investigation of higher order statistics in speech perception

The idea that environmental statistics are important for perception has been developed over a long time (Field, 1987; Simoncell, 2003; Simoncelli & Olshausen, 2001). However, it has not been applied to speech perception research directly. The direct investigation of the relationship between high orders statistics and speech perception would help to increase knowledge on how the auditory system can explore the redundancy in natural speech and how the auditory system can cope with daily complex environments in general.

More and more evidences have shown that the understanding of auditory system cannot rely on the simple stimuli anymore (Plomp, 2002). Controlling of stimuli with different high order statistics would help to investigate the relationship between higher order statistics and perception.

It is quite likely that the unified perception principle, redundancy reduction, is applicable to auditory system (Barlow, 2001; Field, 1987; Lewicki, 2002), even in a very periphery level (Olshausen & O'Connor, 2002). The redundancy exploration needs to be implemented through higher order statistics analysis, as finding the

structure of a stimulus involves more than the relationship two points (second order statistics).

Further neural physiological evidences on how higher order statistics affect periphery neurons are needed to assist the designing of experimental stimuli on speech perception. It has been found that neuron firing patterns can adapt to stimulus statistics (Dean *et al.*, 2005). The combination of neurophysiological experiments with speech perception results could further extend the understanding of our auditory system.

8.2.2 Information capability matching

The SPARSE algorithm tries to reduce the redundancy of speech and use it as stimuli for CI users. Subjects whose speech recognition in noise or in quiet is higher did not get too much improvement. Some of their speech recognition even become slightly worse. The improvement is mostly seen for those whose baseline performance is lower.

One assumption is that there could a mismatch between information processing capability or channel capacity and information sent to listeners. For those whose baseline performance is poor, the information contained in the sparse stimuli could match their limited capability of information processing. And thus their speech recognition score can be improved by the SPARSE algorithm.

The SPARSE algorithm could be specially helpful for the lower end performance CI users. It is possible that the SPARSE can be adjusted to change certain parameters to fit for the capacity of the CI users. This could have a much wider application for speech processing for hearing impaired users. The high variance among CI subjects' speech recognition performance has been seen as a big challenge in CI research. Some CI users can perform as well as normal hearing subjects, while some have great difficulties in understanding speech in any noise. The SPARSE algorithm might help these lower end performance users by making the stimuli sparse and using their limited encoding resources of the auditory system efficiently.

8.2.3 Kurtosis and SNR

Kurtosis has been applied in predicating SNR of a sub-band of signals (Nemer, 1999). In the experiment with the project pursuit algorithm, the increase of kurtosis leads to better speech perception with project pursuit algorithm. One might argue that the increase is simply due to increased SNR. And indeed clearly the SNR increases when the kurtosis increases as the clean signal has less gaussian components. Similarly in the SPARSE algorithm, the increase of kurtosis can be seen as a de-noising process.

However, one has to notice that the aim of increasing kurtosis in both experiments (SPARSE algorithm and projection pursuit) is to purely data driven . The cost function of optimization is only related to the higher order statistics of data. There are no explicit stages of voice active detection in both algorithms.

8.2.4 Parameters of SPARSE

The SPARSE algorithm uses a threshold which was used not only to reduce noise but also some speech components. This seems counter -intuitive, but it is based on sparse coding principles and the redundant property of speech. Specially for cochlear implants, it is known that only few speech components or limited information can be transmitted to the auditory system via electrical stimulation. The macroscopic approach of ICA and PCA might provide new insight on to how to select the most necessary information for speech perception. This is important since it can be applied for general audio coding or for acoustic hearing aids.

8.3 Conclusion and future work

This thesis introduced higher order statistics into speech perception research and it shows that the sparse coding principles could be implemented into current CI processors. The SPARSE algorithm could improve the CI users whose baseline performance is poor.

Further work will focus on the refining SPARSE and making it work in real time. There could also associated neuroscience work, towards better understanding of sparse coding in the auditory system. Sparse coding research on novel speech processing algorithm for cochlear implant may not only help improve the speech recognition performance but also an ideal platform to understand how sparse and what kind of sparse is important for auditory perception.

The combination of neuroscience, psychological behaviour research and statistical machine learning methods could provide new insight on how our auditory system works.

Chapter 9

Appendix A: Results of subjective experiments

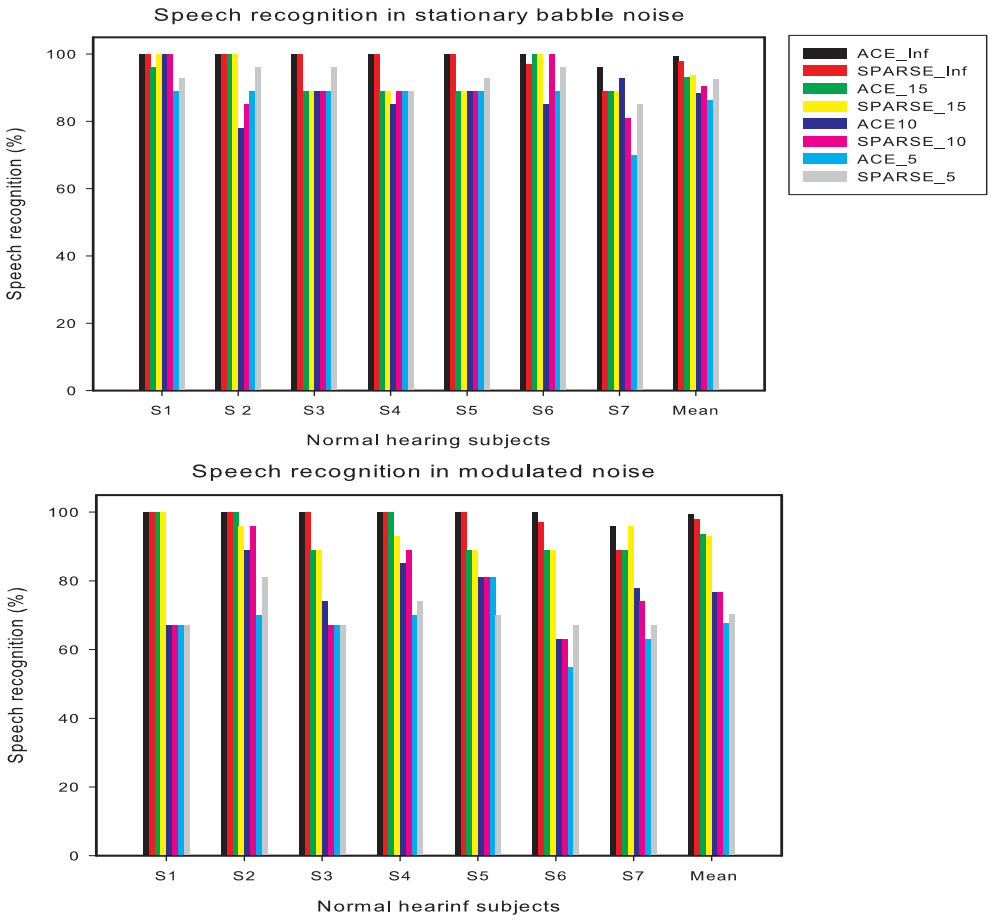


Figure 9.1: *Normal hearing subject recognition in babble noise. The upper panel is the babble noise condition and the lower panel is the modulated noise condition for all the subjects. ACE_Inf represents the condition of quiet with the ACE algorithm; SPARSE_15 represents the* 15 *dB SNR condition with the SPARSE algorithm; masking of modulated noise is stronger than the babble noise. SPARSE is helpful when the SNR is low or the speech recognition performance is relatively poor. There are ceiling effects for normal hearing subjects, specially conditions of babble noise.*

Table 9.1: Normal hearing listeners in quiet.

subjects	ACE_Inf	$SPARSE\_Inf$
S1	1	1
S2	1	1
S3	1	1
S4	1	1
S5	1	1
S6	1	0.97
S7	0.96	0.89
Mean	0.99	0.98

Table 9.2: Normal hearing listeners in babble noise. 'SP' is the abbreviation for SPARSE. The number after each algorithm (SP and ACE) is the SNR in dB.

subjects	ACE_15	SP_15	ACE_10	SP_10	ACE_5	SP_5
S1	0.96	1	1	1	0.89	0.93
S2	1	1	0.78	0.85	0.89	0.96
S3	0.89	0.89	0.89	0.89	0.89	0.96
S4	0.89	0.89	0.85	0.89	0.89	0.89
S5	0.89	0.89	0.89	0.89	0.89	0.93
S6	1	1	0.85	1	0.89	0.96
S7	0.89	0.89	0.93	0.81	0.7	0.85
Mean	0.93	0.94	0.88	0.90	0.86	0.93

Table 9.3: Normal hearing listeners in modulated babble noise. 'SP' is the abbreviation for SPARSE. The number after each algorithm (SP and ACE) is the SNR in dB.

subjects	ACE_15	SP_15	ACE_10	SP_10	ACE_5	SP_5
S1	1	1	0.67	0.67	0.67	0.67
S2	1	0.96	0.89	0.96	0.7	0.81
S3	0.89	0.89	0.74	0.67	0.67	0.67
S4	1	0.93	0.85	0.89	0.7	0.74
S5	0.89	0.89	0.81	0.81	0.81	0.7
S6	0.89	0.89	0.63	0.63	0.55	0.67
S7	0.89	0.96	0.78	0.74	0.63	0.67
Mean	0.94	0.931	0.77	0.77	0.68	0.7

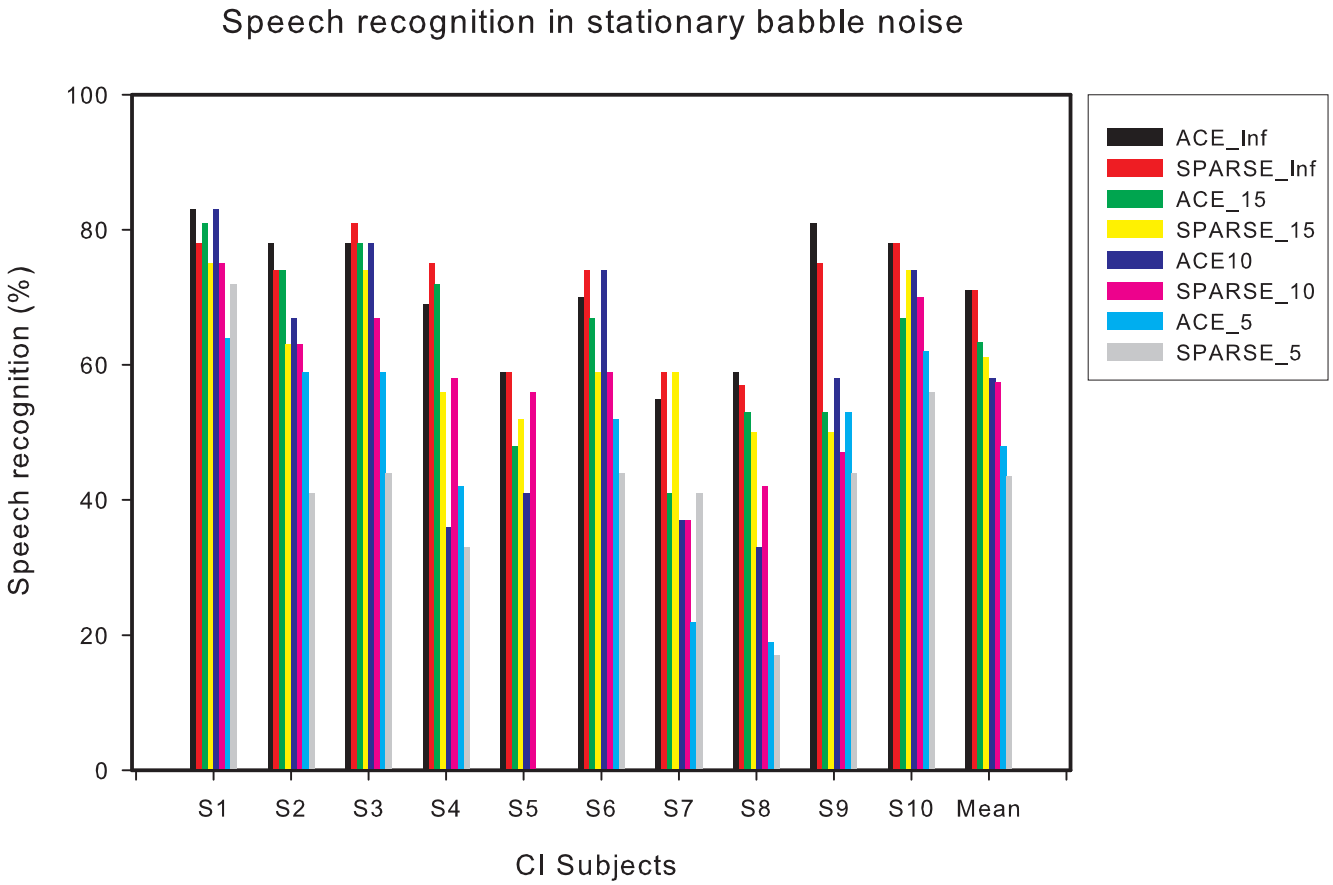


Figure 9.2: *CI users recognition score in babble noise. The label uses the format of 'ALgorithm_SNR'. i.e. 'ACE_Inf' represents using ACE algorithm in quiet; 'SPARSE_15' represent the condition of using SPARSE algorithm in 15 dB SNR.*

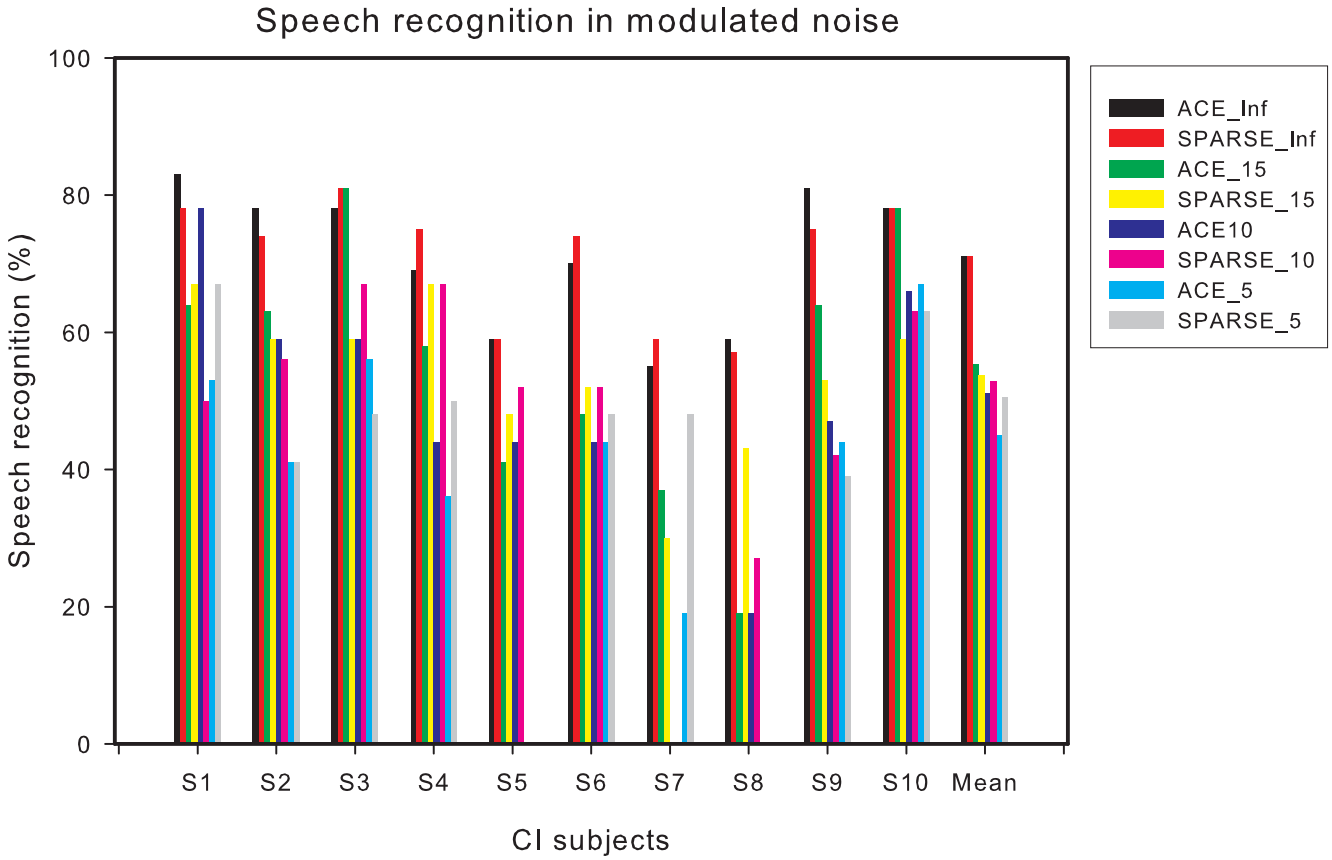


Figure 9.3: *CI users recognition score in modulated babble noise. The label uses the format of 'ALgorithm_SNR'. i.e. 'ACE_Inf' represents using ACE algorithm in quiet; 'SPARSE_15' represent the condition of using SPARSE algorithm in 15 dB SNR.*

Table 9.4: CI users listening VCV words in quiet.

subjects	ACE_Inf	$SPARSE\_Inf$
S1	0.83	0.78
S2	0.78	0.74
S3	0.78	0.81
S4	0.69	0.75
S5	0.59	0.59
S6	0.7	0.74
S7	0.55	0.59
S8	0.59	0.57
S9	0.81	0.75
S10	0.78	0.78
Mean	0.71	0.71

Table 9.5: CI users in babble noise. 'SP' is the abbreviation for SPARSE. The number after each algorithm (SP and ACE) is the SNR in dB.

subjects	ACE_15	SP_15	ACE_10	SP_10	ACE_5	SP_5
S1	0.81	0.75	0.83	0.75	0.64	0.72
S2	0.74	0.63	0.67	0.63	0.59	0.41
S3	0.78	0.74	0.78	0.67	0.59	0.44
S4	0.72	0.56	0.36	0.58	0.42	0.33
S5	0.48	0.52	0.41	0.56		
S6	0.67	0.59	0.74	0.59	0.52	0.44
S7	0.41	0.59	0.37	0.37	0.22	0.41
S8	0.53	0.5	0.33	0.42	0.19	0.17
S9	0.53	0.5	0.58	0.47	0.53	0.44
S10	0.67	0.74	0.74	0.7	0.62	0.56
Mean	0.63	0.61	0.58	0.57	0.48	0.44

Table 9.6: CI users in modulated babble noise. 'SP' is the abbreviation for SPARSE. The number behind each algorithm (SP and ACE) is the SNR in dB.

subjects	ACE_15	SP_15	ACE_10	SP_10	ACE_5	SP_5
S1	0.64	0.67	0.78	0.5	0.53	0.67
S2	0.63	0.59	0.59	0.56	0.41	0.41
S3	0.81	0.59	0.59	0.67	0.56	0.48
S4	0.58	0.67	0.44	0.67	0.36	0.5
S5	0.41	0.48	0.44	0.52		
S6	0.48	0.52	0.44	0.52	0.44	0.48
S7	0.37	0.3			0.19	0.48
S8	0.19	0.43	0.19	0.27		
S9	0.64	0.53	0.47	0.42	0.44	0.39
S10	0.78	0.59	0.66	0.63	0.67	0.63
Mean	0.55	0.54	0.51	0.53	0.45	0.51

Chapter 10

Appendix B: NIC streaming

The streaming here refers to sending the electrical sequences, which are saved in a computer to CI users. NIC (Nucleus Implant Communicator) refers to a set of software and hardware, developed by Cochlear company to support the streaming process. The hardware mainly includes two parts. One is the programming pod and the other is L34 research speech processor. The NIC software can generate specific electrical stimulus sequences or send commands to implement streaming.

In our experiment, we used NIC to send the sequences of SPARSE and ACE to CI users. As shown in the Fig. 10.1, sequences of VCV words can be generated with ACE or SPARSE algorithm. The sequences for each individual are based on not only the strategy but also their owm mapping files, defining many important parameters for streaming such as the most comfortable level (C level), the threshold (T level) and number of active electrodes and so on.

An example of streaming commands in MATLAB :

% prepare streaming
client = NICstreamClient;
% 'l34-cic3-0'is for virtual device. 'l34-cic3-1' is for real stimuli. Selecting different type of processor (by CIC3 for CI24M, or CIC4 for freedom)
client = initialiseClient(client, 'l34-cic3-1');
% Send the sequence to be streamed.
client = sendData(client, data.sequence);

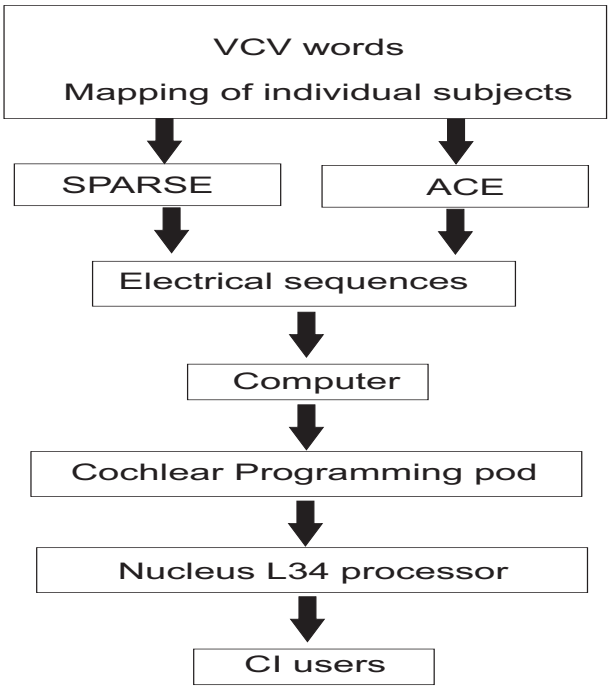


Figure 10.1: *NIC streaming process. The electrical sequences can be generated based on SPARSE or ACE algorithms and CI users' mapping file, which defines the active electrodes, most comfortable level and many other critical parameters for streaming. The sequences can then be saved as files in a computer. The streaming command can then load the sequences and send the sequences from computer to CI subjects internal receiver via the L34 research processor and the programming pod.*

%Start the streaming
client = startStream(client);
%Wait until the streaming has finished. STATUS= StreamStatus;
[*client*, *status*]= streamStatus(client);
while (STATUS.idle = status)
pause(0.5);
[*client*, *status*] = streamStatus(client);

```
end
% Stop the system
client = stopStream(client);
```

References

Allen, J. (1994). How do humans process and recognize speech? *IEEE Trans Speech Audio Proc*, **2**, 567–577. 14

Arons, B. (1992). A review of research in the area of multi-channel and spatial listening with an emphasis on techniques that could be used in speech-based systems. *Journal of the American Voice I/O Society*, **12**, 35–50. 4

Assmann, P.F. (1995). The role of formant transitions in the perception of concurrent vowels. *J. Acoust. Soc. Am*, **97**, 575–584. 35, 91

Atick, J.J. (1992). Could information theory provides an ecological theory of sensory processing? *Network: Computation in Neural Systems*, **3**, 213–251. 17

Attneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev.*, **61**, 183–193. 9, 10, 11, 17

Bacon, S., Fay, R.R. & Popper, A.N. (2004). *Compression: From Cochlea to Cochlear Implants*. Springer. 85, 87, 93

Barlow, H. (1959). Sensory mechanisms, the reduction of redundancy and intelligence. In *National Physical Laboratory symposium No. 10*. 10, 17, 18

Barlow, H. (1989). Unspervised learning. *Neural Computation*, **1**, 295–311. 17

Barlow, H. (2001). Redundancy reduction revisited. *Network*, **12**, 241–53. 2, 9, 10, 11, 17, 18, 42, 94, 126

Barlow, H.B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*. 17, 46, 54

Barlow, H.B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? *Perception*, **1**, 371–394. 16, 17, 18

Bekesy, G.v. (1948). On the elasticity of the cochlear partition. *J. Acoust. Soc. Am*, **20**, 227–241. 35

Bell, A.J. & Sejnowski, T.J. (1996). Learning the higher-order structure of a natural sound. *Network-Computation in Neural Systems*, **7**, 261–267. 19, 66

Bell, A.J. & Sejnowski, T.J. (1997). The independent components' of natural scenes are edge filters. *Vision Res*, **37**, 3327–3338. 19, 20, 66

Bentler, R. & Chiou (2006). Digital noise reduction: an overview. *Trends in amplification*, **10**, 71–82. 67

Bregman, A.S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press. 4, 6, 37

Bronkhorst, A.W. & Plomp, R. (1992). Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *J. Acoust. Soc. Am*, **92**, 3132–3139. 67

Brown, G.J. & Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech and Language*, **8**, 297–336. 6

Bruce, V., Green, P.R. & Georgeson, M. (1996). *Visual perception: Physiology, psychology and ecology*. Psychology Press Ltd. 7

Cherry, E. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am*, **25**, 975–979. 4

Clark, G. (2003). *Cochlear Implants: Fundamentals and Applications*. Springer. 91

Coker, C. & Umeda, N. (1974). Speech as an error correcting process. In *Speech communication seminar, SCS-74, Aug. 1-3*, 349–364. 13

Cooke, M. (2003). Glimpsing speech. *Journal of Phonetics*, **31**, 579–584. 15

Cooke, M. (2006). A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*, **119**, 1562–1573. 1, 15, 25, 29, 114

Cooke, M. & Ellis, D.P.W. (1998). The auditory organization of speech in listeners and machines. Tech. rep., International Computer Science institute. 6

Cooke, M. & Okuno, H. (1999). Introduction to the special issue on computational auditory scene analysis. *Speech Communication*, **27**, 155–157. 37

Cooke, M.P. & Brown, G.J. (1993). Computational auditory scene analysis - exploiting principles of perceived continuity. *Speech Communication*, **13**, 391–399. 6, 37

Dahlquist, M., Lutman, M., Wood, S. & Leijon, A. (2005). Methodology for quantifying perceptual effects from noise suppression systems. *International Journal of Audiology*, **44**, 721–732. 67

Dean, I., Harper, N.S. & Mcalpine, D. (2005). Neural population coding of sound level adapts to stimulus statistics. *Nature Neuroscience*, **8**, 1684–1689. 127

Diehl, R.L., Lotto, A.J. & Holt, L.L. (2003). Speech perception. *Annu Rev Neurosci*, **55**, 149–179. 37, 85

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *J. Acoust. Soc. Am*, **97**, 585–592. 14

Eddington, D.K. & Dobelle, W.H. (1978). Auditory prothesis research with multiple hannel intracochlear stimulation in man. *Ann Otol Rhinol Laryngol suppl*, **87**, 1–39. 91

Field, D.J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J Opt Soc Am A*, **4**, 2379–2394. 19, 126

Field, D.J. (1994). What is the goal of sensory coding? *Neural Comput.*, **6**, 559–601. 2, 18, 20, 21, 44, 48, 52, 71, 72, 114

Fletcher, H. (1953). *Speech and hearing in communication*. New York: Van Norstand. 14

Fletcher, H. & Galt, R. (1950). The perception of speech and its relation to telephony. *J. Acoust. Soc. Am*, **22**, 89–151. 14

Fowler, C. (1996). Listeners do hear sounds, not tongues. *J. Acoust. Soc. Am*, **99**, 1730–41. 47

Fowler, C.A. (1986). An event approach to the study of speech perception from a directrealist perspective. *J. Phon.*, **14**, 3–28. 47

Friesen, L.M., Shannon, R.V., Baskent, D. & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *J. Acoust. Soc. Am*, **110**, 1150–1163. 83

Gardner, B. & Martin, K. (1994). Hrtf measurements of a kemar dummy-head microphone. 61

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S. & Dahlgren., N.L. (1993). Darpa timit acoustic phonetic continuous speech corpus cd-rom. NIST, Md. 27

Greenberg, S., Ainsworth, W.A., Popper, A.N. & Richard., R.F. (2004). *Speech processing in the auditory system*. Springer. 14, 73, 74, 85, 91, 93

Greenwood, D. (1990). A cochlear frequency-position function for several species - 29 years later. *J. Acoust. Soc. Am.*, **87**, 2592–2605. 90

Hartmann, R., Topp, G. & Klinke, R. (1984). Discharge patterns of cat primary auditory fibers with electrical stimulation of the cochlea. *Hear Res*, **13**, 47–62. 23, 67

Helmholtz, H. (1925). *Physiological Optics. Volume III. The Theory of the Perceptions of Vision*. Washington, DC: Optical Society of America. 22, 89

Helmholtz, H.V. (1863). Die lehre von den tonempfindungen [on the sensations of tone]. *Braunschweig: F. Vieweg und Sohn.*. 3

Hoyer, P.O. & Hyvarinen, A. (2002). A multi-layer sparse coding network learns contour coding from natural images. *Vision Res*, **42**, 1593–1605. 12

Hyvarinen, A. (1999). Sparse code shrinkage: denoising of nongaussian data by maximum likelihood estimation. *Neural Comput*, **11**, 1739–1768. 55, 97, 109

Hyvarinen, A. & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Netw*, **13**, 411–430. 48, 56, 102

Hyvarinen, A., Hoyer, P. & Oja, E. (1998). Sparse code shrinkage for image denoising. vol. 2, 859–864 vol.2. 113

Hyvarinen, A., Gutmann, M. & Hoyer, P.O. (2005). Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in V2. *BMC Neurosci*, **6**, 12. 2

Hyvarinen, J., A. Karhunen & Oja, E. (2001). *Independent component analysis*. 23, 52, 94, 96, 99

James, C.J. & Hesse, C.W. (2005). Independent component analysis for biomedical signals. *Physiological Measurement*, **26**, R15–R39. 56

Kasturi, K., Loizou, P.C., Dorman, M. & Spahr, T. (2002). The intelligibility of speech with "holes" in the spectrum. *J.Acous.Soc.Am.*, **112**, 1102–1111. 14

Kiang, N.Y. & Moxon, E.C. (1972). Physiological considerations in artificial stimulation of the inner ear. *Ann Otol Rhinol Laryngol*, **81**, 714–730. 23, 67

Kvale, M.N. & Schreiner, C.E. (2004). Short-term adaptation of auditory receptive fields to dynamic stimuli. *J Neurophysiol*, **91**, 604–612. 66

Laming, D. (1997). *The measurment of sensation*. Oxford university press. 43

LeBlanc, J. & Len, P.D. (1998). Speech separation by kurtosis maximization. In *IEEE Int. Conf. Acoustics, Speech, Signal Processing*. 25

Lennie, P. (2003). The cost of cortical computation. *Curr Biol*, **13**, 493–497. 18

Lewicki, M.S. (2000). Learning optimal codes for natural images and sounds. *Proceedings of the SPIE - The International Society for Optical Engineering*, **4119**, 185. 20, 36

Lewicki, M.S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, **5**, 356–363. 20, 36, 67, 72, 126

Li, G. & Mark, E.L. (2006). Sparseness and speech perception. In *Interspeech 2006–ICSLP*. 117

Li, K., Swamy, M. & Ahmad, M. (2005). An improved voice activity detection using higher order statistics. *Speech and Audio Processing, IEEE Transactions on*, **13**, 965–974. 25

Liberman, A.M. (1996). *Speech: a special code*. MIT Press, Cambridge, MA; London. 73, 75

Liberman, D.P.C., A. M. & Cooper, F.S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, **1**, 153–167. 35

Liu, J., Zhao, C., Zou, X. & Zhang, W. (2005). An approach of speech enhancement by sparse code shrinkage. *Neural Networks and Brain, 2005. ICNN&B '05. International Conference on*, **3**, 1952–1956. 23

Loizou, P.C., Dorman, M.F., Tu, Z. & Fitzke, J. (2000). Recognition of sentences in noise by normal-hearing listeners using simulations of speak-type cochlear implant signal processors. *Ann Otol Rhinol Laryngol Suppl*, **185**, 67–68. 3

Loizou, P.C., Lobo, A. & Hu, Y. (2005). Subspace algorithms for noise reduction in cochlear implants. *J. Acoust. Soc. Am*, **118**, 2791–2793. 99

Massaro, D. (1975). *Experimental psychology and information processing*. Chicago: Rand Mc Nally. 38

Match, E. (1886). *The analysis of sensation, and the relation of the physical to the psychical*. Chicago, IL: Open Court. 22

Miller, G.A. & Licklider, J.C.R. (1950). The intelligibility of interrupted speech. *J. Acoust. Soc. Am*, **22**, 167–173. 14

Moore, B.C. (2003a). *An Introduction to the Psychology of Hearing, Fifth Edition*. Academic Press. 3, 89

Moore, B.C. (2003b). Speech processing for the hearing-impaired: successes, failures, and implications for speech mechanisms. *Speech Communication*, **41**, 81–91. 3

Nemer, E. & Goubran, S., R.and Mahmoud (2001). Robust voice activity detection using higher-order statistics in the lpc residual domain. *Speech and Audio Processing, IEEE Transactions on*, **9**, 217–231. 25

Nemer, E., Goubran, R. & Mahmoud, S. (1999). Snr estimation of speech signals using subbands and fourth-order statistics. *Signal Processing Letters, IEEE*, **6**, 171–174, 1070-9908. 25, 31

Nemer, R.M.S., E.; Goubran (1999). Snr estimation of speech signals using subbands and fourth-order statistics. *Signal Processing Letters, IEEE*, **6**, 171–174. 128

Olshausen, B.A. & Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**, 607–609. 20, 46, 72

Olshausen, B.A. & Field, D.J. (2004). Sparse coding of sensory inputs. *Curr Opin Neurobiol*, **14**, 481–487. 18, 19, 25, 44, 45, 67, 68, 71, 87

Olshausen, B.A. & O'Connor, K.N. (2002). A new window on sound. *Nat Neurosci*, **5**, 292–294. 18, 22, 72, 126

Pearson, K. (1892). *The Grammar of Science*. London: Scott. 22

Pisoni, D.B. (1985). Speech perception: Some new directions in research and theory. *J. Acoust. Soc. Am*, **78**, 381–388. 36

Plomp, R. (2002). *The intelligent ear: on the nature of sound and perception.* 34, 35, 39, 40, 85, 126

Potamitis, I., Fakotakis, N. & Kokkinakis, G. (2001). Speech enhancement using the sparse code shrinkage technique. *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, **1**, 621–624 vol.1. 23

Purwins, H., Blankertz, B. & Obermayer, K. (2000). Computing auditory perception. *Organized Sound*, **5**, 159–171. 8

Ricketts, T.A. & Hornsby, B.W. (2005). Sound quality measures for speech in noise through a commercial hearing aid implementing "digital noise reduction". *J. Am. Acad. Audiol.*, **16**, 270–277. 67

Rieke, F., Warland, D., Steveninck, R.R. & Bialek, W. (1999). *Spikes: Exploring the Neural Code*. The MIT Press. 16, 21, 43

Schwartz, O. & Simoncelli, E.P. (2001). Natural signal statistics and sensory gain control. *Nat Neurosci*, **4**, 819–825. 46

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423 and 623–656. 11, 41, 42

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J. & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, **270**, 303–304. 12, 13, 14, 114

Shannon, R.V., Jensvold, A., Padilla, M., Robert, M.E. & Wang, X. (1999). Consonant recordings for speech testing. *J. Acoust. Soc. Am*, **106**, L71–L74. 27

Simoncell, E.P. (2003). Vision and the statistics of the visual environment. *Curr Opin Neurobiol*, **13**, 144–149. 126

Simoncelli, E.P. & Olshausen, B.A. (2001). Natural image statistics and neural representation. *Annu Rev Neurosci*, **24**, 1193–1216. 126

Simpson, S.A. & Cooke, M. (2005). Consonant identification in n-talker babble is a nonmonotonic function of n. *J. Acoust. Soc. Am.*, **118**, 2775–2778. 15

Smaragdis, P. (2001). *Redundancy reduction for computational audition, a unifying approach*. Ph.D. thesis, Massachusetts Institute of Technology. 22, 38, 48, 52, 72

Smith, E.C. & Lewicki, M.S. (2006). Efficient auditory coding. *Nature*, **439**, 978–982. 46

SPIB (2000). Tno, soesterberg, http://spib.rice.edu/spib/data/signals/noise. 116

Stickney, G.S., Zeng, F.G., Litovsky, R. & Assmann, P. (2004). Cochlear implant speech recognition with speech maskers. *J. Acoust. Soc. Am*, **116**, 1081–1091. 83

Stone, J.V. (1993). *Independent Component Analysis: A Tutorial Introduction*. Bradford Book, London. 2, 23, 52, 54, 57, 60, 62, 66

Strange, W., Jenkins, J.J. & Johnson, T.L. (1983). Dynamic specification of coarticulated vowels. *The Journal of the Acoustical Society of America*, **74**, 695–705. 14

Warren, R.M., Riener, K.R., Bashford, J.A. & Brubaker, B.S. (1995). Spectral redundancy: intelligibility of sentences heard through narrow spectral slits. *Percept Psychophys*, **57**, 175–182. 14

Willmore, B. & Tolhurst, D.J. (2001). Characterizing the sparseness of neural codes. *Network*, **12**, 255–270. 25

Wilson, B. & Dorman, M. (2007). The surprising performance of present-day cochlear implants. *Biomedical Engineering, IEEE Transactions on*, **54**, 969–972. 82

Wilson, B.S., Finley, C.C., Lawson, D.T., Wolford, R.D., Eddington, D.K. & Rabinowitz, W.M. (1991). Better speech recognition with cochlear implants. *Nature*, **352**, 236–238. 91, 92

Yang, B. (1998). Vowel perception by formant variation. *J. Acoust. Soc. Am*, **103**, 3093+. 35