The Cocktail Party Problem: Solutions and Applications

The Cocktail Party Problem: Solutions and Applications

By

Karl Wiklund, B.Eng.Scty, M.A.Sc.

A Thesis
Submitted to the School of Graduate Studies
In Partial Fulfillment of the Requirements
for the degree of
Doctor of Philosophy

McMaster University

Doctor of Philosophy (2009): McMaster University
Department of Electrical and Computer Engineering
Hamilton, ON

TITLE: The Cocktail Party Problem: Solutions and Applications.
AUTHOR: Karl Wiklund, B.Eng.Scty, M.A.Sc.
SUPERVISOR: Dr. Simon Haykin.
NUMBER OF PAGES: xi, 149

Abstract

The human auditory system is remarkable in its ability to function in busy acoustic environments. It is able to selectively focus attention on and extract a single source of interest in the midst of competing acoustic sources, reverberation and motion. Yet this problem, which is so elementary for most human listeners has proven to be a very difficult one to solve computationally. Even more difficult has been the search for practical solutions to problems to which digital signal processing can be applied. Many applications that would benefit from a solution such as hearing aid systems, industrial noise control, or audio surveillance require that any such solution be able to operate in real time and consume only a minimal amount of computational resources.

In this thesis, a novel solution to the cocktail party problem is proposed. This solution is rooted in the field of Computational Auditory Scene Analysis, and makes use of insights regarding the processing carried out by the early human auditory system in order to effectively suppress interference. These neurobiological insights have been thus adapted in such a way as to produce a solution to the cocktail party problem that is practical from an engineering point of view. The proposed solution has been found to be robust under a wide range of realistic environmental conditions, including spatially distributed interference, as well as reverberation.

Acknowledgements

I would like to sincerely thank my supervisor, Dr. Simon Haykin, for his support and guidance during the course of this degree. I would also like to thank Dr. Suzanna Becker, Dr. Ian Bruce, and Dr. Laurel Trainor for their many helpful suggestions, without which this process would have been considerably more difficult.

I would also like to thank my parents for their patience and support, as well as my friends in Hamilton and elsewhere for their encouragement, as well as for making this entire process much easier to bear.

Table of Contents

Abstract ..... **iv**
Acknowledgements ..... **v**
Table of Contents ..... **vi**
List of Tables ..... **x**
List of Acronyms ..... **xi**

Chapter 1 The Cocktail Party Problem **1**
1.1 The Cocktail Party Problem 1
1.2 Solving the Coktail Party Problem 6
1.3 Outline of the Thesis and Contribution to the Literature 8

Chapter 2 Auditory Organization and the Cocktail Party Prob **10**
2.1 Introduction 10
2.2 The Auditory Periphery 11
2.3 Higher level Representation and Streaming 14
2.4 The Effects of Reverberation 25
2.5 Summary and Conclusions 29

Chapter 3 Computational Auditory Scene Analysis **31**
3.1 Computational Auditory Scene Analysis 31
3.2 Time-Frequency Masking 32
3.3 Computational Acoustic Cues 38
3.4 Cue Fusion and Stream Segregation 48
3.5 Summary and Conclusions 51

Chapter 4 Fuzzy Logic Cocktail Party Processor **53**
4.1 Introduction: The Cocktail Party Problem Revisited 53
4.2 Approaches to Cue Fusion and Uncertainty 54
4.3 A New Approach to Cue Fusion 60
4.4 Control 69
4.5 Post Processing Using Spectral Subtraction 74
4.6 Summary of Novelty 76
4.7 Example Results 77
4.8 Summary and Conclusions 80

Chapter 5 Coherent Independent Components Analysis **81**
5.1 The Limitations of CASA 81
5.2 Indepenent Components Analysis 82
5.3 Coherent Independent Components Analysis 86
5.4 Coherent ICA From First Principles 94
5.5 Summary and Conclusions 101

Chapter 6 Objective and Subjective Evaluation **105**
6.1 Objective Evaluation 105
6.2 Subjective Evaluation 116
6.3 Discussion of Results 124
6.3 Summary and Conclusions 126

Chapter 7 Summary and Future Work **131**
7.1 Summary 131
7.2 Future Work 132
7.3 Conclusion 137

Appendix A **138**
Appendix B **140**
References **141**

List of Figures

Figure 2.1 The frequency response of gamma-tone filters..... 13
Figure 2.2 An example of voiced speech..... 21
Figure 2.3. The speech onset.....24
Figure 2.4. A sample room impulse response.....25

Figure 3.1. Speech onset estimator..... 43
Figure 3.2. ITD spread for SNR = ∞ dB.....46
Figure 3.3 ITD spread for SNR = 0dB.....46
Figure 3.4 ITD spread for SNR = -∞ dB.....47
Figure 3.5. The formation of a binary mask using logical operations. 50

Figure 4.1. A quadrilateral membership function..... 64
Figure 4.2 The fuzzy "most" membership function.....64
Figure 4.3. A basic flowchart..... 66
Figure 4.4. A basic flowchart..... 66
Figure 4.5. With two directional microphones mounted 74
Figure 4.6. The original target signal..... 78
Figure 4.7. The observed mixture with three interfering talkers 78
Figure 4.8. The estimated signal using the original CPP 79
Figure 4.9. The target estimate using the new fuzzy CPP algorithm..... 79

Figure 5.1. Microphones mounted on KEMAR..... 83
Figure 5.2. Signals recorded from two closely spaced microphones..... 84
Figure 5.3. Pedersen's arrangement..... 85
Figure 5.4. Diagram of the operation of the cICA algorithm 87
Figure 5.5 Comparison of different Generalized Gaussian probability distributions..... 91
Figure 5.6 Sample target signal 97
Figure 5.7 Received noisy mixture. 98
Figure 5.8. The intermediate signal 98
Figure 5.9. The recovered target using FCPP-ICA..... 99
Figure 5.10. HRIR directivty patterns vs. frequency..... 100
Figure 5.11. Divergence of the cICA algorithm 101
Figure 5.12. The frequency domain cICA algorithm..... 102
Figure 5.13 Diagram of the combined system. 103

Figure 6.1. PBSE under light reverberation..... 108
Figure 6.2 PBSE under heavy reverberation.....107
Figure 6.3. Average performance in different environments..... 109
Figure 6.4 Performance of the FCPP under heavy reverberation 110
Figure 6.5. Performance of the FCPP under light reverberation 111

Figure 6.6. FCPP in different environments 111
Figure 6.7 A comparison of the FCPP and the PBSE algorithms 112
Figure 6.8. The same comparison as in Figure 6.7, but under heavy reverberation 112
Figure 6.9. ICA+FCPP speech separation 113
Figure 6.10. The user interface for the CMOS audio test 119
Figure 6.12 Group 2 average CMOS scores. 121
Figure 6.13. Group 1 results by subject 122
Figure 6.14. Group 2 results by subject 122

List of Tables

Table 3.1. Pitch Estimation Accuracy vs. SIR 41
Table 3.2. ITD Estimation Accuracy vs. SIR. 45

Table 5.1. Senisitivity of cICA to model mismatch 90

Table 6.1. Comparison of real-time algorithms 115
Table 6.2. Comparisons with non-real time methods 116
Table 6.3 CMOS Scoring 117
Table 6.4. Comparison types for subjective trials 119

List of Abbreviations

ACF	Autocorrelation Function
AI	Articulation Index
ASA	Auditory Scene Analysis
CASA	Computational Auditory Scene Analysis
CCA	Canonical Correlation Analysis
CCF	Cross-correlation Function
cICA	Coherent Independent Components Analysis
CMOS	Comparative Mean Opinion Score
CPP	Cocktail Party Problem
FCPP	Fuzzy Cocktail Party Processor
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
GUI	Graphical User Interface
HINT	Hearing in Noise Test
HRTF	Head-Related Transfer Function
ICA	Independent Components Analysis
IID	Interaural Intensity Difference
ITD	Interaural Time Delay
PBSE	Perceptual Binaural Speech Enhancement
RIR	Room Impulse Response
R-HINT-E	Realistic Hearing in Noise Test Environment
SACF	Summary Autocorrelation Function
SIR	Signal to Interference Ratio
SNR	Signal to Noise Ratio
STFT	Short-Time Fourier Transform

Chapter 1

The Cocktail Party Problem

1.1 The Cocktail Party Problem

The problem of listening to an individual sound source in a noisy environment is one that is familiar to all of us. Consider, for example, trying to carry out a conversation with another person or a group of people whilst attending a cocktail party. In order to do so effectively, you will need to be able to somehow extract the speech of the talker that you are interested in from the background noise. In this environment, neither the desired speech, nor the background noise will ever be constant, as sources will be constantly appearing, disappearing, or moving around the room. Similarly, your own attention may shift from one speaker to another as the conversation progresses, or your attention may be otherwise drawn by some other, more distant sound.

This, in a nutshell, is the essence of the cocktail party problem (CPP) as described by Colin Cherry[1]. Of specific interest in relation to this problem is the fact that the human auditory system is remarkably adept at solving it. Somehow, we are able to pay selective attention to a desired acoustic source that is present in a noisy environment, and follow it against a background of numerous competing sources. We are also able to shift

this attention both consciously and sometimes unconsciously in response to the environment. Of special importance to Cherry was the fact that, at the time, no machine had so far been constructed, or even envisioned, that was capable of solving this problem.

It is only at the present time, many decades after Cherry's observation that researchers are beginning to come to grips with the problem and propose at least partial solutions. This has stemmed, in part, from a greater understanding of the neurobiological solutions used by our own brains, and also in part by technological advances that allow us to explore possible applications that could benefit from solving the cocktail party problem.

Consider, for example, the following purely technological problems that are a matter of interest for consumers, and hence for engineers:

> 1) It is well known that in cocktail party or otherwise noisy environments, hearing impaired listeners are at a great disadvantage in deciphering speech. As a result, it would be desirable to remove as much of the unwanted interference as possible so as to improve speech intelligibility.
>
> 2) Electronic surveillance of a suspect in a noisy restaurant or similar environment could also benefit from a greater enhancement of what was being said.
>
> 3) Active noise control devices filter out significant noise sources to the exclusion of everything else. However, often other environmental sounds may be of crucial importance to the wearer; if possible, these sounds should be preserved and enhanced.
>
> 4) Telephone conversations may take place in a variety of noisy environments, from automobiles to crowded shopping malls. Clearly, it would be desirable to remove as much of the interference as possible in order to improve the clarity of the conversation.

These four examples are by no means exhaustive, nor should speech be considered the only possible signal of interest. Indeed, there are many other applications such as automatic music transcription[2], where audio segregation and streaming is likely to be of considerable benefit. It should be understood, therefore, that solutions to the cocktail

party problem are not of a purely theoretical or biological interest. A satisfactory solution may lead to an impressive array of applications in the real world.

However, in spite of such a motivation, it has only been in recent years that researchers have started to make headway into developing machines that are capable of coping with the cocktail party environment, and none of these approaches has either the capabilities or the versatility of the human auditory system[3][4]. To understand why this is so, it is essential to understand the nature of the acoustic problem and how it relates to signal processing.

Note that while the goal of "extracting a single acoustic source from a noisy background of competing sources" is simple to state, it is less easy to define the actual nature of the problem. It is not immediately clear, for example, what is signal or what is noise, nor how they should be distinguished from one another. Even given such a description, neither the desired source nor the interferers are likely to remain constant in terms of their statistical specification. After all, speech and music are both known to be highly non-stationary. Speech-on-speech interference also presents the added difficulties of both overlapping spectral regions, as well as very similar long-term statistical descriptions for both the target and the interference signals[5][6][56].

In addition, many scenarios of interest are likely to take place indoors, or within some kind of enclosure (such as an automobile). In this case, there is the added complication of reverberation. The effect of reverberation is to produce a received signal that is a combination of both the target sound as well as numerous scaled and delayed versions of itself. This results in the target signal being smeared in time, with both fine-scale and large-scale structures being distorted[7]. In humans, this can reduce the

perceived intelligibility of speech, although we possess neurobiological mechanisms that are surprisingly good at eliminating the effects of reverberation[8]. As this mechanism is still only poorly understood, there is as of yet no satisfactory computational analogue, and thus conventional signal processing algorithms are vulnerable to the distorting effects of reverberation.

As a result of these difficulties, researchers have been mostly stymied in their efforts to solve the cocktail party problem using the familiar toolsets of digital signal processing. While such algorithms have, to a greater or lesser extent, been proven useful in many different types of applications, their use in a problem as broadly defined and as variable as the cocktail party problem has been found to be highly problematic. The types of processing that have been applied to the problem of speech enhancement and noise reduction may be broadly described as falling into one of three groups:

1)**Wiener Filtering/Spectral Subtraction**

The basic form of the Wiener filter is expressed by:

$$H_s(\omega) = \frac{S_x(\omega)}{S_x(\omega) + S_n(\omega)} \tag{1.1}$$

In the above equation, $S_x(\omega)$ is taken to be the power spectral density of the target signal, $S_n(\omega)$ is the noise spectral density, and the resulting gain for a given frequency ω is $H_s(\omega)$. In the case of many conventional speech enhancement algorithms, the spectral densities of (1.1) can be replaced by their frame-based estimates[5], allowing the algorithm to track at least some non-stationarities in the target and noise signals. Typically, this is accomplished through some sort of recursive estimation that exploits

speech pauses in order to estimate the noise spectrum. However, owing to the fact that in the cocktail party environment there is no clear way of distinguishing the signal from the noise, and the level of non-stationarity is so great, this method generally performs poorly in such environments.

2)**Spatial Filtering**

Spatial filtering techniques make use of an array of microphones configured in some pattern that is determined by the needs of the application. The purpose is to use the array to focus on a signal coming from a known direction, whilst simultaneously filtering out strong signals coming from other directions. Adaptive techniques can be used where either the target or interference directions are imperfectly known or apt to change. However, for an array of N microphones it is generally possible to only eliminate at most N-1 interferers. As a result, both the number and size of the required microphone array can place practical limits on its usefulness, especially in applications where any speech enhancement device is expected to be wearable. In addition, the practical effectiveness of the spatial filtering is generally reduced in environments where there are many interfering sources, or where reverberation is present in the environment[11].

3)**Independent Components Analysis**

Independent Components Analysis (ICA) posits an array of microphones, each of which receives a signal that is a linear superposition of independent sources. Using Bell and Sejnowski's information-theoretic criterion, it is possible to recover the original sources if the number of signals N is less than or equal to the number of

sensors[12][13]. The original formulation of ICA assumed constant linear mixing matrix and the absence of noise, with the mixing model being expressed as:

$$\mathbf{y}(t) = \mathbf{A}\mathbf{x}(t) \tag{1.2}$$

where the vector of source signals is $\mathbf{x}(t)$, the mixing matrix is $\mathbf{A}$, and the vector of received signals is $\mathbf{y}(t)$. In reality, such a scenario rarely exists, and while a number of methods[14][15][21] have been devised to deal with more practical scenarios that involve background noise, convolutive mixing (i.e. reverberation) and time-varying mixing scenarios, the results have been largely inadequate for the cocktail party environment. Of additional importance is the slow convergence speed of these algorithms, which entirely rules out their use in real-time applications.

1.2 Solving the Cocktail Party Problem

Each of the algorithms outlined in the preceding section have their strengths and weaknesses. However, the fact that they have proven to be inadequate to solving the cocktail party problem indicates that a significant shift in thinking is required. Instead, a more fruitful line of inquiry may begin by asking how it is that the human auditory system is so adept at solving this problem in spite of the difficulties outlined previously. Indeed, this question has been the focus of much past research in the fields of neurobiology, psychology, computer science and engineering. As a result, it has become apparent that a clear understanding of the neurobiological and psychological foundations of human auditory perception is absolutely necessary if this signal processing problem is to be solved effectively.

The psychoacoustic process of separating received auditory signals into source-specific streams was described in detail by Bregman, who termed the phenomenon

'Auditory Scene Analysis' (ASA)[16]. Bregman summarized much of the previous research on the subject and outlined many of the acoustic cues used by the human auditory system to accomplish this task. Bregman's book did not dwell extensively on the neurobiological mechanisms involved, and this was at least partly due to a lack of knowledge available at the time.

From an engineering point of view, it is essential to understand that such knowledge is primarily useful as a guide. It is neither necessary nor desirable to merely replicate in silicon or software what nature does with neurons. The human nervous system is structured so differently from a conventional computer[8] that any slavishly imitative system is unlikely to be of anything other than theoretical interest. As a result, an understanding of the precise computational mechanisms of human auditory perception can be laid aside in favour of an understanding of what it is that is being perceived. If it is known what information is being extracted by the brain, then computational analogues may be devised that are suitable for conventional application platforms.

This computational approach to ASA has naturally been termed "Computational Auditory Scene Analysis, or CASA[2]. While this has been an active area of research for some time, few realistic applications have so far emerged. This has so far largely been a result of failing to adequately consider all of the challenges posed by the environment, as well as the constraints posed by the applications themselves. For example, in [17] the authors propose a system that requires repeated iterations over a long speech sample. In another paper[18], where the problem of reverberation is explicitly addressed, it is assumed that a fixed dereverberation filter has been learned ahead of time. In other systems, reasonable performance can be obtained for some environments, but not for

others. In particular, the level of signal distortion can rise dramatically with increasing levels of background noise or reverberation, offsetting any gains in intelligibility.

Of particular importance as well is the fact that very few authors have considered the problems associated with real-time operation. Many approaches assume that the entire speech signal is available prior to processing, and can be iterated upon if necessary. While this is possible in some applications, it is not universally so. Any kind of wearable system meant for speech/hearing enhancement, for example, is required to function in real-time.

1.3 Outline of the Thesis and Contribution to the Literature

This thesis has emerged out of an attempt to apply the concepts of CASA to realistic applications using small, portable platforms designed to be wearable or to otherwise fit within some small device. These kinds of applications cannot rely on even the computational resources of a standard desk-top computer, and in addition must function in real-time. In such cases, the possible solutions are much more severely constrained. To summarize the behaviour of such a system, we can say that in addition to actually improving speech intelligibility in noise, the system must also meet the following requirements:

1)The device requires limited physical or computational resources. Even in the most generous designs, there are still far fewer resources than are available on a conventional PC.

2)The device operates in real time. Significant processing delays are not tolerable in a hearing aid device. In particular, the time from when a sample is read into the processor and when it is fed out again should not exceed 15 ms.

3)The device outputs cannot be significantly distorted. Processing artefacts such as "musical noise" must be largely eliminated in order for the processed speech to sound sufficiently natural.

4)The device must be able to operate in a wide variety of acoustic environments, and do so with essentially no previous training.

5)Owing to the time-varying nature of the problem, any environmental adaptation must be extremely quick.

This thesis presents a novel approach to implementing CASA concepts that is effective, computationally efficient and firmly rooted in an understanding of realistic acoustic environments.

The principal contributions of this work include the following:

1) A hierarchical grouping of acoustic cues based on robustness The identity of the segments that have been grouped are then constrained, based on the average behaviour of the less reliable cues.

2) A novel approach to decision and data fusion rules that is rooted in fuzzy logic. This has allowed for a system that is substantially more robust to reverberation, as well as a reduction in musical noise.

3) A novel approach to mitigating front-back confusion, using coherent independent component analysis, is proposed.

4) A new SNR adaptive control mechanism was introduced in order to improve the perceptual performance in especially difficult environments.

The outline of this thesis is as follows:

Chapter 2 provides an introduction to auditory scene analysis. This section outlines the work done by Bregman and describes the fundamental concepts of auditory grouping and streaming.

Chapter 3 outlines the computational auditory scene analysis, as well as the computation of the simple auditory cues.

Chapter 4 presents the core of the thesis. This chapter describes the novel fuzzy cocktail party processor (FCPP), and also presents new adaptive routines designed to improve the quality and robustness of the FCPP.

Chapter 5 presents the second major novel contribution to the field. A new spatial processing system based on coherent independent components analysis is presented. This system can be implemented as an optional pre-processor to the system described in Chapter 4.

Chapter 6 brings together the results of Chapters 4 and 5 and discusses the conclusions that can be drawn. Discussion of possible improvements and future research directions is provided.

Chapter 2

Auditory Organization and the Cocktail Party Problem

2.1 Introduction

If the guiding principle of our attempt to solve the cocktail party problem is to mimic the parts of the human auditory system responsible for our own ability to follow speech in noisy and reverberant environments, then an understanding of the neurobiological process is an essential foundation. Fortunately, it is a process that can be broken down into multiple, layered systems that can be described with varying levels of abstraction. This is especially important from an engineering perspective, as our goal is to develop signal processing analogues to neural processes. As a consequence, we are less interested in the precise neural mechanisms than we are in what kind of computations the neurons are engaged in; this is especially true of higher-level grouping mechanisms. The purpose of this chapter then is to examine the underlying acoustic and neurobiological mechanisms that can be used to guide the development of practical engineering solutions.

2.2 The Auditory Periphery

The perception of sound begins in the auditory periphery, which acts as a complex system for transducing acoustic signals into neural impulses. As a whole, the auditory periphery encompasses the outer, middle and inner ears, whose function is briefly described below. A more detailed description of these systems can be found in [19][20].

The Outer Ear and the Middle Ear

The outer ear consists of the external part of the ear (the pinna), the ear canal, and the eardrum (the tympanum). While not part of the transduction system *per se*, the pinnae nevertheless play an important role in spatial hearing. Besides serving to focus the auditory system on sounds in front of the listener, they play a further role in localizing sounds via spectral shaping. The spectrum of high frequency sounds is altered in a direction dependent fashion by the pinnae, which provides acoustic cues for localization regarding elevation, as well helping to distinguish sounds originating from in front of or from behind the listener[9][10]. This fact will be discussed in later chapters as it has important implications for signal processing in the cocktail party environment.

From the pinnae, sound travels down the short length of the ear canal and impinges on the eardrum, causing it to vibrate. These vibrations are transmitted to the fluid-filled cochlea via a trio of small bones called ossicles.

The Cochlea

While the inner ear consists of a number of different parts that perform a variety of functions, it is in the cochlea that the transduction of acoustic signals to neural

impulses actually occurs. The cochlea itself is a snail shaped tube that is filled with an extra-cellular fluid which is used to transmit acoustic impulses down the length of the cochlea. The interior of the cochlea is divided lengthwise by the basilar membrane, on which reside the inner and outer hair cells responsible for converting the physical vibrations of sound into neural spike trains[22].

The basilar membrane varies in both mass and stiffness across its length, and so responds in complex ways to acoustic vibrations in the cochlear fluid. Specifically, it can be likened to a mechanical filter, where certain regions of the membrane respond preferentially to different vibrational frequencies. The base of the cochlea thus responds most to high frequencies, while towards the apex, the response favours lower frequencies[19]. As the hair cells are stimulated by the motion of the basilar membrane, a tonotopic or place-frequency map is created, where certain groups of neurons respond only to certain ranges of frequencies. The resulting time-frequency decomposition is often referred to as a tonotopic map or a cochleagram[2].

The frequency decomposition obtained from the cochleagram resembles the output of a constant-Q filterbank[5] in that the low-frequency filters possess a narrower bandwidth than the higher-frequency filters. In general, the actual auditory filter banks are typically modelled using gamma-tone filters, which are known to provide a close approximation to the true auditory filters. Interestingly, this arrangement seems to be optimal from an information-theoretic point of view as demonstrated by Lewicki[23], a fact which has encouraged its use in application-oriented cochlear models as well. At least one researcher has also found that speech processing algorithms based on

neurobiological systems perform better using this decomposition as opposed to using a uniform frequency decomposition such as the well-known FFT[24].

It should be noted that the output of the auditory nerves is approximately a half-wave rectified version of the input signal. As this representation is also convenient for many signal processing algorithms, it is also found in many simplified computational models as well.

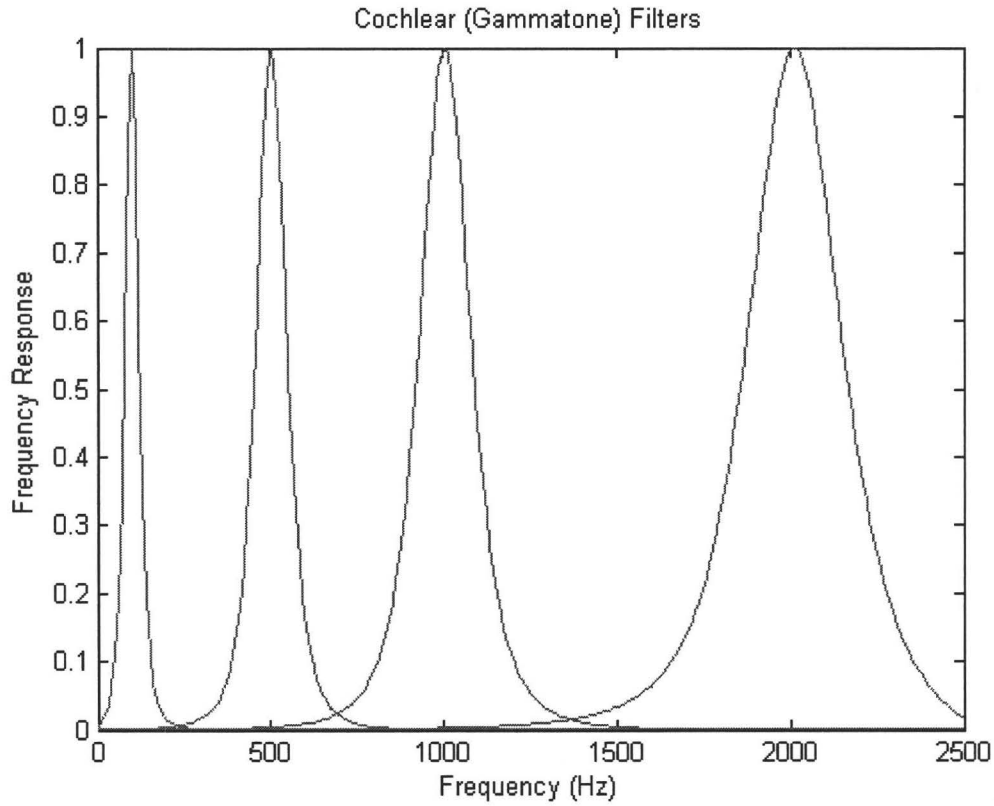

Figure 2.1. An example of the frequency response of four different gamma-tone filters. These filters closely approximate the frequency response of the mechanical filters found in the cochlea.

The cochleagram provides the basic starting point of all attempts to unravel the riddle of the cocktail party problem. While there are additional layers of neural representation beyond the basic time-frequency map provided by the cochleagram, these layers should be considered as being derived from the basic representation described above. More

importantly, the form of processing performed can be readily described in terms of features within the cochleagram, and without significant reference to the underlying physiology.

Within the cochleagram itself, as well as in artificial time-frequency decompositions, two critical facts stand out as motivations for our understanding of higher-level grouping and segregation in noisy environments. The first is that the human speech signal is sparsely distributed in the time-frequency plane[26][27]. This means that within a received mixture, there will be regions where there is little overlap between the target and the interference signals. This can be true even in very noisy environments[25], which allows for the possibility of the auditory system focusing on and tracking just those regions where the desired signal is strongest.

The other salient fact is that the speech signal contains a significant degree of redundancy. Even in situations where portions of the actual speech signal are obscured by interference, there is still often sufficient information present for the human auditory system to make sense of what is being said. The reasons for this are complex, and will be discussed later in this chapter, but the upshot for engineers is that a certain amount of signal fidelity can be sacrificed while still retaining speech intelligibility or recognizability. How much can be sacrificed, of course, will depend on the specific application being considered.

2.3 Higher level Representation and Streaming

Beyond the primitive structures of sound extracted by the cochlea, the auditory system must at some point begin to use such representations in order to build a coherent

picture of the listener's acoustic environment. This is not a straightforward task; using only data from two one-dimensional time series, some kind of picture must emerge of events happening in a three-dimensional world. The difficult task is to somehow combine the available sensory evidence in the right way, so that the resulting conclusions about the scene are more or less correct.

This problem is not altogether different from what is known about brain's analysis of visual scenes. In both cases, the individual neurons, or receptive fields[28], observe only a small part of the scene, and even then often only a particular type of feature. The shapes we perceive at a higher level are built up from those individual perceptual elements. Some kind of rules (likely soft) must exist in the brain to either group visual regions together, or to separate them as distinct objects, based on some perceived similarities or dissimilarities of those regions. It is less clear, however, what constitutes a "region" for a sound. Unlike in vision, acoustic events from separate sources do not occlude each other in the conventional sense, rather, they are merged in an additive fashion where they overlap[16][25]. There is also the additional question of how the differences or similarities of acoustic events should be judged.

Analogies to Visual Processing

The problem of how the human brain makes sense out of the images it sees is one that has been extensively studied. Of particular importance was the question of how a visual scene is assembled from the inputs provided by the highly localized detectors in the eye. To an extent, this question was addressed by the Gestalt school of psychological thought around the turn of the last century[16]. It was the opinion of these researchers

that visual scenes were assembled by the forming of connections between the various regions perceivable by low-level visual receptors.

The Gestaltists' goal was to discover the rules and conditions by which visible elements would be grouped together into larger objects. The rules by which such groupings were carried could ultimately be described as resting on the principles of similarity and continuity. Individual elements that were more similar to each other than they were to other elements in a scene would tend to be grouped together, rather than be perceived as discrete, unconnected entities. Similarity, in this case, could mean colour, proximity, temporal continuity and so forth.

It was argued by Bregman[16] that the brain takes a similar approach to sound and the composition of auditory "scenes". If two sounds in a mixture, for example, change in a similar way over time, or occur in some other closely related way, then it is likely that they originated from the same source. In this way, the occurrence of individual sound sources could be isolated from each other in a process that Bregman referred to as auditory streaming. Within the brain, attention could then be shifted between the various separate streams occurring in the observed acoustic mixture.

Auditory Streaming

Given that our primary experience of sound is through its changes in time, it should be clear that the level of temporal similarity in events is one of the ways in which sound streams are grouped or segregated. As a result, several ways in which the human auditory system makes use of temporal information can be described. However, in addition to time, acoustic events also often occur across frequencies. Therefore, a second

mode of description can be discussed that acts on temporally simultaneous events across frequency bands.

In his book, Bregman distinguished between the two modes by focusing on largely artificial sound experiments. These help to understand the basic mechanisms involved, but it must be understood that real auditory streaming takes place in a much more complicated environment, where the two modes necessarily interact to a high degree. As a result, a full discussion of the related experiments is outside the scope of this thesis. Instead, what follows is meant to be a brief discussion of the grouping mechanisms that Bregman and others have described.

Exclusive Allocation

Perhaps the key notion behind Bregman's theory of auditory streaming is what he terms "the principle of exclusive allocation". The basic concept of this principle is that a perceptual element should not be assigned to more than one object in the scene at a time. An element must be categorized as belonging to one object or the other; it cannot be in both. While the initial motivation for this idea is rooted in visual theory, its auditory counterpart is well-evidenced experimentally[29][30]. This principle is also consistent with Gestaltist ideas that different explanations compete with each other for the assignment of perceptual elements.

However, the principle of exclusive allocation is not an absolute rule. There are instances where it can be violated[31][32]. Why this should be so is understandable given that sounds do not overlap or occlude; they mix in an additive fashion instead. As a consequence it may not be desirable for the auditory system to always exclusively

assign one time-frequency element to either one source or another[16]. The observed examples of the violation of this principle always occur when two sounds occur at the same time, and where the cue information is ambiguous.

Spatial Cues

That we have two ears is a strong indicator of the role of spatial perception in audition. Certainly, it is well known that in noisy environments, the ability to hear with both ears improves the perceived speech intelligibility on the part of the listeners[1]. One interpretation is that listening with both ears allows the brain to choose which ear receives the least noisy signal and follow that. However, from an auditory scene analysis point of view, a more sophisticated form of processing may also be employed.

The idea of stream segregation, in which acoustic events are grouped according to some rule of similarity, provides another approach to understanding the value of binaural processing. It would seem likely, for example, that sounds emanating from the same spatial location at the same time would have originated from the same source. Similarly, simultaneous sounds emanating from different directions should be regarded as coming from different sources, and so not grouped together perceptually.

The primary sources of spatial information used by the auditory system arise from the interaction of the acoustic waveform with the listener's head and the binaural nature of hearing. If a sound source is not directly in front of or behind a listener, but instead impinges on the ears from some other angle, the sound will arrive at the closer ear somewhat earlier and with a somewhat greater intensity. The difference in time between when a sound arrives at one ear as opposed to the other is referred to as the **interaural**

time difference (ITD), and the difference in received intensity is called the **interaural intensity difference** (IID).

These two cues are generally dominant in two different frequency ranges. The interaural time difference is the primary cue below about 1.5 kHz, while its importance is diminished above this point owing to phase ambiguities. Certain types of sounds provide an exception to this rule however. An interaural delay can still be observed from amplitude modulated signals, provided the modulation is slow enough, as well as for non-periodic signals such as impulses. In both cases, the relevant information is extracted from the signal envelope and not from the carrier waveform itself[33].

Above 1.5 kHz, the interaural intensity difference is dominant. Below this range, the acoustic wavelengths are too long for the head shadowing effects that give rise to the IID to be of much consequence.

It turns out that this information is sufficient for reasonably accurate angle-of-arrival estimation by human subjects[34][35]. There are several vitally important ambiguities that must be addressed however. It is apparent that a signal emanating from behind the listener will produce the same ITD and IID as a signal in front of the listener. In addition the IID and ITD cues do not take the elevation of a source into account. Experiments by Wightman and Kistler[36], using probe microphones inserted into the ear, document the effects of varying the azimuth and elevation of a source on the received binaural signal. Of particular importance is that both elevation and azimuth were found to produce ambiguities in the IID and ITD measurements. In general, however, the ITD was found to be more vulnerable to such effects.

Monaural Harmonic Cues

Most natural sounds consists of mixtures of a range of frequencies; pure tones, such as those used in many audio experiments, are rare. However, many such natural sources occupy only a possibly disjoint subset of the available frequency space at any given time. As a consequence, the human auditory system needs to be able to partition this frequency space into different streams. This requires that multiple, simultaneously occurring acoustic events need to be analyzed and grouped according to some set of similarity rules that are valid across frequencies. While this is true of the spatial cues discussed earlier, they are not cues that are inherent to the source sound, and in any case are not monaural in nature. Therefore, they are different enough to warrant discussion in a separate section. There are two cues that are of particular importance to frequency-based separation: pitch and common fate.

The pitch, or periodicity, of a signal is most commonly associated with speech. In the formation of voiced speech, vibrations are produced in the vocal fold, which gives rise to a pulse-like periodic element in vowel sounds. The frequency of these pulses is referred to as the pitch, or fundamental frequency of the speech. Importantly, this periodicity is common across all of the associated harmonics of a speaker at that time. Therefore, any portions of received spectral energy with the same pitch frequency should in general be grouped together as being from the same source. However, it should be noted that this rule is not complete, as sounds with very similar pitch frequencies can occur at the same time. In such cases, additional grouping rules such as those relating to temporal or spatial cues must be relied upon in order to resolve the ambiguities.

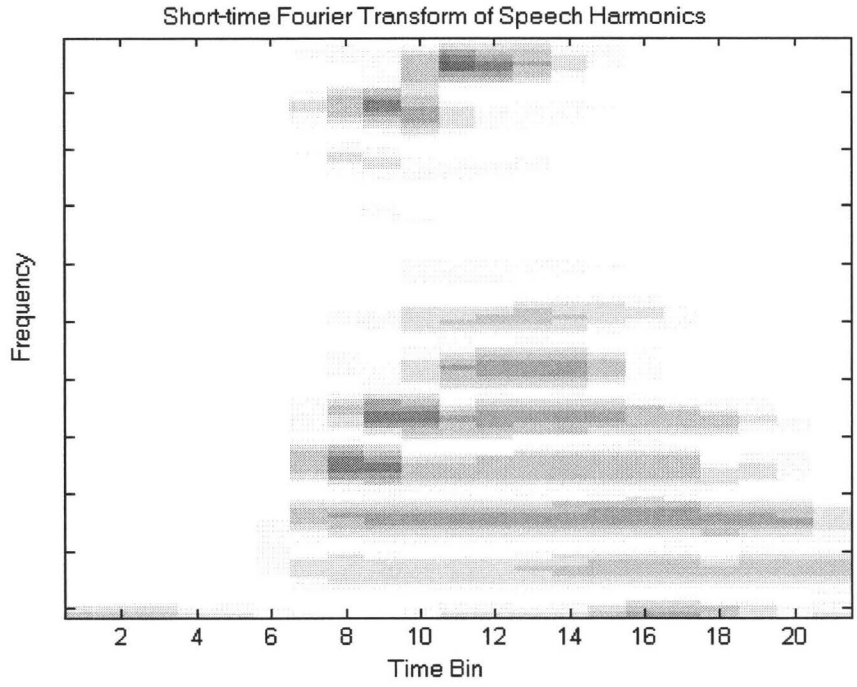

Figure 2.2 An example of voiced speech. The multiple, parallel horizontal lines reflect the presence of the harmonics in the signal.

Related to pitch is the concept of harmonic series. In voiced speech, where pitch is present, harmonics will be generated at integer multiples of the fundamental frequency. For a fundamental f_0, additional harmonics will be present at $2f_0$, $3f_0$, and so on. While this is presented in [16] in part as an alternative to periodicity analysis as a means of identifying the pitch frequency in the human auditory system, many non-speech sounds also exhibit harmonic structure. The physics of vibration dictate that similar harmonic series are generated by many vibrating structures, even if not driven by a common mechanism. As a result, the human auditory system also has the natural ability to group such series, and reject tones that do not naturally belong to it.

As a cue, harmonicity is known to be relatively strongly attended to. Experiments by Broadbent and Ladefoged[37], for example, demonstrated that this grouping

mechanism is able to override spatial cues, at least in some circumstances. In their experiment, different halves of a harmonic series were presented to each ear. Instead of hearing two separate sounds, the series was perceived as a single source emanating from a single direction.

The second major frequency-related cue used by the human auditory system is what is often termed "common fate"[16]. Unlike pitch and harmonic relationships, which assume a more or less steady sound, the idea of common-fate concerns the changes in a sound that occur over time. As in visual scenes, it can be generally stated that objects that move together are related to each other. The constituent components of a ball, for example, will all follow the same trajectory. Similarly, a group of perfectly stationary lights can be made to produce the illusion of motion by varying the individual brightness of the bulbs according to some pattern. In contrast to the visual system's tracking of the motion of objects or changes in their brightness, the auditory system integrates changes in both frequency and amplitude in order to group acoustic events together.

While frequency modulation in sound is not an uncommon phenomenon, its properties are well illustrated by temporal changes in human speech. Not only do all of the harmonics change in the same way, but they change in such a way as to preserve the harmonic relationships. If the fundamental frequency rises from $f_0(t_0)$ to $f_0(t_1)$, then its harmonics must also change correspondingly, so they will still exist at integer multiples of the new fundamental. As it would be unusual for such a group of changes to occur by chance, it is reasonable for the auditory system to group those changing structures together as part of a single source. Interestingly enough, the brain also preserves the idea of harmonically related groups. Experiments in[36] have demonstrated that if the

relevant frequency groups do not change together in such a way as to preserve the harmonic relationships, then the brain is less apt to group those elements together, even if the direction of their change is the same. Grouping by frequency modulation also extends to the concept of pitch, which will naturally vary during the course of a spoken word.

The other way in which a signal can change is through amplitude modulation, or the changes in signal intensity over time. The earliest discussions of amplitude modulation were related to pitch and periodicity analysis. It was observed that intensity changes across the basilar membrane for a given source were correlated across frequencies, a fact that was especially noticeable for voiced speech[16]. Individual parts of the basilar membrane will vibrate at their particular centre frequencies, but for anything other than a pure tone, those vibrations are likely to be modulated by some lower frequency signal (such as pitch). Determining which frequency groups have correlated amplitude modulations is one important basis by which the auditory system groups them, and results in the percept of a single source. This is of particular importance at higher frequencies, where individual harmonics often cannot be resolved due to the broader bandwidth of the cochlear filters at those frequencies[33].

Onset

The **onset** and **offset** of a sound are defined as the sound's start and stop times. Perceptually, of these two cues, the onset is the cue that is most useful, as natural sounds are more likely to have a clearly defined start time than a stopping time. In addition, as the onset occurs first, it will contain the most important information in terms of a

listener's situational awareness. The onset itself is correlated across multiple frequency bands, and this correlation is an important cue employed by the auditory system, which requires a high degree of alignment in order for grouping to occur. Synthetic vowel experiments, conducted by Darwin and Sutherland, have shown that small misalignments in the onset times of harmonic components can substantially alter a person's perception of a vowel[38][39]. However, it has also been found that the nature of the auditory system's grouping process for onsets can depend on the type of stimulus that is presented to it.

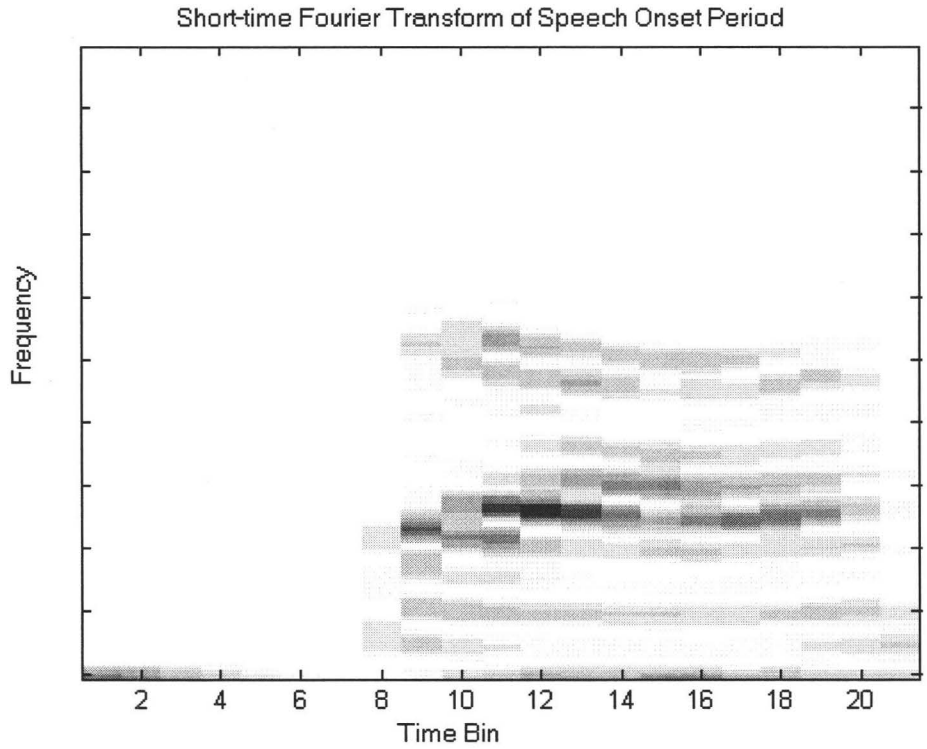

Figure 2.3. An example of a speech onset period. There is a sharp rise in signal energy that carries across several frequencies simultaneously.

2.4 The Effects of Reverberation

In most real-world environments, the sound that a listener perceives is not simply the sound that was transmitted by the source. Instead, what is received is a mixture of the direct sound plus a series of attenuated and time-delayed versions of the original source that have been reflected off of various objects and surfaces in the environment. While this effect is most pronounced in indoor environments, it also can occur in outdoor settings. Because of this fact, humans and other animals have evolved mechanisms for coping with the effects of reverberation. How the auditory system might respond to the resulting corruption of source cues is of particular interest, as reverberation is an ever present reality in many possible applications.

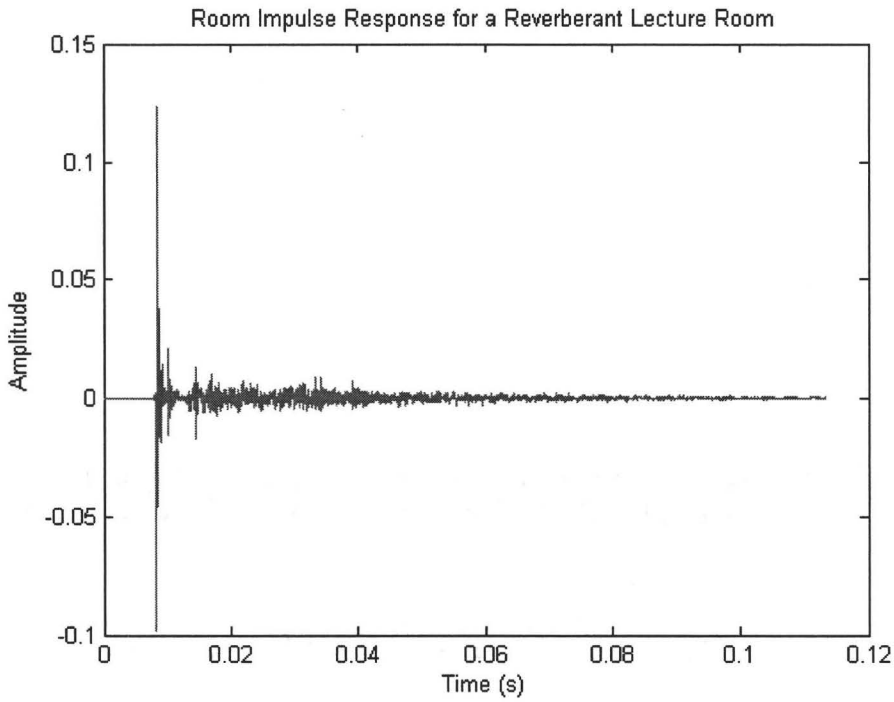

Figure 2.4. A typical room impulse response.

In an enclosed room, for example, we can imagine a single source and a single sensor. For a given placement of the source and the sensor, the effects of the room reverberation can be completely characterized by the room impulse response (RIR), which encapsulates all of the information about the strength and timing of the various received echoes over time (Figure 2.4). The nature of this RIR, and the subsequent effects of reverberation, will vary according to the source-receiver distance, the reflective strength of the walls, as well as the overall size of the room. As a result of this variability, reverberation is difficult to meaningfully characterize from a signal processing point of view except in general, statistical terms.

Effect on Spectral Cues

From an auditory perspective, despite our own ability to make sense of reverberant sounds, reverberation is still a source of confusion with regards to both auditory cues and basic speech intelligibility. Because the received mixture is a sum of both scaled and delayed versions of the original source, a signal such as speech is smeared in time so that its fine temporal structure is less distinct. This affects both the individual components of speech such as stops and frictations[40], as well as a person's overall ability to listen effectively in noise.

The way in which the speech signal is obscured by reverberation can be described as arising from two principal causes: masking effects and the loss of modulation information. Masking effects entail the obscuring of phonemic information through the overlap of echoed parts of speech with the direct sound. This causes earlier parts of a phoneme to obscure the later parts, which can blur transitional segments and offsets. In

addition, a preceding vowel may be echoed through a following consonant[7]. Because the consonant sounds are often shorter, and spoken with less energy, they are readily obscured, leading to a loss of intelligibility.

These masking effects are particularly problematic in the case of speech-on-speech interference. In such cases, both the target and the interferer are smeared into each other, resulting in a significant loss of separability. For example, in double vowel experiments, where the two interfering vowels possess the same fundamental frequency, the human auditory system is able to separate the two vowels accurately, owing to small modulations in the fundamental frequency of both signals[41]. However, in the presence of reverberation, these modulations become much less distinct, leading to a reduction in separability. Other experiments using whole speech sentences indicate that this effect carries over to running speech as well.

In addition to masking, the temporal smearing caused by reverberation also affects the coarser-level features of speech, such as amplitude modulation at the prosodic level. This form of modulation exists at the level of spoken words, and ranges between 2 and 10 Hz. This type of modulation is known to be crucial to speech intelligibility[42], although the reasons for this are poorly understood at present. While this cue is more resilient than either frequency modulation or spatial cues, reverberation still causes a loss of modulation depth, meaning that the dips in signal energy are less apparent than would be the case otherwise[43][25]. This is of crucial significance in noisy, cocktail-like environments, where it has been hypothesized that such dips in the energy of interfering signals permit clearer "glimpses" of the current target sound. Such glimpses allow the auditory system to more precisely estimate such signal attributes such as energy, pitch,

and possibly spatial position as well. Fewer or less complete glimpses due to loss of modulation depth, mean less reliability in these estimates, and hence a reduced ability for the brain to track them through the noisy periods.

Effects on Spatial Cues

Under reverberant conditions, the same sound, although in the form of slightly delayed and attenuated versions, is received from multiple directions. The spatial grouping cues normally used by the auditory system are reduced in effectiveness owing to the increased confusion regarding the actual source direction. This arises from the fact that the distribution of the ITD and IID cues are now broadened, and are no longer wholly consistent. In noisy conditions, this confusion of cues can affect a listener's ability to separate competing talkers[44].

Spatial cues therefore provide only a weak source of information regarding auditory grouping. Monauaral cues therefore seem to be dominant under such conditions, with the ITD and IID playing more or less supplementary roles. Other experiments, such as those by Broadbent and Ladefoged[37] as well by Deutsch[45] seem to confirm this, albeit under more controlled conditions. However, in both cases the conclusions indicate that spectral cues tend to dominate spatial cues whenever some conflict arises between them.

The ability to effectively localize, and hence separate signals, in reverberant environments is quite variable however. Reverberation levels, signal type[47], and sensor placement[45] all have an effect. Of particular interest, though, is the fact that signals with rapid attack times (that is, the speed at which a signal gains in amplitude) are

readily localizable, and that this localizability is independent of the level of reverberation[47]. Rapid signal onsets then can be used to localize source signals for the purpose of auditory grouping, which can also be of value for subsequent grouping and tracking via purely monaural cues. This idea was first expressed by Wallach et al[48], who found evidence that the auditory system selectively applies greater perceptual significance to onsets than it does to the following parts of the sound, which are corrupted by reverberation.

2.5 Summary and Conclusions

This chapter has described the most relevant anatomical and functional features of the human auditory system, and given particular attention to the problem of separating a desired target sound from a noisy mixture. Bregman's research has brought together the results from many disparate experiments on the behaviour of the auditory system. Using a Gestaltist framework, Bregman integrated these results into a coherent theory of auditory scene analysis in order to describe how it is that people are able to hear individual sounds in noisy environments. Essentially, his theory can be described as stating that individual auditory events are grouped together on the basis of some perceived similarties in the elements of the signal.

Bregman's work, however, does not dwell substantially on the problem of reverberation which, in addition to interference, is known to be a significant problem in speech understanding and acoustic signal processing. In auditory processing, reverberation poses such a difficult problem because of the fact that the resulting temporal smearing effects tend to obscure many of the cues that humans would normally

use to segregate the contents of an audio mixture. However, it is known that some of these cues are more resilient than others to the effects of reverberation, and hence it is expected that the auditory system assigns some greater perceptual weighting or ranking to those cues in such environments.

Thus far, these questions have been approached from a qualitative perspective. In part this is because the work that this chapter describes embodies a sizeable volume of literature, and it is more useful here to simply summarize the conclusions found therein. In addition, because the focus of this thesis is on practical applications, it is also useful to ignore many of the neurobiological mechanisms that have been proposed as they offer few useful solutions for the engineer. Instead, the ideas discussed in this chapter are meant to lay the foundations for the next chapter, in which computational approaches to auditory scene analysis are reviewed, and their relationships to possible applications are discussed.

Chapter 3

Computational Auditory Scene Analysis

3.1 Computational Auditory Scene Analysis

In the previous chapter, it was seen that in a cocktail party environment, humans perform a structured analysis on the incoming signals. This permits a form of perceptual grouping to be carried out, which allows individual sources to be extracted from the received mixture. We know from experience that not only is this a task that our brains are remarkably adept at, but that it is a problem that traditional signal processing algorithms are largely unsuited for. As a result, the idea of using the principles of auditory scene analysis as the basis for developing new signal processing algorithms has been an active area of research for some time now.

Computational auditory scene analysis (CASA) is the common term for this area of research. Typically, this work is focused on heavily constrained problems, where the sensor configuration is restricted to either monaural or binaural arrangements. Generally speaking, the goal of such a system is to achieve human-like capabilities of signal separation and understanding in realistic environments. More realistically however, the research problem is often more application-specific, focusing for example on speech enhancement, music transcription, and so on. In such cases, CASA principles are applied

to solve a narrow range of problems in order to overcome the deficiencies of classical signal processing schemes.

In either case it is important to understand what signal separation means in the context of ASA and CASA. For Bregman, the purpose of ASA is to extract individual streams from the auditory inputs so that each stream corresponds to a separate source in the external environment. In a neurobiological context, attention can be switched between these various extracted streams more or less at will. However, in a computational context, the problem is generally easier as the desired target signal is often known in advance and this foreknowledge is related to the specific application in mind. In other words, it can be assumed that for CASA systems at least, there is already some means of distinguishing the desired target from the background noise.

3.2 Time-Frequency Masking

For the actual process of ASA, Bregman described what is essentially a two-stage process. The first stage is the segmentation stage in which the input is decomposed into a number of different time-frequency regions. In the second, or grouping, stage, the individual time-frequency regions are grouped into different streams that are likely to represent a separate acoustic source. The operation of CASA systems is roughly similar, following an analysis-grouping-resynthesis approach, in which the extracted signals are resynthesized in order to produce the desired audio output[2].

In the computational context, the means of time-frequency representation is no longer a train of neural impulses, but rather some signal-processing based approach. This typically involves a filter bank designed to mimic the frequency decomposition properties

of the human cochlea. While this is useful from the point of view of biological modelling, there are also good engineering reasons to use such an arrangement. Generally speaking, gammatone filters[49][50] are used for this task, although other systems are certainly possible.

The discrete, two-dimensional time-frequency representation of signals gives rise to a very intuitive approach to the formation of auditory streams. This natural approach makes use of a mask such that time-frequency bins identified with the target are retained while the energy of the non-target bins is suppressed[51][52]. The idea of masking follows the work of Jourjine[27], who proposed that there was little overlap between the time-frequency representations of two or more sources. This sparseness means that the elements of the mixture that correspond to the interference signals may simply be zeroed out with little or no loss of the target signal's quality.

Wang and others[53] subsequently formulated the notion of mask formation as the primary goal of CASA-based systems. In their analysis, given some time-frequency representation of a mixture, where the target energy in an individual bin is given as $s(t,f)$ and the interference energy is denoted as $n(t,f)$, then the resulting ideal binary mask $m(t,f)$ is computed as

$$m(t,f) = \begin{cases} 1 \text{ if } s(t,f) - n(t,f) > \theta \\ 0 \text{ otherwise} \end{cases} \tag{3.1}$$

where θ represents some threshold, usually chosen to be 0 in order to obtain an SNR-based decision boundary of 0 dB.

In the literature, it has been found that ideal masks of this sort are capable of producing a high-quality audio output signal from an acoustic mixture[51][54]. Work by Brungart et al[55] has shown that a decision criterion set anywhere between -5 and 5 dB

can result in good quality speech separation. If the criterion is set to be much lower, too much noise is retained in the signal, while a higher threshold removes too much of the desired target, resulting in distortion. Work by other researchers has also found that a masking threshold level set below 0 dB is on the whole preferable from the perspective of human listeners[54]. This squares well with observations from the telecommunications field, which indicate that human listeners object less to noise than to distortion[57].

Use of the binary mask, however, presumes that there is little overlap between the time-frequency representations of the target and interference. In practice, however, this is rarely the case, and is certainly false in cocktail party environments, where background babble makes up much of the noise. In addition, the limited time-frequency resolution[58][59] of practical decompositions increases the probability of spectro-temporal overlap. This can lead both to errors in mask estimation, as well distortion artifacts such as musical noise, as different parts of the signal are turned on and off. It has further been shown by Li and Wang[60] that the binary mask is not optimal from the perspective of the SNR under the condition of signal overlap.

As a result, it has been proposed that a real-valued mask be used instead of the binary mask. However, this does make the problem of estimating the mask values more difficult, owing to the fact that the basis for computing the mask value itself is not clear. Li and Wang[60], for example, have proposed using a simple power ratio for computing the ideal real-valued mask on the basis of its similarity to the well-known Wiener filter. However, this presumes that the target and interference signals are already known, a situation that is unlikely to occur in reality. More usefully, Barker et al[61], working in the context of missing data speech recognition, proposed that the masks used in that

application not be calculated solely on the basis of "data present" / "data missing", but rather on the basis of the system's confidence about the reliability of the data it has. Specifically, their system generated a "fuzzy" mask obtained by compressing the estimated SNR in a sigmoidal non-linearity. In such a way, the reliability estimates of time-frequency units near the system's base noise level were assigned a value of ~0 and those with significantly greater energy were assigned a value of ~1, with intermediate signal energy levels falling somewhere in between.

Although Barker's method relied on a stationary or slowly-varying noise environment, the concept of soft masking has also been extended to non-stationary environments. However, in these cases, the authors have made use of statistical models such as Hidden Markov Models[62][51], MaxVQ[75], or Gaussian Mixture Models (GMMs) [63][64][65], which determine the probability that a particular time-frequency bin is associated with a particular output stream. Unfortunately, these methods are computationally intensive, require prior knowledge of the source characteristics, and are in any case not well suited to on-line execution.

The case of the GMM method is illustrative of the weaknesses shared by all of the afore-mentioned statistical methods of mask estimation. A signal $y(t)$ is received from a single sensor, which is known to be a combination of interfering signals and noise such that

$$y(t) = x_1(t) + x_2(t) + n(t) \tag{3.2}$$

where the $x_i(t)$ are the distinguishable signals, and $n(t)$ is the noise term. In the case where there are several different types of possible sources, the probability distribution for

a given source type can be described using the GMM distribution for the short-time spectrum of $x_k(t)$ for the k^{th} source type

$$p(X_k(t) \mid \Lambda_k) = \sum_i u_{k,i} N_C(X_k(t); \lambda_{i,k}) \tag{3.3}$$

where Λ_k is the source type, which is defined by the model parameters $\lambda_{i,k}=\{\mu_{i,k}, \Sigma_{i,k}\}$ of the complex Gaussian distribution $N_C(\cdot)$. The term $u_{k,i}$ represents the individual weighting for each Gaussian function, such that $u_{k,i} \geq 0$ and $\Sigma_i\ u_{k,i} = 1$. The separation algorithm then decides which source type is currently active by comparing the results across all possible models and determining which model is most likely to have occurred. Mask estimation is then carried out by assigning weights that are some function of the probability of the given source model and its desirability as an output.

The drawbacks of this approach can be seen from the form of equation (3.3). Most notably, it is necessary to have access to pre-existing source models. In the case of general speech separation from a non-speech background noise (such as music), it is possible to have a single speech model as is done in [1]. However, for speech-on-speech interference, individual speaker models must be trained from clean data[2]. This is obviously not practical in the kind of general environments that this thesis is considering. Equation (2) also does not take into account the effects of reverberation, or at least changing levels of reverberation which will have unpredictable effects on the received source distributions.

It can be said that it may be possible to work around this need for precise source models either by using more general class models, or by attempting to learn on the fly. However, both of these ideas, while perhaps in some sense *theoretically* possible, are impractical given the constraints under consideration.

With respect to the first proposition, that speaker class models can be substituted for individual speaker models, it should be apparent that the reduction in the complexity of the problem is minimal. Decisions must be made for example, regarding the granularity of the classes, which needs to be sufficiently fine so as to permit good separation between different talkers, and under any environmental conditions than can be expected to arise. In addition, there also needs to be a system that creates the relevant feature vectors, and another that is capable of correctly classifying them (and so producing the correct speaker class). Both of these systems are non-trivial, and require significant memory and processing resources in order to store and compute the relevant data.

The second problem, that of learning on the fly is even more complex, given the reduced data set from which to start with, as well as the need to perform such learning in an environment that is constantly changing. All of the problems described in the previous paragraph are still present, with the added complexity of an on-line learning algorithm that attempts to update the source classes and probabilities on the fly. Given that the environment, the target characteristics, and even the identity of the particular target source are all changing while this learning is being carried out, it would be exceedingly difficult to describe the mathematics of the necessary algorithm, let alone implement it.

The problems associated with the GMM algorithm are also shared by the other model-based methods described earlier. For a small, wearable device it is not practical to store all possible source models (much less know them in advance), and then compute the

associated probabilities for each model. Similarly, it is not clear how the effects of different levels of reverberation can be dealt with in this mathematical framework.

For the more general environments in which humans function and in which it would be desirable for signal enhancement devices to work, it is not possible to know the source distributions in advance, or to iterate over the data. Instead, mask estimation must be performed on the fly, and with only the information that is immediately available to the sensors. For the monaural or binaural systems of the sort that we are interested in, this means that the primary sources of information available are the grouping cues outlined by Bregman, which were discussed in the previous chapter. Fortunately, as will be explained in the following section, there are computationally simple methods of extracting these cues that are rooted in traditional signal processing.

3.3 Computational Acoustic Cues

There are four principal cues used for auditory grouping that have proven to be useful for mask estimation in computational systems. These include **pitch**, **interaural time differences**, **interaural intensity differences**, and sound **onset times**[12]. What follows is a brief description of each of these cues, how they are computed, and what their limitations are with respect to the estimation of auditory masks. It should be stated that the understanding of how these cues behave is crucial to the development of better hearing aid algorithms, as it must be understood when a cue is reliable, as well as how different cues can be best fused.

Pitch

From the point of view of CASA, the fundamental frequency, or pitch, is useful because it is an important grouping cue; auditory streams with the same pitch are likely to be from the same source, and thus should be grouped together. However, this assumes that the pitch can be reliably estimated, even in noisy and reverberant environments. While the problem of detecting and estimating pitch in quiet and non-reverberant environments is one that is well explored, the problem of performing such estimation in more difficult environments is not. Several approaches have been proposed[66][67][68], although in-depth discussions of the related performance issues only seem to be available for the method described by Wu et al. in [68].

In Wu et al.'s approach, the pitch is estimated using two slightly different methods depending on the centre-frequency of the band of interest. If a low frequency band is being explored, a straightforward autocorrelation function is used, taking the form of the equation shown below[68]:

$$\mathrm{ACF}(c, j, \tau) = \frac{\sum_{n=-N/2}^{N/2} r(c, j+n) r(c, j+n+\tau)}{\sqrt{\sum_{n=-N/2}^{N/2} r^2(c, j+n)} \sqrt{\sum_{n=-N/2}^{N/2} r^2(c, j+n+\tau)}} \tag{3.4}$$

where *r(.)* represents the subband signal of interest, *c* is the channel, *j* the time step, and τ is the time lag of the autocorrelation function. For a given time-frequency unit *r(c,j)*, the first peak not located at the τ=0 position should, under ideal conditions, indicate the pitch period of the designated channel. For high frequency signals, the method is the same except that the subband signals, *r(c,j)*, are replaced by their envelopes in order to avoid problems associated with unresolved harmonics[68][33][69].

In many applications such as in[69], the overall signal pitch is then afterwards estimated via the summary autocorrelation function

$$\mathrm{SACF}(j,\tau) = \sum_{c=1}^{M} A(c,j,\tau) \tag{3.5}$$

where the overall pitch period can then be estimated by finding the time lag associated with the largest peak of SACF(j,τ). However, simply doing so is not completely desirable, as it ignores several important aspects of how the pitch signal behaves in reality, and how it is represented in the time-frequency plane. In particular, the following facts pertaining to voiced speech and the autocorrelation method should be considered:

1)Even in an acoustically clean environment, the pitch signal will not be present in all subbands. In noisy environments, some bands will be dominated by different pitch signals, or have no discernible pitch. Such bands should be eliminated prior to performing the summary autocorrelation function as they may reduce the quality of the estimate.

2)For many parts of speech, the pitch signal will vary more or less continuously over time. Information gleaned from this trajectory can aid in correctly discriminating between the target and interferer. It will also aid in grouping time-frequency segments.

3)While pitch is something that is computed monaurally, it can also provide binaural information. Specifically, the target pitch may dominate the time-frequency unit from one ear, but not from the other ear.

4)While the autocorrelation method is easy to compute, it is subject to half-pitch and double pitch errors. That is, the estimated pitch may occasionally be either half of or double, the correct value.

5)The pitch period of rapidly changing pitch is difficult, if not impossible, to estimate correctly in the presence of reverberation. Alternative processing schemes are required in such a case.

6)If the pitch is not changing rapidly, then the autocorrelation can produce a pitch estimate that is robust to both noise and reverberation (see table 3.1 below).

SIR (Light Reverberation) (dB)	# of TF units at +/-5 lags. Left / Right	SIR (Heavy Reverberation) (dB)	#of TF units at +/- 5 lags Left / Right
∞	20 / 25	∞	22 /21
20	21 / 25	20	20 / 19
15	20 / 23	15	18 / 18
10	18 / 20	10	17 / 18
5	14/18	5	13 / 16
0	6 /15	0	7 / 12
-∞	0 / 0	-∞	1 / 0

Table 3.1. Change in correct pitch estimate with changing SNR for three voiced interfering signals. The numbers here are shown for both the left and right ears.

In spite of these potential problems whose design implications will be discussed later, pitch remains the most important cue available in hearing systems. In humans, it seems to be the dominant listening cue in noisy environments, and on a computational level it seems to be more robust than other cues. Therefore, from a design perspective, it is necessary that a workable CASA system incorporate pitch as a primary cue, and use the others in a supplementary role, aiding the segregation decision.

Onset

The value of the acoustic onset cue is that it aids the grouping of time-frequency units in time as well as in frequency. In other words, units that have the same onset are likely to belong to the same stream. In addition, the directional cues immediately following the onset are largely unaffected by reverberation, and so are more reliable than

at other times. In the literature[70][71], the detection of onset times is done by measuring a sudden increase in signal energy across multiple frequency bands.

However, this is not necessarily the best approach as the mechanisms may require additional filtering steps[71] or complicated thresholding procedures[70]. A more efficient and perhaps more reliable way to make use of acoustic onsets is hinted at by the variance of the ITD and IID, which was discussed in the preceding sections. Specifically, the lack of reverberation that accompanies the acoustic onset ensures that the variance of these cues drops markedly following the point of onset. The same is true of the channel-wise cross correlation coefficients as well.

We exploit this fact in our own system for onset detection, which is a simplification of pre-existing methods. In the approach taken in this thesis, the onset is determined by computing the change in channel power over successive frames, which is then compared to a pre-chosen threshold. For the i^{th} channel and the k^{th} frame, the decision function is

$$O_i = x_i(k) > \theta \cdot x_i(k-T) \tag{3.6}$$

which assigns a value of 1 to the function if the relation is true, and 0 if it is false. Unfortunately, under realistic acoustic conditions, the timing and/or existence of a clearly defined onset can be quite variable, so an estimator like (3.6) cannot be seen as wholly reliable. For this reason, the onsets must be summed across frequency channels. In addition, the binary truth value of the onset is carried over to the next frame. That is, if an onset was detected in the previous frame, then the current frame will also be registered as an onset frame regardless of whether the condition in (3.6) was met. Such an approach is

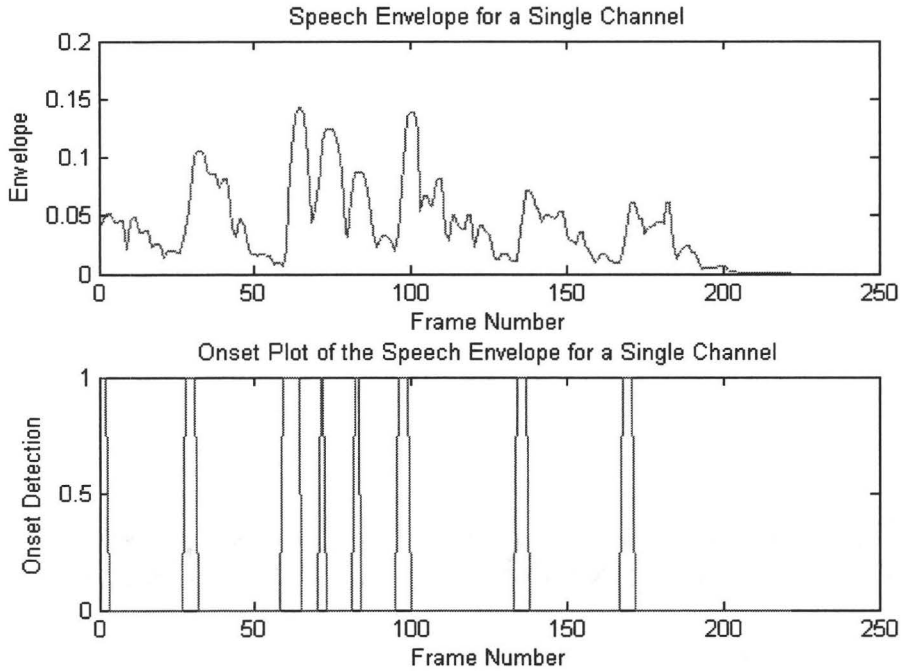

Figure 3.1. The speech envelope for a single frequency channel and the estimated onset periods for that channel calculated using (3.4).

necessary due to the fact that onset periods in the speech envelope occur over multiple frames, and an extra degree of robustness is needed under adverse conditions.

Interaural Time Delay

The interaural time delay operates on low frequencies, below 2 kHz[34], where the wavelength is long enough that phase differences between the received signals at each ear can be measured without ambiguities. Above that frequency, the ITD of the signal envelopes is calculated[54], which also corresponds psychoacoustic evidence[72]. For the purposes of computational systems, this time difference is computed using some type of cross-correlation[69][54], usually taking the following form:

$$\mathrm{CCF}(c,j,\tau)=\frac{\sum_{n=0}^{K-1} r_r(c,j+n)r_l(c,j+n+\tau)}{\sqrt{\sum_{n=0}^{K-1} r_r^2(c,j+n)}\sqrt{\sum_{n=0}^{K-1} r_l^2(c,j+n+\tau)}} \tag{3.7}$$

The overall ITD map can be computed by calculating the summary cross-correlation function in a similar fashion to (3.4). This is a convenient form for most computational systems, as it can be readily calculated, although in reality this form is not ideal. The reason for this rests with the poor temporal resolution available using (3.7), which is well below what is possible in real neural systems[73].

A significant drawback of the ITD is that it is not robust to noise and reverberation. In noisy environments, the information gleaned can be highly misleading, and as a result, the human auditory system does not use it as a significant cue in such situations. By way of example, we may consider the decay in reliability of the ITD cue according to the noise and reverberation levels. In the tables shown below, a target signal was present at an azimuth of 0°, while 3 interfering signals were present at 67°, 135°, and 270°. For a single time period, the tables count the number of frequency bins (out of a possible 32) where the target direction is correctly guessed to within +/- 4 time lags.

SIR (Light Reverberation) (dB)	# of TF units at 0°	SIR (Heavy Reverberation) (dB)	#of TF units at 0°
∞	24	∞	16
20	22	20	18
15	20	15	17
10	12	10	16
5	10	5	15
0	8	0	11
-∞	5	-∞	8

Table 3.2. Change in ITD reliability versus SIR in different acoustic environments. Note the high level of TF units indicating a target at 0° when no such target is actually present.

From the above table and from Figures 3.2 - 3.4, it is clear that the reliability of the ITD measure is highly dependent on the environment. Indeed, in some cases, it is difficult to even determine whether or not the measure is able to distinguish the existence of a real target. Any CASA system making use of this cue must therefore allow for a measure of adaptation to the environment in order to reflect a decrease in confidence in the cue's value for auditory streaming.

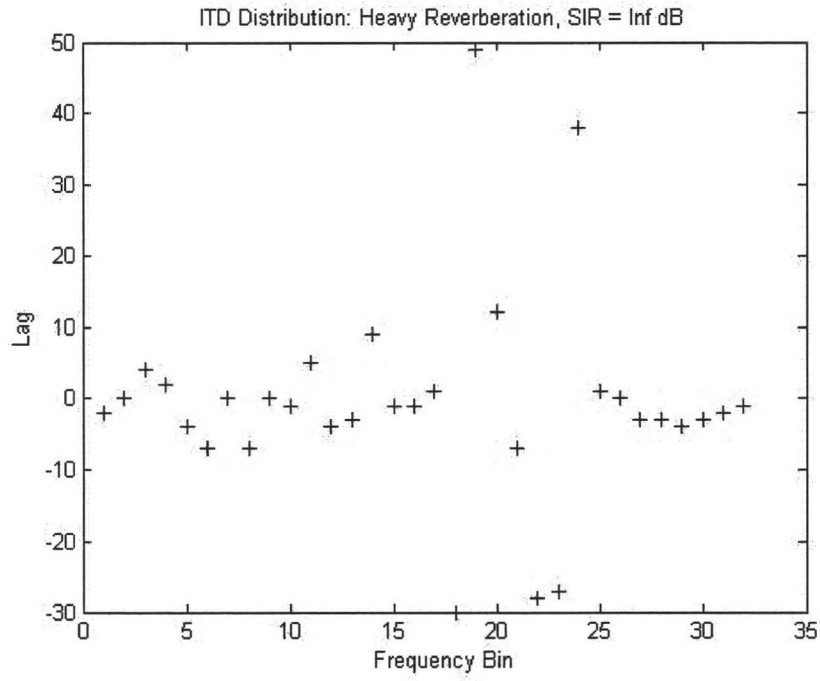

Figure 3.2. Plot of ITD lags in a reverberant environment with one target at 0 ° and no interfering signals.

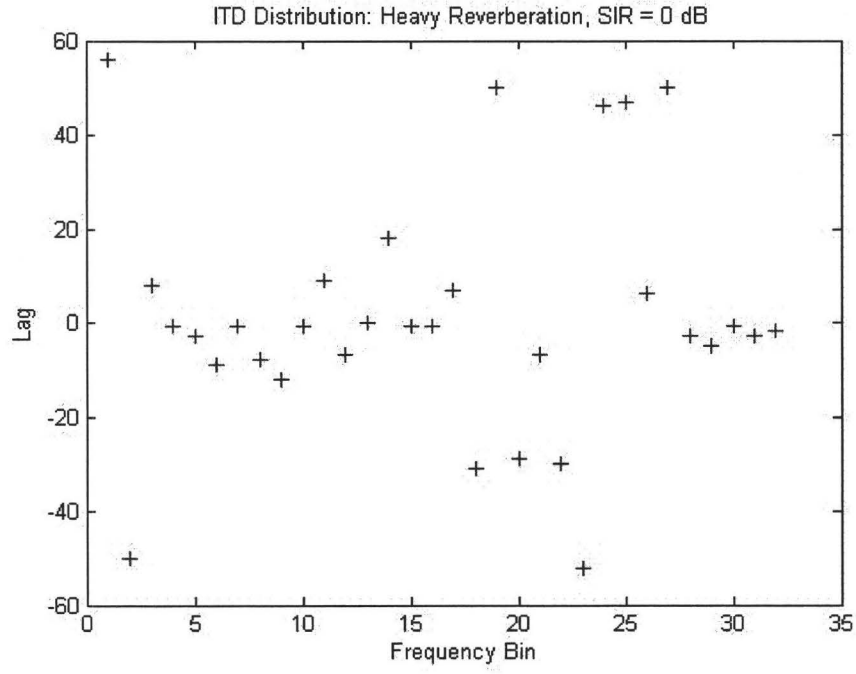

Figure 3.3. The ITD distribution for three interferers at 0 dB, shows that there is a strong clustering near 0 time lag. Further analysis must be done though to determine how to group each time-frequency block

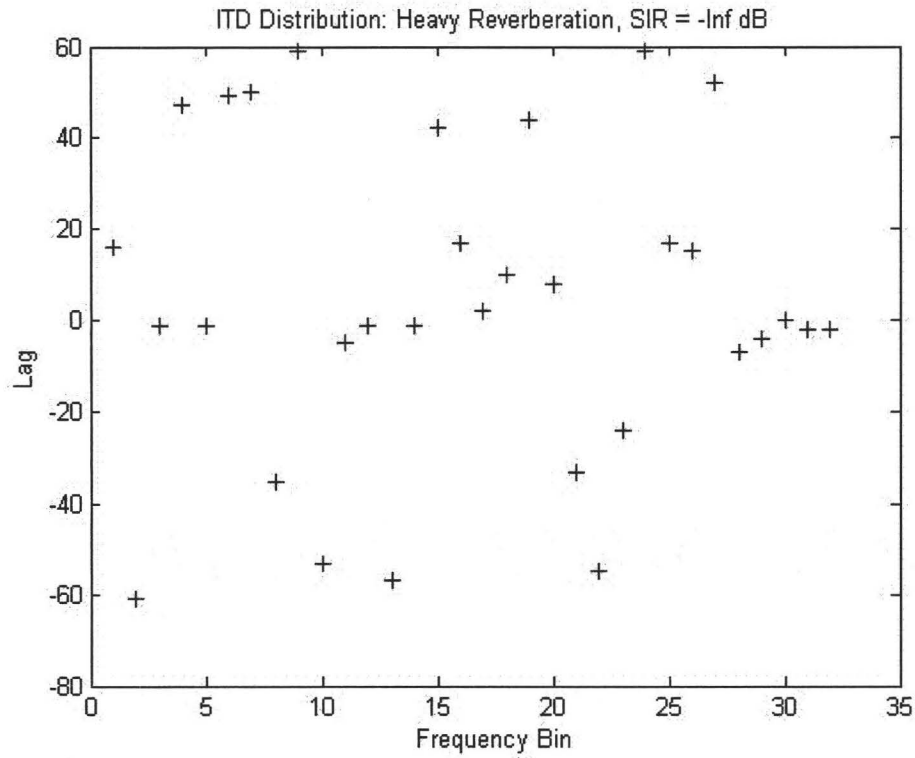

Figure 3.4. Plot of ITD lags in the same environment as above, but with no target signal, and three interferers (located at 67°, 135°, and 270°). Notice that even in the absence of a target signal, there is still a noticeable clustering around the 0° lag position

Interaural Intensity Difference

The Interaural Intensity Difference (IID) is an additional spatial cue that, like the ITD, is well-known and easy to compute. This cue can be computed simply taking the log of the power ratio between the right and left channels[69][46]:

$$\mathrm{IID}(c,m) = \log \frac{\sum_t r_{c,m}(t)^2}{\sum_t l_{c,m}(t)^2} \tag{3.8}$$

The information obtained from this cue is only considered valid for frequencies greater than about 800 Hz. As with the ITD, some care must be taken in the interpretation of this cue and how it relates to the grouping of auditory streams. Due to the presence of noise

and reverberation, there is no simple mapping that can associate the IID value of a given time-frequency region with a source from a particular azimuth.

The nature of the IID variation has so far not been well described in the literature. For example, in [69] and [74], a completely deterministic mapping is used relating the IID to a source azimuth. This mapping was calculated without taking into account the effects of noise and reverberation, and as a result has no useful effect on the quality of the source separation. A somewhat more sophisticated approach is used in [46], where the consistency of the IIDs are checked against the predicted values of the IIDs based on the calculations made using the ITDs. However, it is assumed that the ITD estimates are correct, and the allowable deviations are tightly constrained.

3.4 Cue Fusion and Stream Segregation

The basic acoustic cues that have been discussed are largely common across the CASA literature. Their computational simplicity and robustness, relative to other methods, has largely ensured that these methods have remained standard throughout the field. However, it is obvious that merely computing the values of these cues does not solve the cocktail party problem. Instead, the question of how to solve this problem focuses not so much on the cues themselves, as it does on how they are used and combined in order to make a decision about the identity of a given time-frequency region.

Perceptual Binaural Speech Enhancement (PBSE)

Exactly how this fusion occurs depends on the type of auditory masking procedure to be used. This does not preclude a certain degree of similarity in fusion

depending on the structure of the CASA system as a whole however. For example, in [69] Dong initially developed a CASA system that incorporated binary masking for the purposes of speech enhancement. Subsequent work[74], developed using a similar architecture, extended her results to include real-valued masks in order to reduce the level of musical noise.

In both cases, the system was designed for use in a hearing aid system, and as a result was developed considering many of the same limitations as were outlined in Chapter 1. Overall, both systems assumed a binauaral sensor configuration, and given the application in mind, also assumed that the desired target was always placed in front of the listener. In both systems, the stream segregation system used by Dong may be summarized as follows:

1)A bank of 32 gamma-tone filters is calculated using the method described in [49]. Assuming a 16 kHz sampling rate, the output of each frequency channel is windowed with a 50% overlap. Subsequent cue extraction and fusion is carried out on each time-windowed channel.

2)Calculate the IID and ITD for each time-frequency unit. Map the IID to azimuth according to [69]. Determine their possible association with the target based on the estimated azimuth. In the binary masking system, this means making a hard decision, otherwise the mask value is a function of the angular distance from the desired target position.

3)Calculate the dominant pitch for a particular time-frame. Assume that the time-frequency units corresponding to this either belong to the target, or that no target is present at this time. Determine a time-frequency unit's possible association with the target based on the closeness of its pitch period to the dominant pitch.

4)Using the filtering procedure from [76], determine whether or not an onset has occurred. All onsets occurring at the same time are grouped together.

5)In both cases, all cues are treated as being equally reliable, and there is no grouping across frequency channels. Each frequency band is treated independently of the others.

Data fusion in the PBSE

The original binary masking scheme developed by Dong (the PBSE) used logical 'AND' operations on each of the individual masks obtained by the cue decision algorithms. The type of arrangement used is shown below in Figure 3.5. It should be noted that the pitch

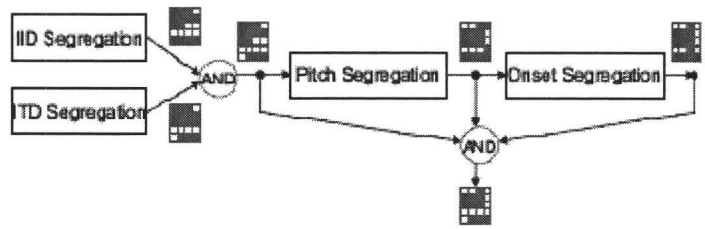

Figure 3.5. The formation of a binary mask using logical operations.

and onset segregation masks are only applied when such conditions are detected. In addition, both the onset and pitch segregation modes are considered mutually exclusive; if the one condition occurs, the other does not, with pitch taking priority. In the real-valued version of the PBSE[74], individual gains are associated with how close each time-frequency bin comes to the desired target region. In this case, the fusion algorithm is much the same as is shown in Figure 3.5, except that the binary 'AND' functions are replaced with real-valued multiplications.

From a perceptual point of view, both cases are problematic. The binary mask pattern is known to be a significant source of audible artifacts owing to the rapid fluctuations in signal strength that is caused by the binary switching of time-frequency bins. A similar problem is responsible for much of the musical noise issues in the real-valued implementation of Figure 3.5. Although it is not as bad as for the binary case, the use of multiplicative gains can still result in rapid fluctuations of the mask values in difficult environments.

Much of the problems with the PBSE are ultimately a result of placing too great a faith in the reliability of the received cues. In less reverberant environments, with a moderate SIR (i.e., >5 dB), the PBSE performs adequately. The real world is not as amenable or controllable as this however. In order to address this problem, and to design a cocktail party processor capable of operating in more realistic scenarios, a redesign of the processor is required. The following chapter discusses such a processor, in which the uncertainty of the acoustic cues was made the central consideration of the mask estimation process.

3.5 Summary and Conclusions

The neurobiological concepts described Bregman in his book "Auditory Scene Analsysis" were primarily intended for an audience of psychologists and neuroscientists. His work however, also has particular relevance for computers scientists and engineers studying machine-based solutions to the cocktail party problem. Of critical importance is the psychoacoustic principle of "exclusive allocation", in which a given time-frequency unit can only be ascribed to a single source at a time. In signal-processing terms, this concept can be readily understood as being analogous to a binary mask, in which the time-frequency units ascribable to a particular source are retained, whilst the others are filtered out. The engineering problem then is constructing a mask that provides a good estimate of the desired signal.

In many ways, it turns out the problem of mask estimation is also very similar to its psychoacoustic counterpart. Various low-level acoustic and signal-based cues can be used to extract information about the identity of a particular time-frequency entity,

information that can be subsequently combined to form larger auditory streams. The cues described by Bregman which include the ITD, IID, pitch and the signal onset time, can all be readily understood in signal processing terms, and can be readily calculated using well-known procedures.

Various researchers, such as Wang, Palomaki, Dong, and others have considered the problem of developing solutions to the cocktail party problem that incorporate Bregman's ideas. Of these, Dong's is the most relevant to the goals of this thesis, since in contrast to other approaches, it was developed as a potential real-time solution. However, as with many other algorithms, her approach does not adequately take environmental effects into account. This results in a poorer level of performance in reverberant environments than is desirable.

Nonetheless, this work is a valuable starting point for further research regarding the problem of developing a wearable cocktail party processor. In the following chapter, modifications to Dong's work are proposed that offer substantial gains in performance, that both improve interference rejection, and allow for operation in environments with higher levels of noise and reverberation than had previously been the case.

Chapter 4

Fuzzy Logic Cocktail Party Processor

4.1 Introduction: The Cocktail Party Problem Revisited

The principal current problem of CASA is not how to estimate the cues needed for grouping, but rather how to make adequate use of them in order to estimate the target speech, and to do so while meeting the desired standard of quality. This is not a straightforward problem given that the information needed for such estimates is often of uncertain quality, and usually time-varying as well. In fact, the statistical distributions that determine how much confidence we can have in the measured cues are also time-varying and ultimately impossible to know. The fundamental difficulty for engineers arises from the generality of the problem; in a realistic scenario involving a wearable device, there are few constraints on the types of sources, their complexity and placement, or on the nature of the acoustic environment itself.

It is therefore not possible to assume that the statistical properties of the sources are known, as is the case in [64][77][78], or that inverse filters aimed at removing reverberation effects can be estimated in a timely manner[21]. Likewise, the actual distributions of the auditory cues themselves are unknown and essentially unknowable within the constraints of the desired application. Therefore, while there is a pressing need

for our CASA system to adequately cope with the aforementioned uncertainties, the system itself must be developed with the knowledge that probabilistic representations of these uncertainties are of limited use.

However, as noted in the Chapter 3, there are perceptual reasons for using softmask techniques in order to estimate the target signal. Instead of making hard binary decisions about the identity of a time-frequency unit, these techniques assign a real value to the corresponding mask element based on the probability of the time-frequency unit belonging to the target. While good results have been reported using such methods, they rely on the existence of known probabilistic models, which are unavailable for the applications we are considering. As a result, this thesis proposes to use non-probabilistic measures of uncertainty for constructing the time-frequency masks and incorporate them into a novel, hierarchical system of cue fusion and stream segregation.

4.2 Approaches to Cue Fusion and Uncertainty

As was shown in Chapter 3, the cues received by both the auditory system and any proposed cocktail party processor contain significant uncertainties. These uncertainties affect all cues, and will also vary according to the properties of the current acoustic environment. In order to cope with such ambiguity, it is necessary to make use of a cue fusion scheme that explicitly allows for using such uncertain evidence.

Bayesian Methods

For a cocktail party processor based on time-frequency masking, the goal of the processor is to estimate the mask based on the information that is observed regarding the acoustic environment. In concrete terms, this involves making a determination as to

whether a given time-frequency unit $X_{k,t}$ should be identified as belonging to the target stream based on the information provided by the acoustic cues. For the softmask approach, this is not simply a binary decision, but instead the gain applied to $X_{k,t}$ depends on the level of confidence, or the probability that the unit should be identified as belonging to the target source.

To estimate the mask in such a way, it is necessary to determine the probability that $X_{k,t}$ is part of the target given the observed cues. That is, it is necessary to know

$$p(X_{k,t} = \text{Target} | C_1, \ldots, C_N) \tag{4.1}$$

for some set of observable cues C_i. For such problems of conditional probability, it is necessary to make use of Bayes' theorem[103] in order to obtain the desired result. For a pair of simple, discrete events, A_i and B, where $A_i \in \{TRUE, FALSE\}$, the theorem can be written as

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_j P(B \mid A_j)P(A_j)} \tag{4.2}$$

where B is the event that the probability of A_i occurring is being conditioned on.

For the case under consideration in this thesis, the event A_0 is that the time-frequency unit $X_{k,t}$ belongs to the target stream, and event A_1 indicates that it does not. However, for this case the cues are not discrete events, but rather continuous variables, and must be described in terms of their probability density functions. This leads to the following expression for Bayes' theorem given a single cue

$$P(X_{k,t} = A_0 \mid C_1 = c_1) = \frac{p_{C_1}(c_1 \mid X_{k,t})P(X_{k,t} = A_0)}{p_{C_1}(c_1 \mid X_{k,t} = A_0)P(A_0) + p_{C_1}(c_1 \mid X_{k,t} = A_1)P(A_1)} \tag{4.3}$$

where $p_{C_1}(c_1)$ is the probability distribution of the cue C_1. The extension of this formula needed to cover multiple cues is relatively straightforward, and can be found in[104].

This form of estimation is well known in the statistical literature, and often yields good solutions to many problems, as it is optimal so long as the required probabilities are known. Interestingly, there is also evidence that some form of Bayesian estimation is performed in neuro-sensory systems[105], which could make it a good match for any proposed cocktail party processor. If this is a successful strategy in biological systems, it stands to reason that it may be successful for a machine as well.

However, this method is not without problems for practical implementations, especially with respect to the cocktail party problem. Most important is the question of how to estimate the probability distribution functions in (4.3). For example, given the problem of determining whether or not a given time-frequency unit $X_{k,t}$ is part of the target signal based only the spatial cues (IID and ITD), consider what prior knowledge must first be available. Following (4.3), the probabilities of the IID and ITD values must be known in advance as a joint probability function conditioned on $X_{k,t}$. Learning such room-specific distributions on the fly is undesirable for the same reasons discussed in §3.2. Room classification is likewise an infeasible solution for the same reasons of complexity. In particular, it must be understood that the detection and classification of reverberation is not as straightforward as classifying the features of speech, which are at least reasonably well defined.

The basic parameters of these probability functions depend on both the intensity of the noise and its related spectral characteristics, as well as the level of reverberation. At the same time, some determination of the prior probability of the target being present

must also be known in the form of $p(X_{k,t}=A_0)$. All of this, of course, assumes that the basic forms of the probability distribution functions are also known in the first place.

In real biological systems, this prior knowledge is acquired through a combination of evolution and neural adaptation. As a result, even when the neural estimates are not strictly optimal, a "good enough" approximation is still possible. However, the problem under consideration here is not a theoretical one, but a practical one: from a designer's point of view how can this information be acquired and thence encoded so as to be suitable for use in an embedded application? There are, of course, a number of adaptation methods that could be applied to this problem, ranging from hand-tuning the probabilities to optimizing a set of Gaussian Mixture Models. In any of these cases, a significant amount of work is called for.

Consider, for a moment, what would be involved in such a task. Sufficient data must first be collected in order to adequately describe the base probabilities inside a room in order to adequately characterize it. This procedure must also be repeated in as many rooms as is needed to characterize the space of all possible rooms. Starting from this base data set, probability distributions must then also be calculated for a sufficiently representative set of scenarios. It is true that optimizing the probabilities associated with target selection can be done via repeated simulations, although this leaves open the problem of the how the system will make use of this information during operation. If the probabilities are conditioned on there being knowledge of a given scenario (which may change), then some way of detecting and classifying the scenario must also be built into the system. Working without such knowledge on the other hand, eliminates the

possibility of being specific enough to achieve optimal results, but does not greatly decreasing the computational complexity.

It was stated earlier in this section that it was likely that the brain uses some form of Bayesian estimation in order to perform some aspects of sensory processing. At first blush, this would seem to be at odds with everything that was subsequently said about the practical utility of Bayesian theory. However, it must be remembered that in Chapter 1, it was stated that the architecture of the brain is fundamentally different from that of a digital computer. In particular, unlike a computer, the brain can make use of many neurons working in parallel. In [8], Eliasmith and Andersen describe in mathematical detail how such a collection of neurons is able to form the summations and probability distributions needed to form the optimal estimates of Bayesian theory. In addition, Dayan and Abbott[28] as well as Gerstner and Kistler[113] also explain how learned parameters may be physiologically encoded into the weight-space of a neural network (and thus its processing path). The parallel and distributed nature of neural coding therefore permits a very efficient approximation of Bayesian processing.

By contrast however, commercially available digital signal processors cannot make such parallel connections. Unlike true neural networks, they can perform one or at most two, calculations at a time. Also unlike neural networks, there is no physiological encoding of parameters directly in the path of computation. Parameters and data must instead be stored elsewhere, and if external memory is needed to store large datasets such as feature vector descriptions, or classification information, then one must consider the extra time consumed by memory accesses.

For a body of work such as this thesis, which is as concerned with design as it is with theory, practical implementability must be taken as an over-riding constraint. For this reason, **the theoretical optimality of Bayesian methods is immaterial if these methods cannot be run on a realistic platform**. Implementability here cannot simply be an afterthought that is considered once the mathematics is done; instead, it must be central to how the cocktail party processor is conceived of from the very beginning. From an engineering perspective therefore, an alternative approach must be chosen that is better able to meet the system requirements outlined in Chapter 1 of this thesis.

Fuzzy Logic

An alternative to Bayesian estimation is to express how well the values of the cues match up against some linguistic description of the possible categories (e.g. *Target, Interference, etc*). The attributes of the cues can then be expressed as a set of assertions whose "truth" value is a real number in the range of [0,1]. Data fusion then is a series of *IF-THEN* rules that involve carrying out some logical operations on the data in order to estimate the desired quantity or classification.

This approach is immediately attractive to the problem at hand for several reasons. First and foremost, the process of cue fusion can very intuitively be described by just such a set of *IF-THEN* rules that use somewhat vague linguistic definitions. For instance, returning to the example of using the ITD and IID cues to estimate the presence of a target signal, instead of calculating all the relevant probabilities, we can simply state that a target is present if *most* of the ITDs AND *most* of the IIDs indicate a signal

originating directly in front of the sensors. Similar rules can be described for incorporating the other cues as well.

From an implementational point of view, this approach is also very attractive, as it results in code that is very easy to follow, and thus modify or debug. In addition, the operations involved in fuzzy data fusion are also very simple from a computational point of view, meaning that only minimal computational resources need to be dedicated to this task. Bayesian estimation by contrast is computationally simple only for the special case where all of the probability distributions can be approximated by Gaussian functions. It is not yet clear that that is the case.

However, it can be said that this approach is somewhat arbitrary in that the number, shape, and boundaries of the mappings that convert the raw data into fuzzy expressions are not as connected to real mathematical concepts as probability functions are. While it is true that probabilistic approaches will produce optimal estimates when properly measured and based on correct assumptions, this state of affairs is not easy to achieve for the problem under consideration. As with fuzzy logic, a practical Bayesian cocktail party processor will have to rely, to a certain extent, on arbitrary approximations and assumptions.

Current Design and Future Work

Given the facts so far discussed, the use of fuzzy logic presents a simple-to-understand and easy-to-implement solution to the problem of data fusion for the cocktail party problem. As will be seen later, this simple approach to data fusion not only works, but in fact produces very credible results without requiring any sophisticated

optimization. More importantly, the fuzzy logic solution applied here is also very simple computationally, which is critical given the need for any proposed cocktail party processor to function on a platform with limited computing resources available.

This is not to say that more cannot be done. In this thesis, Bayesian methods have not been explored beyond their initial rejection for design reasons. It is however possible that such methods could be applied either in the context of this application, or in the context of another application where computational resources are not so limited. In addition, the optimization of both the fuzzy logic and Bayesian approaches is not a problem that has been fully addressed. Future work focused on improving the performance of the cocktail party processor should consider this to be a problem worth addressing.

4.3 A New Approach to Cue Fusion

In spite of the many unknowns that must be taken into account, certain important facts about the cues are known, and it is around these facts that a new approach to CASA must be built. In particular, it has already been demonstrated that the estimation of pitch is robust to the effects of noise and reverberation. In actuality this difference in performance is well-known, although it has not been significantly addressed in the literature, except as a straightforward statement of fact.

From results described in [7], it is known that pitch estimation is robust to reverberation provided the pitch changes slowly enough. It was also demonstrated in [70] that for onset periods within the speech envelope, the localization cues remain robust in the presence of reverberation. Based on this knowledge a new cue fusion method is discussed in this thesis that takes into account both the differing levels of cue robustness

as well as the inherent uncertainty of cue estimation in real acoustic environments. Specifically it is based on the previously discussed observations regarding the behaviour of these cues, and encompasses two basic ideas:

1)The most acoustically robust cues are the most important in terms of grouping. Less robust cues should be used in a supplementary role in order to constrain the association of the primary cues.

2)The variability of the cue distribution means that the interpretation of the cues must be in terms of the mean and variance over several channels, and not in terms of any individual time-frequency units.

The first idea, that of placing more emphasis on the most reliable cues, is fairly straightforward. As has been previously discussed, both the pitch and the onset are such cues. In both cases though, there can be significant ambiguity as to how to segregate auditory streams into target and interference signals. It is possible, for example, that at a given instant, the dominant pitch will be from an interfering signal rather than the target, in which case additional cues must be used to constrain the identity of the stream. Neither the pitch nor the onset can by itself resolve the problem of stream identity; they are both monaural cues and are thus ambiguous with respect to direction. Therefore, the initial grouping should be made using these robust cues, while their specific identification must be made using the less reliable directional cues.

This ultimately brings up the second point, that of how to use uncertain cues to produce an estimate of the target. Supplementary to the robust cues are the weaker cues such as ITD and IID, which display much greater vulnerability to noise and reverberation than the major cues used for the initial grouping. However, as the nature of the environment precludes measuring the statistical relationships, we are left with certain general rules about the cues that can best be expressed linguistically.

Formally, this new approach can best be described using the methods of fuzzy logic. This allows a way of expressing membership and fusion rules where the relationships are not clear-cut, and where the amount of information is inadequate for probabilistic forms of reasoning. For cue-fusion in CASA systems, one pitch grouping rule can first be expressed linguistically as follows:

> IF *most* pitch elements are *near* 0° AND the individual units are *near* 0°, THEN these elements belong to the target.

The italicized words are linguistic concepts that can be expressed numerically as fuzzy membership functions[79][80]. These functions range within the interval [0,1] and indicate the degree to which the inputs satisfy the linguistic relationships such as *most*, *near* and so on. Numerically, the individual membership functions can be expressed in a number of ways such as Gaussian functions:

$$\mu(x) = e^{\frac{-(c-x)^2}{2\sigma^2}} \tag{4.4}$$

Membership rules like (4.4) and others can be used to describe the approximate azimuth of the position in terms of ITD and IID, where c describes the centre of the set and σ controls the width. A more useful form of a fuzzy membership function is provided by the quadrilateral function shown in Figure 4.1. This function has an advantage over (4.4) in that it is simpler to compute, and as a result was used for all symmetric type membership functions in the actual implementation of our system. Other membership functions include operators defining linguistic expressions like "many" or "most". In these cases, a limiting function such as that shown in Figure 4.2 was used.

The fusion rules themselves are expressed in terms of the fuzzy counterparts of the more conventional binary logic operators such as AND, OR, etc. In fuzzy terms for example, the AND operator (t-norm) used to describe the simple fusion rule above can

be expressed as either[81]

$$A(x) \text{ AND } B(y) = \min(\mu_A(x), \mu_B(y)) \tag{4.5}$$

or

$$A(x) \text{ AND } B(y) = \mu_A(x) \cdot \mu_B(y) \tag{4.6}$$

where $\mu(.)$ indicates the membership functions for the respective fuzzy sets. For the OR operator, the corresponding *s*-norm is used in its place. Experimentation with both types of operators has so far found that while (4.6) generally leads to better interference rejection, its use leads to greater amounts of musical noise than (4.5). As a result, for the remainder of this thesis, it can be assumed that the *t*-norm of (4.5) is being used any time the fuzzy logic AND operator is being used.

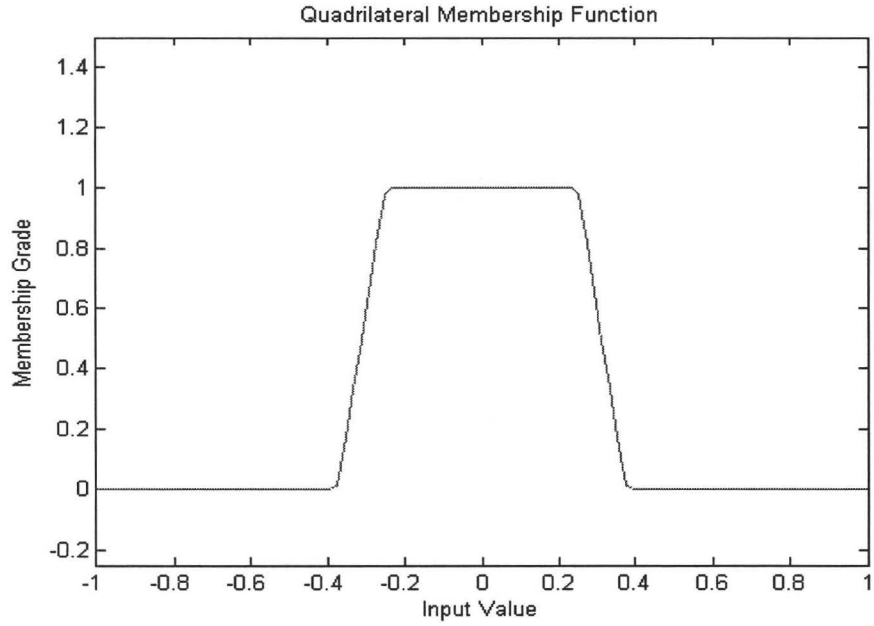

Figure 4.1. A quadrilateral membership function useful for symmetric relations. This function is suitable for operations involving a measurement "in the vicinity of" some target number.

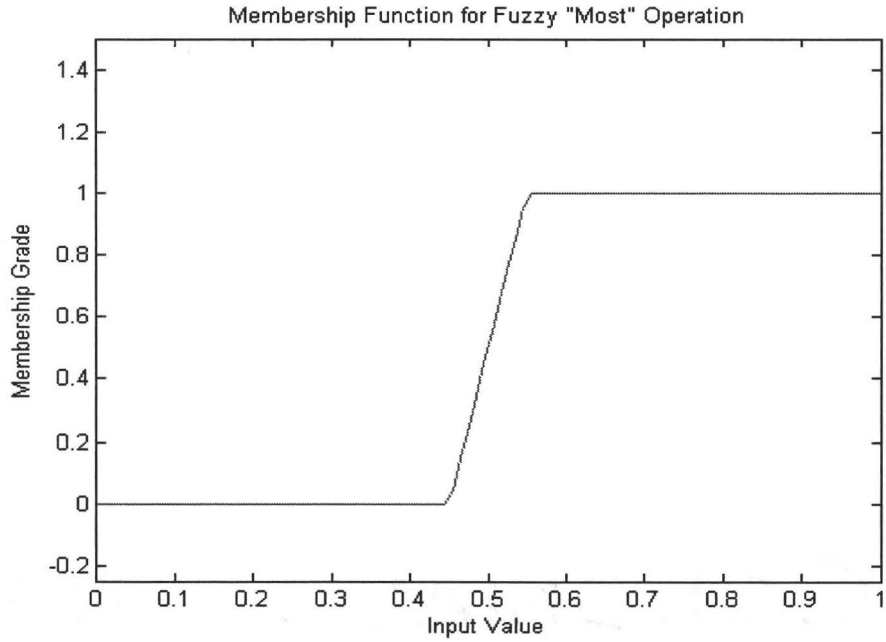

Figure 4.2. A limiting function useful for implementing the fuzzy "most" membership function.

For each of the fuzzy rules, the truth value of that rule is taken to be the level of confidence that the system has about the time-frequency element's identity. This confidence level in turn is directly related to the value of the softmask for the element in question. These fuzzy rules, however, are not universal. Each of the cues behaves differently, and must therefore be subject to different fusion and grouping rules. Figures 4.3 and 4.4 summarize the basic decision making process used by the new algorithm, and the following sections discuss these decisions in more depth with respect to their actual computational organization.

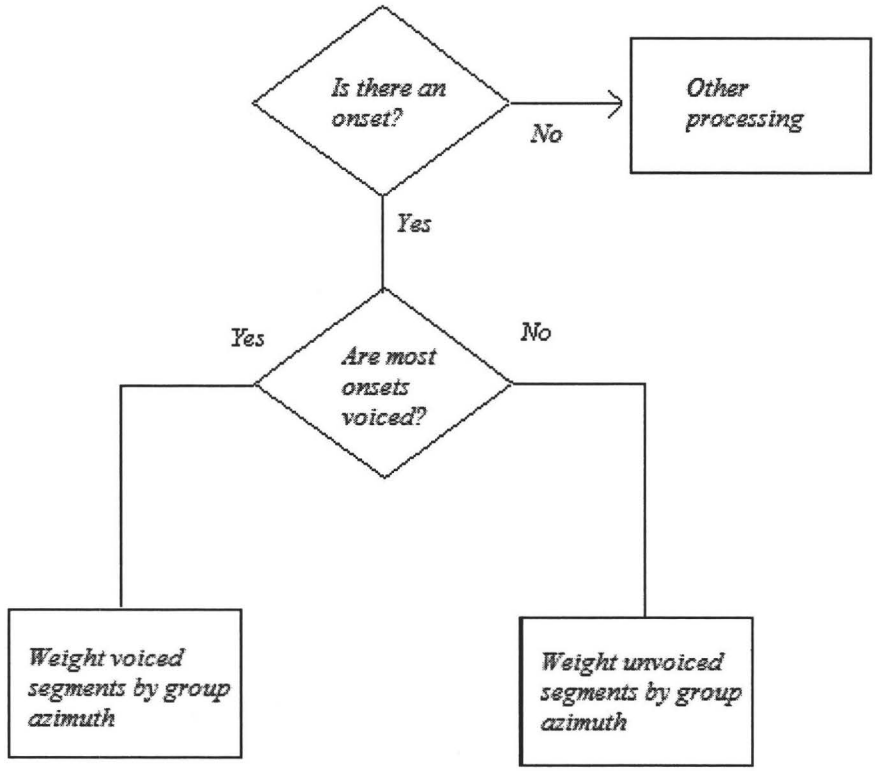

Figure 4.3. A basic flowchart describing the processing steps for input envelopes exhibiting an onset period.

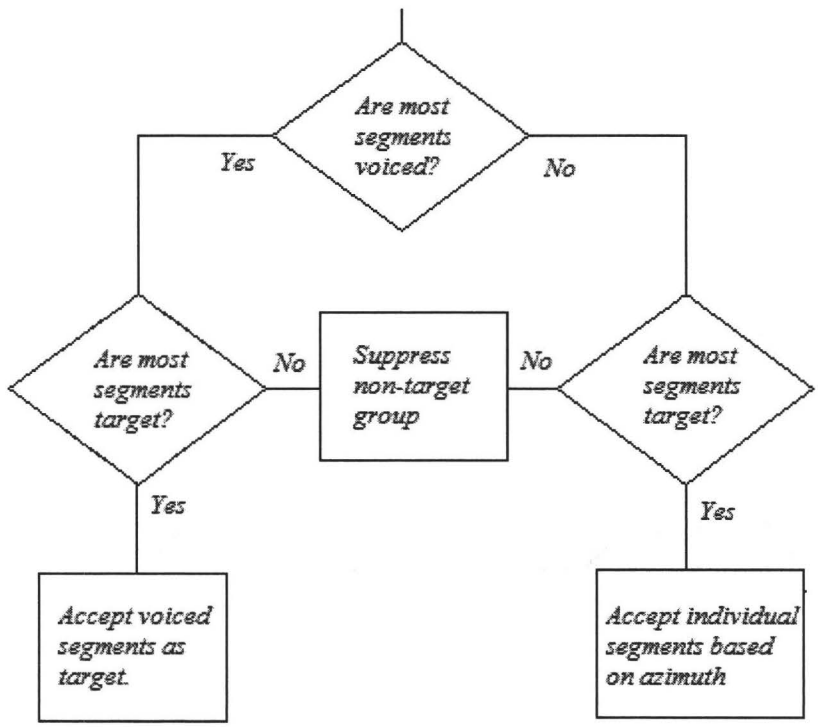

Figure 4.4. A basic flowchart describing the processing for non-onset periods.

Onset

For an individual frame, the onset cue is calculated according to (3.6). The number of frames exhibiting an onset at that time are then summed up, and subjected to the fuzzy operation

If *many* onsets have been detected, THEN "Onsets" is TRUE (4.7)

In this case, the fuzzy *many* operation is computed in the same way as the *most* operation (see Figure 4.2), albeit with a lower threshold. The result of condition (4.7) is further refined for unvoiced signals using an additional condition:

> If (*most* onset ITDs are target AND *most* onset IIDs are target) AND the current frame is an onset frame, AND the front-back power ratio is *high* THEN the current frame is target.

For voiced signals with an onset, the fuzzy condition is similar, except that all frames with the same pitch as the onset frames are also judged to be part of the target stream. Similarly, the onsets cue is also used to reject onset groups, when most members of the group are identified as not being close to the target azimuth.

Pitch

The approach to pitch estimation taken in this thesis is a modification of the method described in Chapter 3. Whereas the method used by Dong and others simply calculates the summary autocorrelation function and extracts the pitch from that result, the new system incorporates a pitch detection system in each channel, so that only channels that are likely to contain voiced speech are used to calculate the dominant pitch. The method used in this thesis is meant to provide a rough estimate of the degree of modulation present in the windowed sample, by calculating the relative modulation depth. Using (4.8), this simply involves calculating the relative difference between the maximum and minimum of the channel-wise autocorrelation function.

$$\delta(c,j) = \frac{\max_{\tau} \text{ACF}(c,j,\tau) - \min_{\tau} \text{ACF}(c,j,\tau)}{\max_{\tau} \text{ACF}(c,j,\tau)} \tag{4.8}$$

In order to ensure plausible results, the range of this function is restricted to the likely range of pitches for human voicing. This function is a variation of the ideas described in [68][82] and has been simplified in order to reduce the computational complexity. Use of this method ensures that only voiced segments are used for the following calculations, and minimizes the risk of spurious pitch peaks. A simple thresholding procedure can then be used for the purposes of detection. For the work in this thesis, a value of $\delta(c,j)>0.7$ was found to provide an adequate level of performance.

In addition to the problem of eliminating non-pitch sequences from the estimation, it must also be noted that the basic auto-correlation function is limited in resolution and also subject to estimation errors that arise from the effects of noise and reverberation. Within the summation of (3.3) for example, two individual sources may be summed together in such a way as to create a single broad peak, whose centre may or may not coincide with the actual pitch of either source[54][46]. This problem favours a sharpening of the auto-correlation function to only preserve a small region around each channel's peak. However, this focus should not be too sharp in order to avoid the small channel-by-channel differences in estimated peak time-delay that can also arise.

As a result, instead of using the basic summary auto-correlation function (SACF), or the much narrower method used by Dong, we create a "skeleton" auto-correlation function in which in the time-delay corresponding to peak-value of the channel's auto-correlation function is used as the centre for some radially-symmetric function. This results in the modified SACF:

$$\mathrm{SACF}(j,\tau) = \sum_{c=1}^{M} \phi(A(c,j,\tau)) \tag{4.9}$$

where ϕ is the radial function. The original version of this approach was developed by Palomaki[46] and Roman et al[54] for the purposes of source azimuth estimation and used a Gaussian function. However, computational limitations may render such a choice undesirable. Instead, a simple piece-wise continuous function with finite support can produce comparable results.

From the SACF, it is then possible to determine whether a dominant pitch is present, and if so, whether or not it corresponds to the desired target direction. These two steps can be combined into one single fuzzy rule, by applying the "many" membership function in conjunction with the spatial cues according to the following rule:

If *most* of the pitched frames are near 0°, THEN "pitch" is TRUE

Once it has been established that the dominant pitch corresponds to the source from the target direction, it is a simple matter to either set all of the frames with the same pitch to be target, or to reject the pitched frames that are most likely to be corrupted by noise. This latter task can be accomplished simply by using the AND or OR connectives on the spatial cues in a channel-wise fashion. For the work in this thesis, the former method was chosen because it minimized the level of audible distortion, while still ensuring a good level of interference rejection.

Directional Cues

As mentioned earlier, the role of the less-reliable directional cues is to constrain the possible grouping choices that arise from the onset and pitch cues. In that regard, the use of the ITD and IID is essentially built into the decision rules for the cases where

either signal onsets or voiced speech are present. However, in cases where neither onsets nor voiced speech are detected, the directional cues are the only ones available. In that case, the directional cues must also be used to establish the existence of a target in addition to being used for speech segregation. The fuzzy rule used for this case is as follows:

> If *most* ITDs AND *most* IIDs are *near* 0°, AND the current TF element's ITD AND IID are *near* 0°, THEN the TF element belongs to the target.

In other words, this rule only groups a given time-frequency element as target if it has been determined that it is likely that a target signal is present, and if both the current element's IID and ITD are consistent with that element actually belonging to the target.

4.4 Control

The reliability of the cues that have been discussed so far, as well as the reliability of the fusion mechanisms used to extract the target source from the mixture, depend on the acoustic environment in complex ways that are difficult to quantify. In a general sense though, it can be said that the quality of the separation that is achievable depends on both the signal to interference ratio and the level of reverberation. This quality must also be discussed in two separate ways: the degree of interference suppression, and the elimination of unpleasant artifacts in the filtered signal.

With increasing noise levels, both measures of quality suffer, and there comes a point at which not only does the interference suppression fail to improve the quality of the speech, but that it actually reduces it by introducing very noticeable artifacts. As a result of these difficulties, some control mechanism is necessary to regulate to what degree the interference suppression is applied and even if it should be applied in the first place.

In [74], the use of an adaptive smoothing parameter was proposed as a means of combating musical noise. This involved smoothing the calculated gain coefficients over time in the following manner:

$$\hat{\rho}(t,j) = \beta(t,j) \cdot \rho(t,j) + (1 - \beta(t,j)) \cdot \hat{\rho}(t,j-1) \tag{4.10}$$

where ρ is the gain calculated by applying the fuzzy fusion conditions, $\beta(j)$ is a time-varying smoothing parameter, and $\hat{\rho}(j)$ is the smoothed gain estimate. The smoothing parameter was adjusted on the basis of the estimated SNR, the form of which is described in Dong's aforementioned report.

While this approach did reduce musical noise, there still remained a significant problem with this form of distortion. As a result, the work contained in this thesis made use of a new approach that has resulted in a noticeable improvement in speech quality. Instead of the single control equation described in (4.10), the control problem has been broken into two separate mechanisms, each of which addresses a different part of the suppression/distortion trade-off. In the new approach, the smoothing formula of (4.10) is retained, although its purpose is different. Instead of adapting to the estimated SNR, the smoother adapts to the signal envelope instead[83]. This is accomplished by allowing the smoothing parameter to take on only two different values, which result from onset and non-onset periods, as shown by

$$\beta(t,j) = \begin{cases} HIGH \text{ if } onset = TRUE \\ LOW \text{ if } onset = FALSE \end{cases} \tag{4.11}$$

The change in smoothing parameter reflects the different degrees of cue reliability in the two components of the envelope. At the signal onsets, which are minimally contaminated by reverberation, the directional cues are at their most reliable, and should

be adapted to most quickly. The time periods after the onset have a greater degree of reverberation present in the signal, which lowers the reliability of the directional cues. However, due to the continuity of the speech envelope, the target time-frequency units are more likely to be in the same frequency band as the onsets, so the adaptation rate should be reduced. For this application, values of *HIGH=0.3* and *LOW*=0.1 were found to produce good results.

The second aspect of the control problem performs the original intent of the smoothing term that was introduced in [74], which is to control the problem of musical noise. In (4.10), the intent was to average out the musical noise via smoothing, at the cost of decreased adaptivity as well as a greater amount of interference. The problem of trading off the adaptation performance of the cocktail-party processor was addressed by making the smoother adapt to the signal envelope instead of the SNR. The problem of musical noise and similar artifacts can then be addressed, not by smoothing, but by selectively adding in the unprocessed background noise. Specifically, the final gain calculation for the controller is expressed as

$$g(t,j) = \hat{\rho}(t,j) + \neg\hat{\rho}(t,j) \cdot FLOOR \tag{4.12}$$

where $g(t,j)$ is the gain for the jth frame, $\hat{\rho}(t,j)$ is the smoothed gain estimate from (4.10), $\neg\hat{\rho}(t,j)$ is its complement, and *FLOOR* is some pre-defined minimum gain value. Equation (4.12) in essence works like a fuzzy **Sugeno controller** [84] because the value $\hat{\rho}(t,j)$ is not merely a gain estimate, but in fact represents the truth-value of the fuzzy conditionals that were described in the previous section.

The value of the minimum gain *FLOOR* is adaptive and depends on the estimated signal-to-noise ratio. For high SNRs, *FLOOR* is set to be low, and increases with

increasing estimated SNR. It should be stated that reliable estimation of the SNR remains problematic, because the reliability of the estimator is itself also strongly dependent on the SNR. In the current version of the software, which uses a softmask approach to interference suppression, it is not wholly possible to simply group accepted and rejected time-frequency bins. Instead, the division of target and interference power rests on the degree of confidence with which the fuzzy conditionals accepted or rejected a given time-frequency bin. This method calculates the power only where the confidence in the algorithm's acceptance or rejection is high. In other words, the value of $\hat{\rho}(t,j)$ or $\neg\hat{\rho}(t,j)$ must be high in order for the bin to be considered for SNR calculations. Once the bin has been accepted as either target or interference, the SNR is calculated normally:

$$SNR(t) = 10 \cdot \log_{10} \frac{\sum_j \left\| \hat{\rho}_s(t,j) \right\|^2}{\sum_j \left\| \neg\hat{\rho}_i(t,j) \right\|^2} \tag{4.13}$$

In the estimator of equation (4.13), $\hat{\rho}_s(t,j)$ are the target frames, and $\neg\hat{\rho}_i(t,j)$ are the interference frames.

Owing to the possibility of strong temporary variations in the background noise level and the fact that in the longer term, the background noise power is likely to be relatively constant, it is therefore necessary to smooth the SNR estimate over time. Doing so prevents unusual variations in the signal power that can be very annoying in the presence of highly non-stationary interferers such as speech. Instead, the SNR estimate is smoothed using the same basic formula as shown in (4.3), except that the smoothing parameter is a constant, and it is set to $\beta_{SNR} = 0.05$. This parameter must be set for such a high level of smoothing because the ambient noise level is expected to vary

slowly; a lower level of smoothing tends to respond more to short-term variations in the interfering signals, and thus cause too much fluctuation in the output signal power.

4.5 Post Processing Using Spectral Subtraction

The cue estimation and fusion routines that have been described so far are unfortunately ambiguous with respect to noise sources located behind the listener. The directional cues that are used to discriminate between target and interference are unable to distinguish between front and back owing to the symmetry of the problem. For this reason, another method must be applied in order to distinguish between front and back sources. Compared to the previous work on the problem of interference suppression, the solution to this problem is relatively simple, involving only a pair of rearward-facing directional microphones and the basic spectral subtraction algorithm which is described in [6]. The two rearward-facing directional microphones can be placed in relatively close proximity to the existing forward-facing ones, as is shown in Figure 4.5.

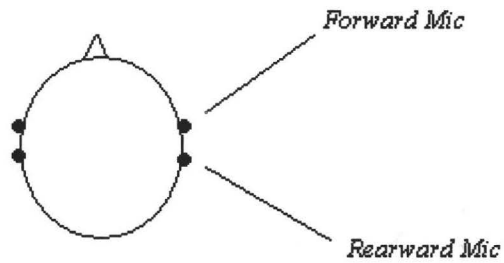

Figure 4.5. With two directional microphones mounted on each side of the wearer's head, the problem of front-back confusion can be resolved.

For this problem, a very simple algorithm was found to produce adequate results. This proposed algorithm simply assumes the signal-to-noise ratio is directly calculable

from the power ratio of the front and back microphones. This results in the gain for a given time-frequency unit to be calculated as

$$SNR(t,j) = \frac{P_{front}(t,j)}{P_{back}(t,j)}$$
$$Gain_{ss}(t,j) = \sqrt{\frac{SNR(t,j)}{1+SNR(t,j)}} \tag{4.14}$$

where $P(t,j)$ is the power in the frame at time t and frequency bin j for both the front and back microphones. The resulting gain to be applied is $Gain_{ss}(t,j)$ which is smoothed over time in the same manner as (4.10), although with a constant, rather than variable smoothing factor. Equation (4.14) is applied as a post-filtering procedure as it performs very poorly if applied before the initial interference suppression algorithm.

On a frame-by-frame basis, the gains derived from (4.14) are smoothed over time. Unlike the smoothing parameters of the mask estimator, however, the parameter for the post-processor remains fixed in time. Specifically, the smoothing rule for the jth frequency channel is expressed as

$$Gain_{total}(t,j) = 0.65 \cdot Gain_{total}(t-1,j) + 0.35 \cdot Gain_{SS}(t,j) \tag{4.15}$$

Owing to the binaural nature of the system, this calculation is repeated for each ear, resulting in two separate gain factors. This leaves open the question of whether the gains should be applied separately, or combined in some way. Informal listening experiments have found that the most perceptually satisfying approach is to apply the smallest of the two proposed gains.

4.6 Summary of Novelty

Based on the design originally devised by Dong, we have made the following changes in order to improve the performance of the cocktail party processor:

1)The cues are grouped according to a hierarchy that is based on the robustness of those cues. The identity of the segments that have been grouped are then constrained based on the average behaviour of the less reliable cues.

2)The grouped channels are now considered as a whole, and not as individual elements.

3)The fact that the directional cues are more robust during onset periods has been incorporated into the design. This was accomplished by making the smoothing rate adaptive to the signal envelope.

4)The decision and data fusion rules were reformulated in terms of fuzzy logic operations. This allowed for a change in the nature of the fusion rules, which substantially reduced musical noise.

5)A new SNR adaptive control mechanism was introduced in order to improve the perceptual performance in especially difficult environments.

6)The front-back ambiguity present in the original CPP design has been greatly mitigated via a spectral subtraction block that makes use of two additional rearward facing microphones.

Of particular importance is the fact that items (3) through (6) are unique in the literature. The items of most significance should be taken to be (4) and (6), which are particularly novel. No other CASA systems have made use of fuzzy logic to cope with the uncertainties inherent in realistic environments, nor have other CASA systems meaningfully considered the problem of front-back ambiguity. These two contributions are not only novel, but are the ones most responsible for the significant improvements in signal quality in the system described in this thesis.

4.7 Example Results

Figures 4.6-4.9 show the results of a single trial of both the original CPP, as well as our improvements to it. In these examples, there is a male target talker located in front of the listener and three other interfering talkers (two male and one female) elsewhere in the room. Specifically, the female talker is located at an azimuth of 67°, while the other two male talkers are located at azimuths of 180°, and 270° respectively. The power level of the interference was set to be equally distributed across the interferers, and was set so that the input SNR was ~1dB. The scenario was created using the R-HINT-E software using the impulses of a reverberant hard-walled lecture room.

Combining the spatially distributed interference with the presence of reverberation allowed for the testing of the FCPP and PBSE algorithms under more or less realistic conditions, which is what makes these examples interesting. In the example plots, the target signal of Figure 4.6 is shown to be significantly obscured by the three interferers, with only some parts of the target signal envelope being clearly dominant. As this is speech-on-speech interference, all of the signals present are highly non-stationary, and over the long-term occupy essentially similar spectral regions.

Application of the PBSE and FCPP both allow for a significant improvement in SNR. In the case of the PBSE (Figure 4.8), the output SNR is 4.8 dB when averaged over several successive overlapping windows of 2048 samples. Applying the same approach to the output of FCPP (Figure 4.9), the measured output SNR is 7.9 dB. This represents an audible improvement in the quality of the output sound in terms of both the level of interference as well as for the audibility of musical noise. A more complete evaluation of

the performance of the PBSE and FCPP algorithms is deferred to Chapter 6, where both objective and subjective criteria will be used in assessing output quality.

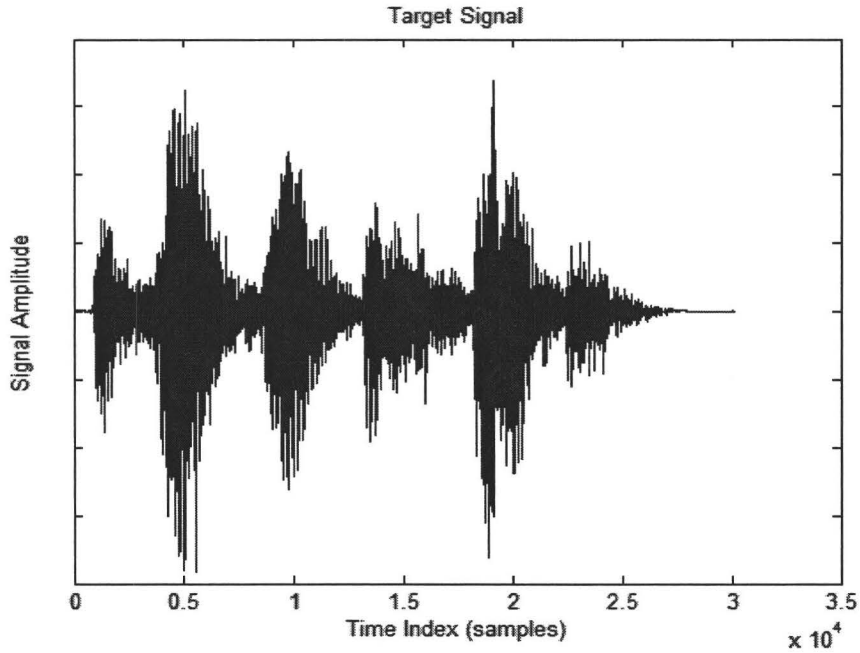

Figure 4.6. The original target signal as recorded in a reverberant lecture room.

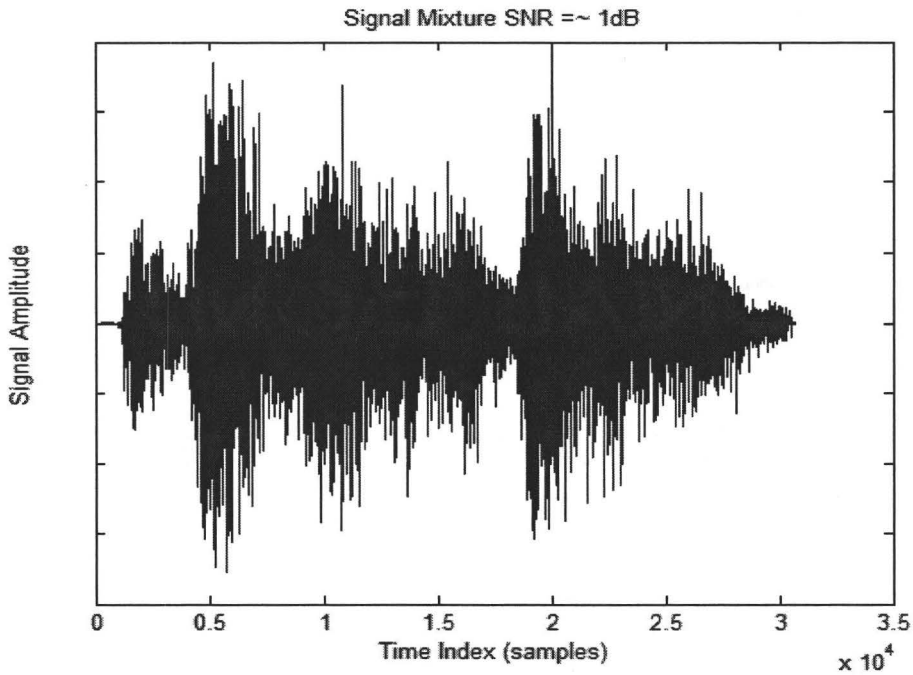

Figure 4.7. The observed mixture with three interfering talkers

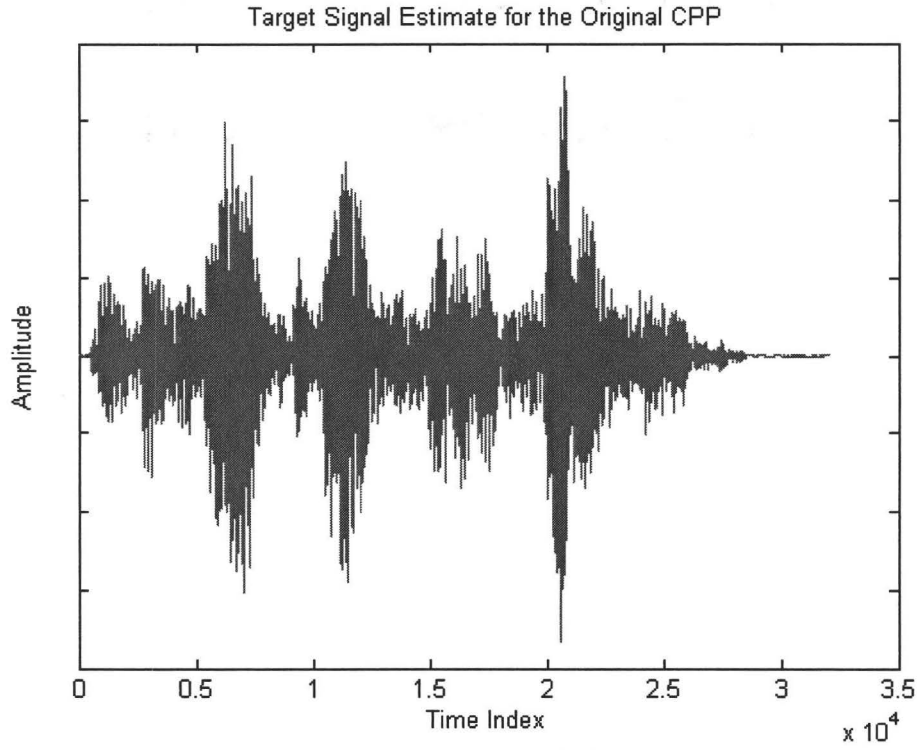

Figure 4.8. The estimated signal using the original CPP algorithm developed by Dong.

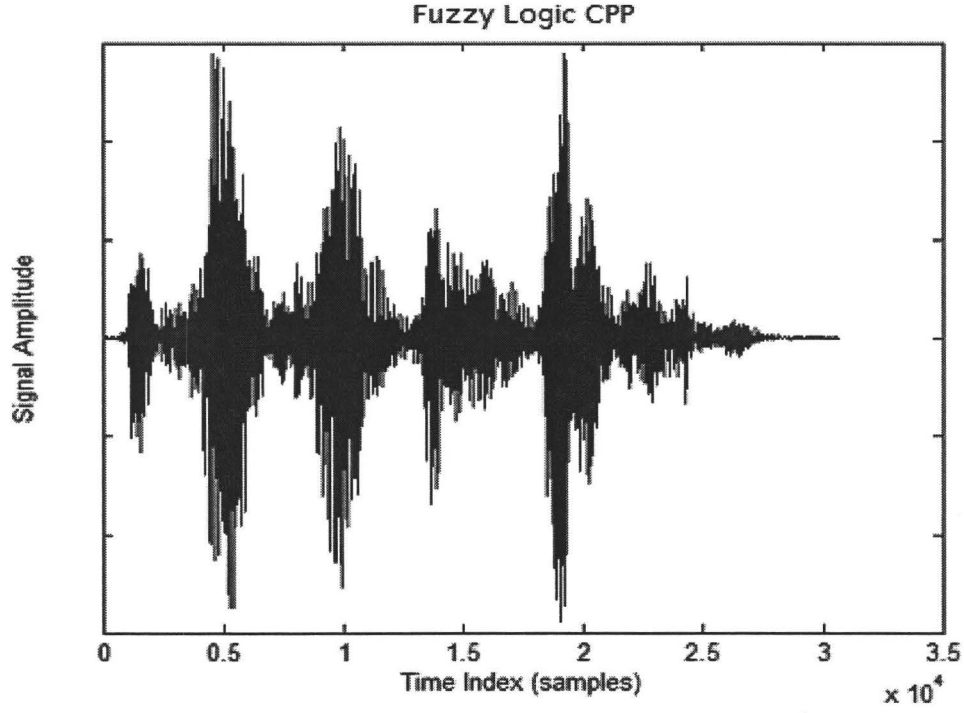

Figure 4.9. The target estimate using the new fuzzy CPP algorithm. Note the reduced levels of background noise in this figure as compared to the previous one.

4.8 Summary and Conclusions

Environmental factors such as noise and reverberation mean that the low-level acoustic cues used by the auditory system are often unreliable. In addition, it is impossible to know just how unreliable the cues are. This double uncertainty makes it extremely difficult to use many of the standard methods of statistical estimation. Instead, non-probabilistic methods of data fusion can be employed so that the essential rules of the auditory segregation and fusion processes are obeyed, while at the same time taking into account the unreliability of the information that is being received.

In this chapter, fuzzy logic was proposed as the fusion mechanism. Although this approach was initially chosen for its intuitiveness, it also represents a computationally simple solution to the problem of cue fusion. It also extends the work done by Dong in that the fusion mechanisms proposed by her (the binary 'AND' and the multiplication of gains) can viewed as special cases of the fuzzy t-norm.

Additional improvements have been made in front-back discrimination, and in the adaptation algorithms. Combined with the fuzzy logic fusion algorithms, these changes result in improvements in the following areas: interference rejection, reduction of musical noise, and graceful degradation. As shown in the example of Figures 4.8 and 4.9, there is a noticeable improvement in output SNR. However, a more complete analysis that includes perceptual criteria is carried out in Chapter 6 of this thesis.

Chapter 5

Coherent Independent Components Analysis

5.1 The Limitations of CASA

While the cocktail party processor does indeed work very well, the problem of improving the performance beyond the current limits is still very important. On its own, the performance of the fuzzy cocktail party processor declines significantly in multi-talker environments when the SNR goes below a range of around –1 to 0 dB. In such environments there is more uncertainty in the identification of the target vs. the interferer, and it is more likely that the dominant signal will not be the desired target. Therefore, one of the important goals that was set for further research was to extend the functionality of the FCPP so that it could function effectively in higher noise environments.

In particular, it would be desirable to eliminate as much of the interference from the received signals as possible before feeding them into the CASA processor. Such a scheme could therefore increase the quality of the output sound by both reducing some of the actual interference, as well as improving the reliability of the cue estimates. The overall effect would thus be to improve the quality of the resulting time-frequency mask.

Instead of using CASA techniques, therefore, such a pre-processor must be based on more traditional signal-processing methods that complement the kind of processing used in CASA. However, any such method must also take into account the limitations that were described in Chapter 1. That is, the pre-processor must function under the constraints of real-time processing, limited computational resources, and the need for a small, wearable device that can process sound binaurally. This last requirement is of particular importance as it rules out all but the simplest of microphone arrays.

5.2 Indepenent Components Analysis

The general approach[12][13] to blind source separation through independent components analysis (ICA) involves estimating N unknown independent source signals $\mathbf{s}(t)$ from a mixture of M recorded signals $\mathbf{x}(t)$. In the basic formulation of ICA it is assumed that the received mixtures are instantaneous linear combinations of the source signals as is shown in (5.1)

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{v}(t) \tag{5.1}$$

where $\mathbf{A}$ is the unknown M x N mixing matrix. The goal of ICA is to find a de-mixing matrix $\mathbf{W}$ such that that

$$\hat{\mathbf{s}}(t) = \mathbf{W}\mathbf{x}(t) \tag{5.2}$$

is the vector of recovered sources.

In most real-world acoustic applications, this model is inadequate, as it takes neither time-delays due microphone spacing nor the effects of room reverberation into account. Instead of the simple linear mixture of (5.1), the received mixtures are in fact a sum of reflected and time-delayed versions of the original signals, a situation which is

much harder to handle. Algorithms based on the linear mixing model of (5.1) are therefore inadequate for such a general problem.

However, it was shown by Ito[85], Celik and Cauwenbergs[86] and by Pedersen et al[87] that if the microphone spacing is small enough, then the problem of convolutive mixing disappears. This was demonstrated by Pedersen in [87], and also confirmed by our own experiments. In our experiments, three closely spaced in-the-ear microphones were used to record data as part of the R-HINT-E project[100]. The arrangement of the microphones is shown in Figure 5.1, and the subsequent recordings are shown in Figure 5.2. It can be clearly seen that the actual signal differences are relatively minor, and that there is no meaningful time-delay between them.

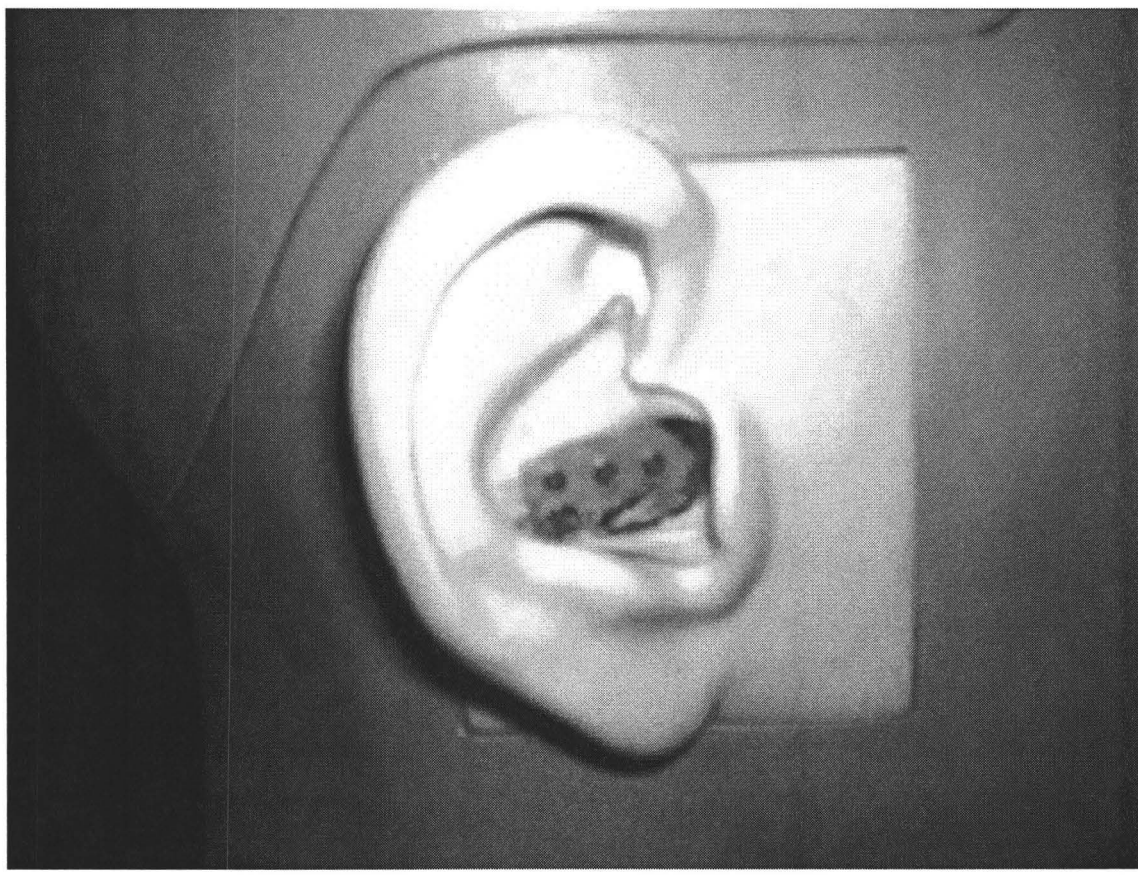

Figure 5.1. The use of closely spaced microphones limits the problems of time-delays and convolutive mixing.

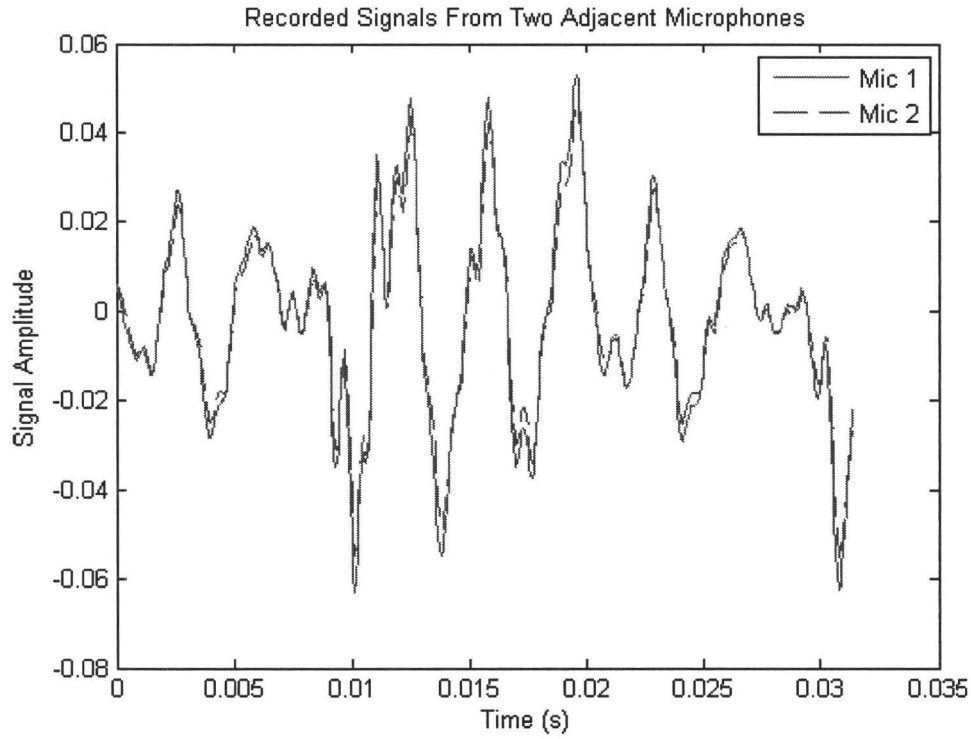

Figure 5.2. The signals recorded from two closely spaced microphones are nearly identical. The room impulse responses used for this recording were from a hard-walled reverberant lecture room.

Using directional microphones, the ICA problem can be solved using only the linear model of (5.1). Pedersen's model is particularly important given its similarity to the sensor arrangement that was proposed in Chapter 4 of this thesis. Specifically, Pedersen proposed the use of two closely-spaced directional microphones, as is shown in Figure 5.3. In the previous chapter, it was also proposed that two directional microphones mounted on each ear be used in order to solve the problem of front-back confusion. While the system discussed in Chapter 4 is more flexible in that it depends much less on the microphone spacing, it is also more limited in what it can accomplish because it is based on spectral subtraction.

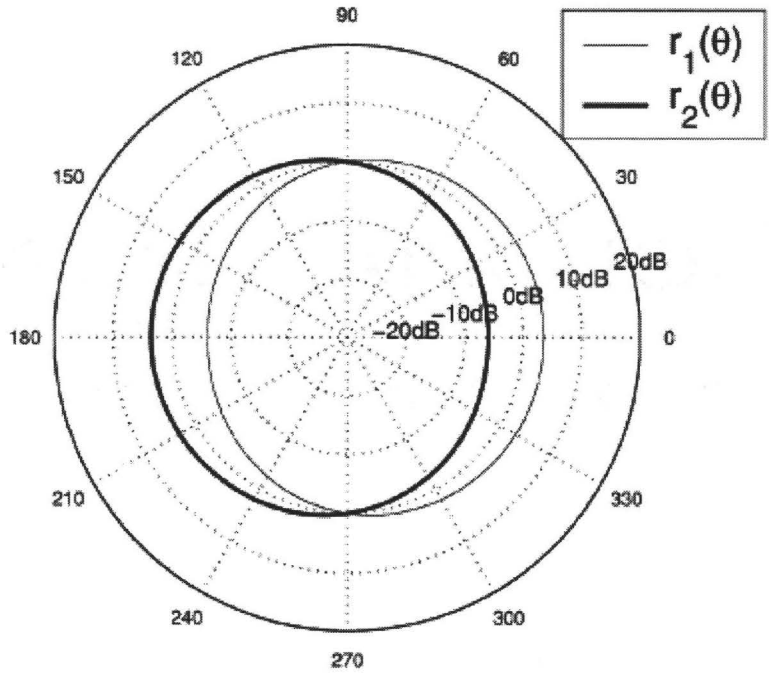

Figure 5.3. Pedersen's arrangement involved two closely spaced directional microphones pointing in different directions (Source: Pedersen, 2008).

By ensuring that the ear-mounted microphones are close enough together, it is possible to develop the kind of pre-processor described in the previous section. Because each ear would possess the same dual microphone arrangement, the binaural signals needed by the CASA system would be available for processing by that unit in the form of the pre-processor's outputs. For this system, it is not necessary for the ICA algorithm to provide full separation; all that is required is at least some removal of the unwanted interference.

However, if the ICA algorithms for each ear are allowed to adapt independently of each other, local variations in signal intensity between the left and right sensor groups will lead to some disparity in the estimated source signals. Further, given the well-known ambiguities of ICA with respect to both magnitude and permutation, there is no guarantee that the sensors on each ear will extract the desired signal at the same strength or even

that the output signals will be the same. Some additional constraints must therefore be added in order to ensure that both of the signals estimated by the ICA pre-processor are the desired target signals, and that the outputs do not confuse the CASA algorithm by distorting the acoustic cues.

5.3 Coherent Independent Components Analysis

Looking at the above problem differently allows us to put forward a possible solution. In the scenario described above, the unconstrained adaptation of the demixing filters for each ear is undesirable. There is no constraint that can prevent undesirable differences between the left and right microphone groups, if the filters for each ear are allowed to adapt independently of each other. To prevent this, any adaptation algorithm must be binaural in nature, allowing the left and right sensors to communicate in some way, so that the two groups of filters converge to a common solution.

This kind of problem was explored by Kan in the context of sensory processing in neural networks[91]. The approach he developed incorporated two different approaches to ICA, InfoMax[88] and Imax[89], into a single algorithm that he termed coherent ICA (cICA)[90]. The purpose of the algorithm was to perform signal separation on two differently mixed, but related, sets of data such as might occur in the human auditory system. The transformed outputs from each network are required to be maximally statistically independent of each other, while at the same time the mutual information between the outputs of the two different networks is to also be maximized (see Figure 5.4).

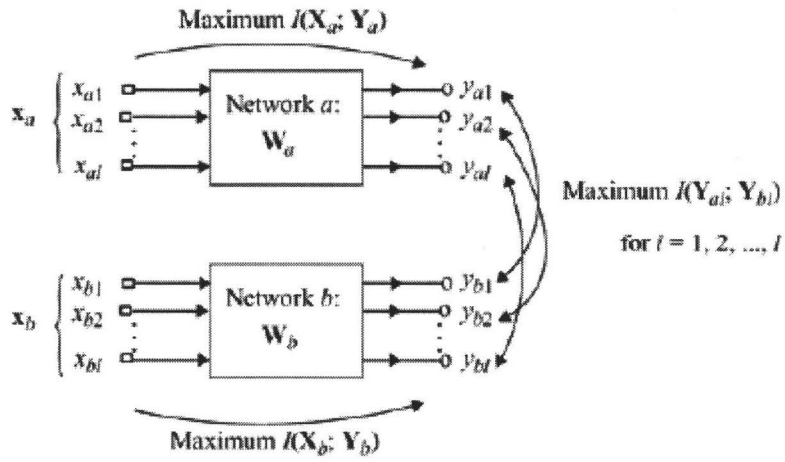

Figure 5.4. Diagram of the operation of the cICA algorithm. The maximization of the information flow across the network ($I(\mathbf{x}_a, \mathbf{y}_a)$) is traded off against minimizing the mutual information between the outputs of the two networks ($I(y_{ai}, y_{bi})$) (Source: Haykin and Kan, 2007).

Mathematically, this results in the cost function shown in equation (5.3) which is to be

$$J_{cICA} = I(\mathbf{x}_a, \mathbf{y}_a) + I(\mathbf{x}_b, \mathbf{y}_b) + \sum_i \lambda_i I(y_{ai}, y_{bi}) \tag{5.3}$$

maximized over the network weights $\boldsymbol{W_a}$ and $\boldsymbol{W_b}$. The summation is carried out across all of the elements of each output vector, and the parameter λ_i weights the relative importance of signal separation within the individual networks versus the coherence across the two sets of outputs.

Using the mathematical copula in conjunction with Sklar's theorem[92][93], Kan developed a mathematically elegant solution to the problem that also allowed for a considerable increase in computational efficiency. Working from the assumption that the approximate statistical distribution of the signals is known, Kan's work proceeded as follows: using the definition of the mutual information in conjunction with Sklar's theorem and a coherence parameter of $\lambda_i=1$, Kan rewrote the cost function of (5.3) as

$$\begin{aligned} J_{cICA} &= \sum_i E[\log \hat{p}_{Y_{ai}}(y_{ai})] + \sum_i E[\log \hat{p}_{Y_{bi}}(y_{bi})] + \sum_i E[\log c(u_{ai}, u_{bi})] \\ &= \sum_i E[\log \hat{p}_{Y_{ai}}(y_{ai}) \hat{p}_{Y_{bi}}(y_{bi}) c(u_{ai}, u_{bi})] \\ &= \sum_i E[\log \hat{p}_{Y_{ai}Y_{bi}}(y_{ai}, y_{bi})] \end{aligned} \tag{5.4}$$

where the function *c(.)* is the copula for the model distributions $\hat{p}(.)$ of the random variables Y_{ai} and Y_{bi}.

In [90] and [91] the authors chose to use a generalized Gaussian distribution to demonstrate how cICA could reduce the blind source separation problem to a simple algorithm. The generalized Gaussian distribution was chosen because of its broad applicability to a variety of problems, including modelling the statistics of speech signals. For a pair of vectors from the individual de-mixing matrices, this results in the following algorithm.

$$\begin{aligned} \Delta \mathbf{w}_{ai} &\propto \frac{\alpha}{1-\rho^2}(y_{ai} - \rho y_{bi})(y_{ai}^2 - 2\rho y_{ai} y_{bi} + y_{bi}^2)^{\frac{\alpha}{2}-1} \\ \Delta \mathbf{w}_{bi} &\propto \frac{\alpha}{1-\rho^2}(y_{bi} - \rho y_{ai})(y_{bi}^2 - 2\rho y_{bi} y_{ai} + y_{ai}^2)^{\frac{\alpha}{2}-1} \end{aligned} \tag{5.5}$$

where $y_{ai} = \mathbf{w}_{ai}^T \mathbf{x}_a$ is the estimated source, and is a product of the *i*th column vector of $\mathbf{W}_a$ with the corresponding input vector $\mathbf{x}_a$. The parameter α is the so-called "shape-parameter", which defines the sparseness (kurtosis) of model probability density. The other parameter ρ, is the correlation coefficient derived from the basic definition of the multivariate generalized Gaussian distribution. This parameter controls the degree of correlation between $y_{a,i}$ and $y_{b,i}$; a large value for ρ favours a more coherent structure being learned across the two networks, while a smaller value favours greater statistical independence within the outputs of each network. In addition to the weight update

equation of (5.5), each of the updated weight vectors is subsequently normalized prior to the next iteration.

Practical Performance Issues

Combined with the use of closely-spaced directional microphones, cICA has the potential to solve the problems discussed in sections 5.1 and 5.2. However, there are two significant performance considerations that must be taken into account. The first is whether or not Kan's incorporation of an underlying statistical signal model affects the performance of the cICA system in more generalized environments. In addition, while the use of closely-spaced microphones solves the problem of convolutive acoustic mixing, this problem is re-inserted because of the need to use a second pair of microphones on the other side of the wearer's head.

Copula ICA

Leaving aside the issue of coherency across two different networks, the issue of using the modelling approach for blind source separation used by Kan should first be looked at in isolation. In such a case, an experimental assessment is relatively straightforward. By setting ρ = 0, the algorithm of (5.5) adapts without regard for coherency, allowing a baseline for the evaluation of the non-coherent version of ICA algorithm (which will here be termed copula independent components analysis[94], or coICA).

As an experiment, two super-gaussian signals were generated using the function

$$s_i(t) = n_i(t) \cdot |n_i(t)|^{0.1} \quad i = \{1,2\} \tag{5.6}$$

where $n_i(t)$ is a normally distributed random signal. These signals were mixed using the linear mixing model of (5.1). For 100 random trials, the effects of three different shape parameters were compared in terms of the algorithm's ability to successfully recover the source signals. Each instance of the source signals was 10,000 samples long, and the algorithm was allowed to run for 100 iterations over the full data set with a constant learning rate of η=0.0015. It was found that while convergence occurred after about 16 iterations in all cases, the quality of source separation was strongly dependent on the shape parameter used, as is shown in Table 5.1

Shape Parameter (α)	**Mean Output SIR (dB)**	**Variance**	**Minimum SIR**	**Maximum SIR**
1.3	6.8	4.8	5.0	9.48
1.7	11.56	3.38	9.35	13.11
1.9	6.87	24.2	0.02	22.2

Table 5.1. The sensitivity of the copula method to different distributional models is shown above.

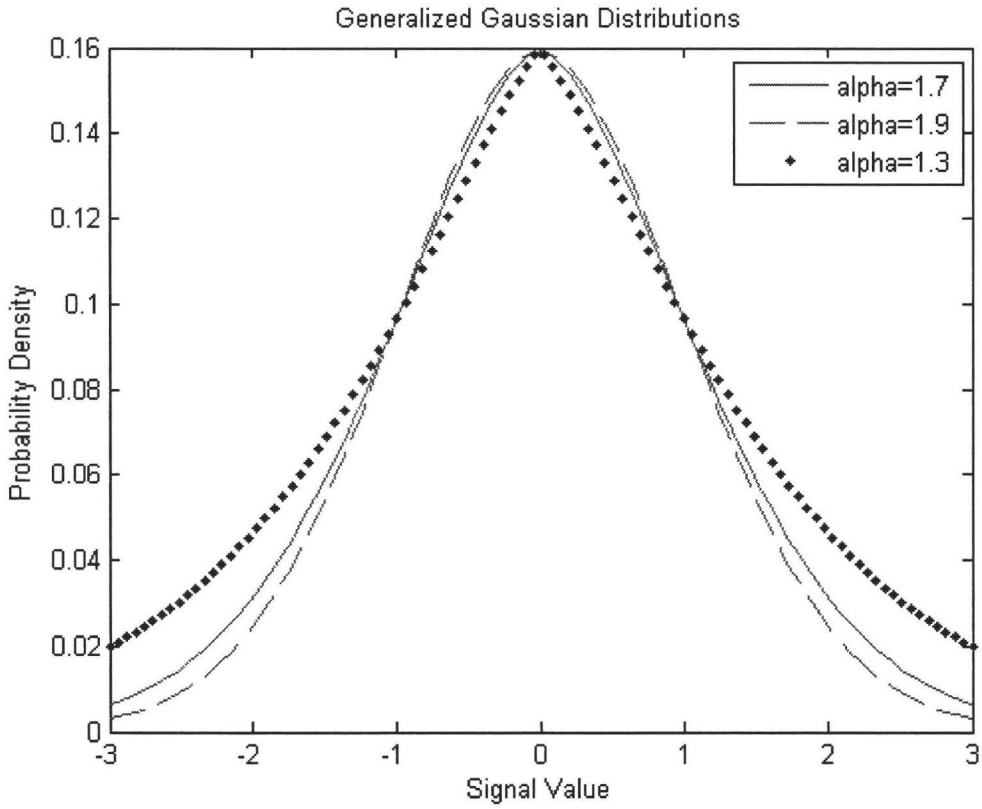

Figure 5.5 Comparison of different Generalized Gaussian probability distributions for the coICA experiment. Note the overall similarity especially for the last two cases.

It should be noted that the differences in the modelled pdf for the values of α chosen for this experiment are not large. Figure 5.5 compares the generated pdf models for the generalized Gaussian distributions used in the above experiment. The conclusion we should draw from this is that the baseline performance for the copula version of ICA, and thus for the original formulation of cICA is overly sensitive to the model distribution. This stands in contrast to the usual formulation of ICA, which is typically only sensitive to the sign of the kurtosis, that is, whether a signal is sub- or super-Gaussian. In terms of implementation in an acoustic signal processing device subject to a wide range of environments and signal types, the narrow performance range of Kan's original formulation is clearly inadequate.

Correspondence Between Imax and cICA

The original formulation of the cICA is also problematic in that it presents a redundant formula. The goal of cICA as shown in Figure 5.4 is to perform a combination of Infomax and Imax such that the outputs of one sensor group are independent of each other, whilst also ensuring maximum statistical dependency between sensor groups. However, it can be shown that the Imax algorithm is capable of performing this task on its own, without needing the Infomax component.

Given two neural networks with input vectors $\mathbf{x}_a$ and $\mathbf{x}_b$, and corresponding outputs $\mathbf{y}_a$ and $\mathbf{y}_b$, the Imax algorithm seeks to maximize the mutual information between $\mathbf{y}_a$ and $\mathbf{y}_b$ by adjusting the appropriate weight vectors. The mutual information can be expressed mathematically as[12]

$$I(y_a; y_b) = h(y_a) + h(y_b) - h(y_a, y_b) \tag{5.7}$$

However, the mutual information can also be written as[12]

$$I(y_a; y_b) = h(y_a) - h(y_a \mid y_b) \tag{5.8}$$

Expressed in this form, the redundancy inherent in the cICA algorithm becomes clear. The equation of (5.8) is maximized by simultaneously maximizing the entropy of the output $\mathbf{y}_a$ as well as minimizing the entropy of $\mathbf{y}_a$ conditioned on $\mathbf{y}_b$. However, from the formulation of the Infomax algorithm it is known that maximizing the entropy of the output of a neural network is equivalent to performing ICA. At the same time, the minimization of the second term works to maximize the correspondence between the outputs of the two networks. As a result, the Imax algorithm is able to perform the essential functions of cICA on its own.

This, however, is not the full story. Kan's form of coherent ICA does differ from this result depending on how the Imax and Infomax terms are weighted. In equation (5.3), for example, it is possible to apply a greater degree of weighting to one term or the other. By doing so, it is possible to place a greater emphasis on independence than on maximizing the mutual information. This allows for a greater range of results to be achieved using cICA than can be achieved using Imax alone.

The consequences of this result are limited by the problem that is being considered. In the above argument, it was assumed that only information theoretic criteria were being applied. However, in many realistic acoustic problems, the application of these criteria to filtering is not straightforward. It is well-known in the literature for example that the basic formulations of ICA do not work well in the presence of ambient noise or convolutive mixing. In such cases either extended versions of ICA must be employed that are suitable for such environments, or else ICA must be dropped in favour of more robust filtering algorithms[21].

For the problem outlined in the beginning of this chapter, which requires something very like cICA, the realities of the environment must be taken into consideration. Real acoustic environments feature both ambient noise and convolution, which renders direct use of ICA problematic, especially if the available computational resources are limited. This makes it difficult to apply either the cICA or the Infomax algorithms as they were originally formulated.

In the following section a novel form of coherent ICA is proposed to cope with these challenges. This work builds on Kan's original concept by proposing that the Imax component be replaced with a correlational term for greater robustness. In this case

however, the redundancy described above is lost. Correlation, unlike Imax, is unable to find independent components, and as a result, the Infomax term of Kan's formulation must be retained.

5.4 Coherent ICA From First Principles

In order to deal with the combined issues of convolutive mixing and the need to reduce the algorithm's dependence on the accuracy of an assumed statistical model, it is helpful to consider the cICA problem as it was originally defined. Reproducing (5.3) below,

$$J_{cICA} = I(\mathbf{x}_a, \mathbf{y}_a) + I(\mathbf{x}_b, \mathbf{y}_b) + \sum_i \lambda_i I(y_{ai}, y_{bi})$$

it can be seen that both of the first two terms concern only adjacent microphone channels. This means that the linear mixing assumption is still at least approximately valid, and that these terms can be replaced with any one of several well-known ICA algorithms.

In the experiments conducted later in this thesis, it was found that the super-Gaussian forms of these algorithms were valid for typical cocktail-party environments containing both speech and music. It was also found that a windowed version of Chicocki and Ubenhauen's algorithm[95] performed best, converging substantially quicker than the natural gradient algorithm[96] or Infomax[88]. The gradient-based nature of [95] also ensures better tracking performance than FastICA[21].

In practical use, it is important to properly initialize the ICA filters in order to achieve the best performance. The initial filters should be chosen to be close to the average desired solution, in order to both minimize the convergence time, as well as to ensure that the ICA algorithm converges to the correct solution. It is a trivial matter to

initialize the ICA filters because the geometry of the problem is well understood; the sources ahead of the listener are considered to correspond to the target, while those emanating from behind the listener are grouped with the interference and should be eliminated. The initial filters should simply reflect this fact, drawing their coefficient values from the known directivity of the microphones being employed, or else from direct experimentation on sample scenarios.

Envelope Correlation

With respect to the problem of convolutive mixing when comparing the outputs of the two microphone groups, it is important to reconsider what information is being compared. In the case of standard ICA, where mutual information is being minimized, or in this case, maximized across channels, the problem of developing a practical coherent ICA algorithm is not an easy one. However, the concepts of mutual information or statistical independence are concerned with high-order statistics in addition to the 1st and 2nd-order statistics used in most classical signal processing algorithms. As the estimation of lower-order statistical information is faster and more robust to noise, limiting the third term of (5.3) to only consider 2nd-order information (correlation), will both simplify the problem, and improve performance. This modification to (5.3) is shown below:

$$J_{clCA} = I(\mathbf{x}_a, \mathbf{y}_a) + I(\mathbf{x}_b, \mathbf{y}_b) + \sum_i \lambda_i E[y_{ai} y_{bi}] \tag{5.9}$$

The resulting formula shown in (5.9) unfortunately still suffers from the problems of convolutive mixing and time-delays discussed earlier, as it uses the raw waveforms. The signal envelope must therefore be substituted in place of the raw signal in order to avoid this problem, because it is relatively robust to noise and reverberation[7].

For the sake of computational simplicity, the signal envelope is approximated in each individual frame as the summation of full-wave rectified elements of that frame. This results in the envelope approximation

$$\tilde{y}_{ai} = \sum_{j=1}^{N} \left| y_{ai,j} \right| \tag{5.10}$$

where for sensor group a the N elements of the frame from the ith input channel are summed after the application of the ICA spatial filters. Applying this to the cost function of (5.9) results in the new cost function

$$J_{clCA} = I(\mathbf{x}_a, \mathbf{y}_a) + I(\mathbf{x}_b, \mathbf{y}_b) + \sum_i \lambda_i E[(\tilde{y}_{ai} - \mu_{ai})(\tilde{y}_{bi} - \mu_{bi})] \tag{5.11}$$

where the envelopes are calculated as above, and the sample means of the windowed and rectified vectors are used as the mean values in the cross-covariance term.

Unfortunately, simply adapting on this cost function does not generally produce good results. The reason for this is that the power of the outputs is unconstrained, which results in a constant growth in the magnitude of the ICA filters. In order to solve this problem, a fourth term can be added to the cost function which penalizes such growth by constraining the output power of the filtered signals to be close to unity:

$$J_p = \left| 1 - \sum_{j=1}^{N} y_{ai,j}^2 \right| \tag{5.12}$$

Where cost function of (5.12) is to be minimized with respect the filter weights. This constraint is similar in concept to the power constraints used in some canonical correlation analysis (CCA) algorithms[97].

The final cost function to be maximized can therefore be written as

$$J_{clCA} = I(\mathbf{x}_a, \mathbf{y}_a) + I(\mathbf{x}_b, \mathbf{y}_b) + \sum_i \lambda_i E[(\tilde{y}_{ai} - \mu_{ai})(\tilde{y}_{bi} - \mu_{bi})] - \gamma \sum_i \left| 1 - \sum_{j=1}^{N} y_{ai,j}^2 \right| \tag{5.13}$$

with the scalar term γ simply representing the weighting of the power constraint. Despite its apparent complexity, the resulting algorithm performs well, and still allows for fast convergence when using gradient ascent. Tests conducted in both low and high reverberation environments with different interferer locations and signal types revealed that the above algorithm's performance was more or less constant over a broad variety of conditions (see Chapter 6 for more details).

Applying this algorithm to the same kind of problem as show in Figures 4.6-4.9, a similar set of plots can be generated. In this case though, instead of an input SNR of 1 dB, the addition of the cICA block allows for much lower input SNRs. The test demonstrated in Figures 5.6-5.9 for example, is carried out with an input SNR of -7.2 dB.

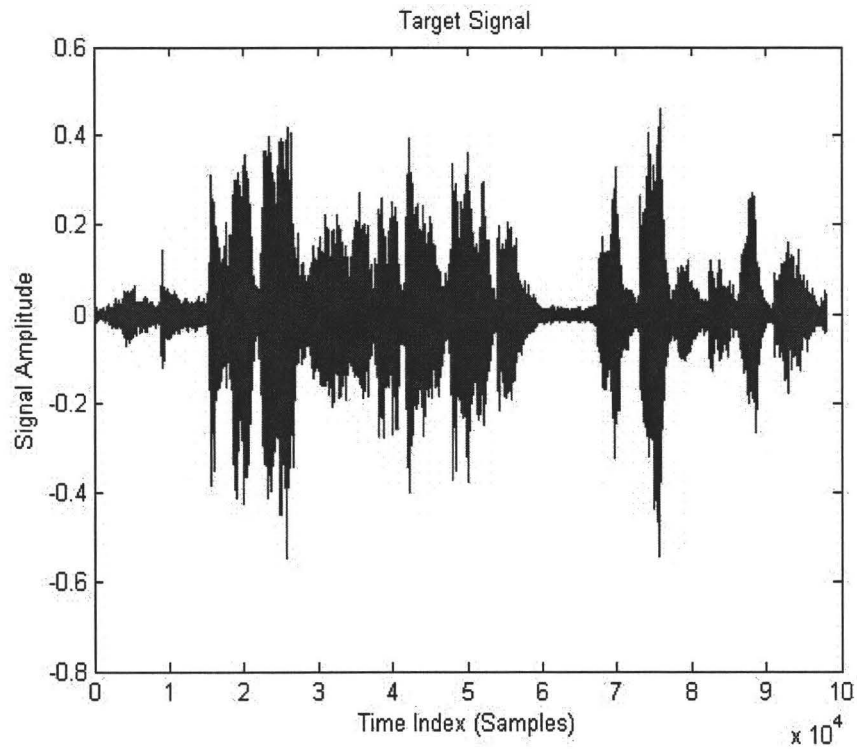

Figure 5.6 A target signal recorded in a reverberant room.

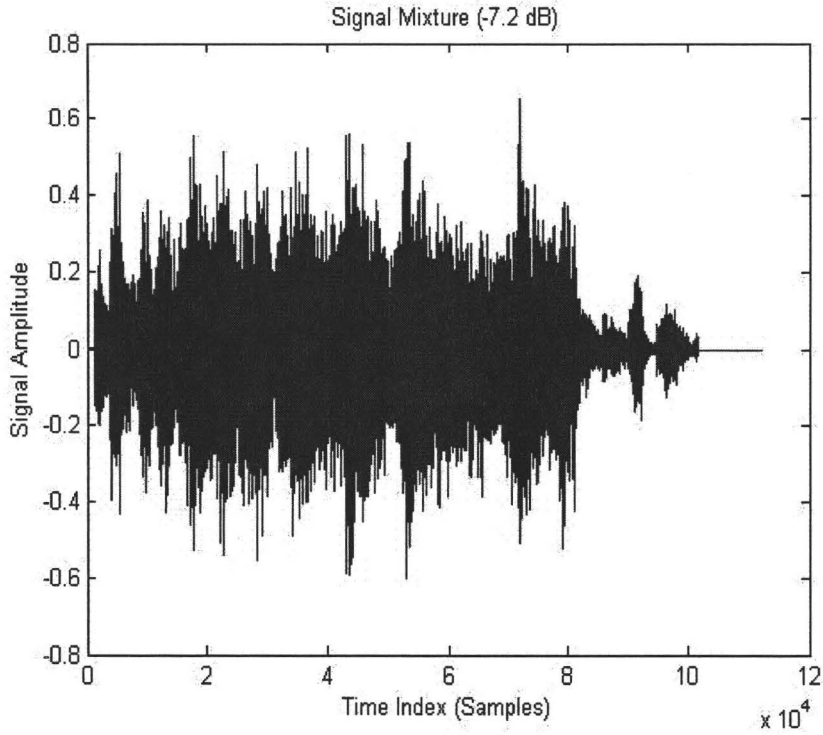

Figure 5.7 The noisy mixture.

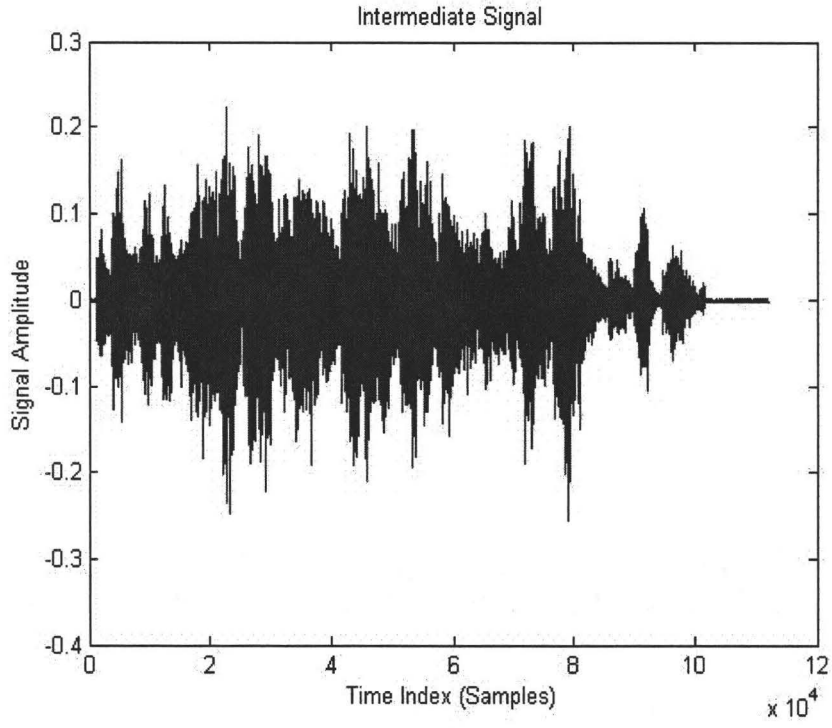

Figure 5.8. The intermediate signal after pre-filtering with cICA, and before applying the FCPP algorithm.

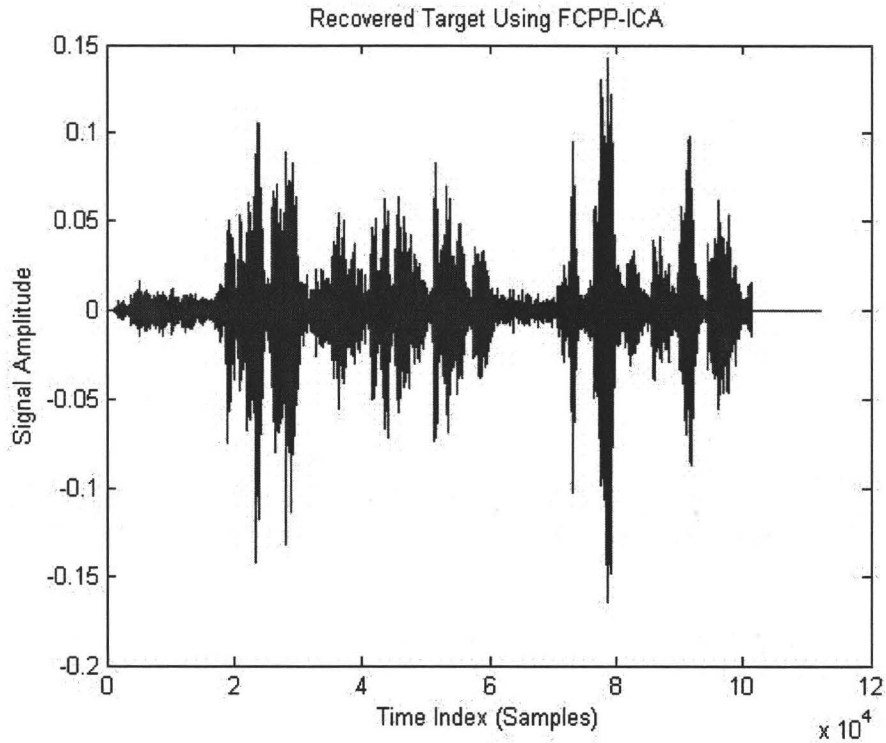

Figure 5.9. The recovered target is a good approximation of the original.

Properties of Microphones

The work presented in this thesis on cICA so far has implicitly assumed the use of ideal microphones. By ideal, it is meant that device properties such as the directivity of the microphones do not change with frequency. In reality, most miniature directional microphones have a directivity index and gain response that is not constant with respect to the frequency[98][99]. For example, in Figure 5.10, the directional response of a single omni-directional microphone is shown in relation to the source frequency. It is important to note that both the microphone and the physical mounting (e.g. a wearer's head) can contribute to variations in directivity with frequency.

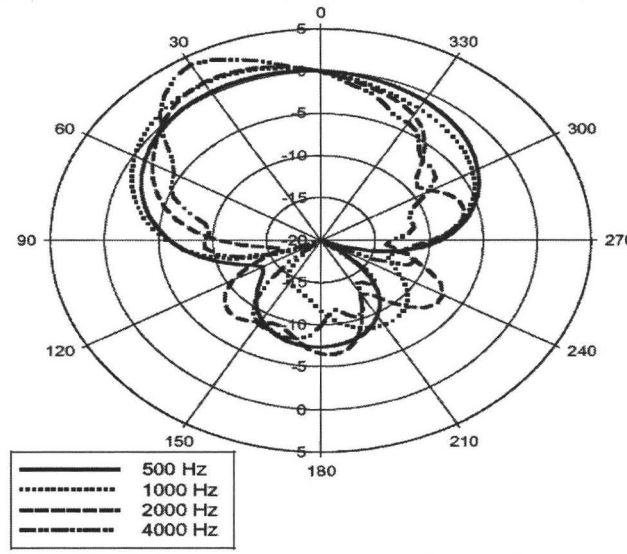

Figure 5.10. The directivity pattern for different frequencies of an ear-mounted omni-directional microphone (Source: Compton-Conley, 2004).

These frequency-based variations can be problematic for the straight time-domain implementation of (5.11). In that case, a single ICA filter is applied across all frequencies based on the assumption that the microphone response is flat. Experiments conducted in the course of developing this thesis, however, reveal that if this assumption is violated, then the time-domain cICA algorithm will diverge. To demonstrate this, a simple simulation was conducted using data collected from the R-HINT-E corpus[100][101]. A simple filtering operation was used to alter the flat-response characteristics of the microphones into a pair of directional microphones whose directivity increases with frequency. Specifically, the base directional gain was assumed to be 1 dB at 100 Hz, and then increased to a maximum directional gain of 4 dB at 1000 Hz. Over several repeated presentations of the same stimulus, it can be seen in Figure 5.11, that the ICA filter slowly diverges

Fixing this problem is simple and follows the well-known work on frequency domain versions of ICA[21][106]. All that needs to be done is for the cost function of

(5.11) to be applied in a channel-wise fashion. That is, an independent set of ICA filters can be applied to each channel or group of channels in order to prevent the filters from diverging during adaptation. The drawback of course is an increase in computational complexity, although this can be minimized by forcing the ICA filters to adapt to a group of channels where the microphone response is known to be similar. The placement and size of such frequency regions will vary between microphones, although in general there is greater variation in the lower frequency ranges than in the higher ones.

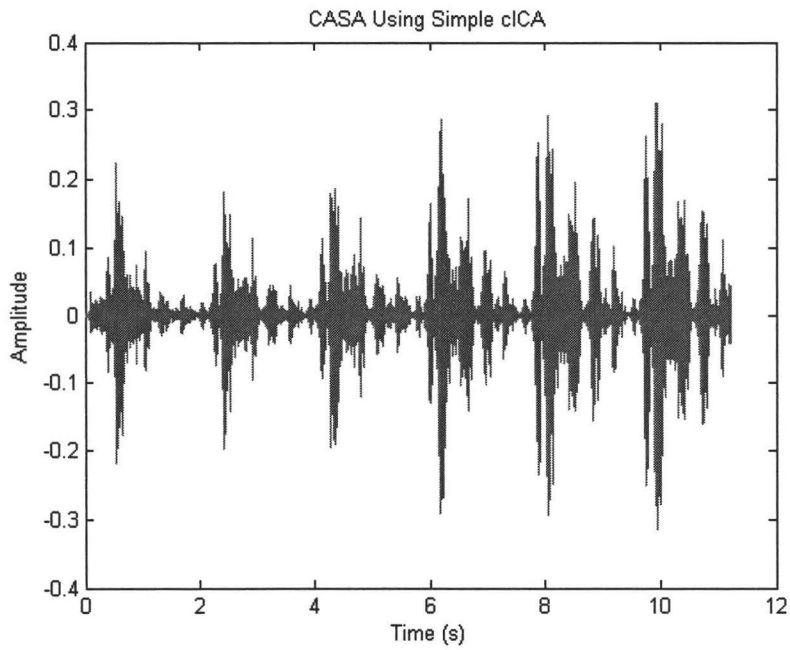

Figure 5.11. With an artificially distorted directivity pattern, the basic clCA algorithm may diverge.

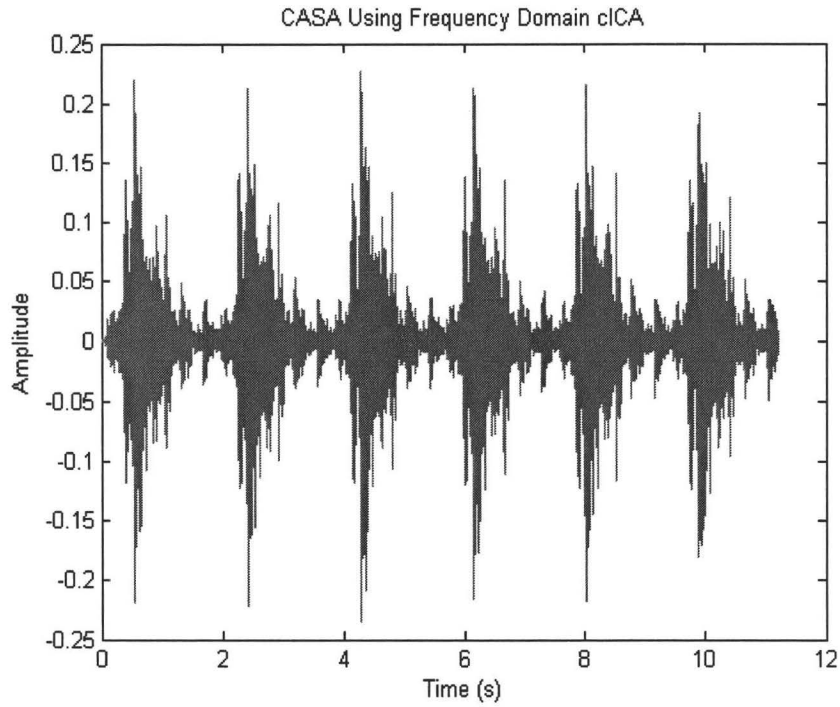

Figure 5.12. Using a frequency-domain implementation of the cICA algorithm can prevent divergence in cases where there is a significant change in directivity with frequency.

5.5 Summary and Conclusions

The problem of enhancing the capability of the fuzzy cocktail party processor depends primarily on reducing the amount of interference present in the signal before presenting it to the FCPP. Ideally, this should be done using some noise-reduction system that operates in a fashion that is complementary to the functioning of the FCPP. This avoids the problem of simply replicating the flaws of one system in the other.

For this purpose, the concept of Coherent Independent Components Analysis, as developed by Kan, has proven to be particularly useful. Building on the idea he developed, and combining it with additional research by authors such as Pedersen[87], a new approach has been developed in this chapter that combines the virtue of "coherency"

in cICA with the practical possibilities afforded by the use of closely spaced directional microphones.

A novel approach to cICA was therefore proposed that also took the realities of the acoustic environment into account. The new system replaces the previous information-theoretic coherency term (based on Imax) with a term that instead incorporates the cross-correlation of the output signal envelopes. The use of the cross-correlation was motivated by the need for robustness in the face of reverberation and time-delays, as was the use of the signal envelopes.

Using this approach, it was shown that a fast and robust form of cICA could be implemented that also performed extremely well in terms of reducing the interference at the inputs. As a result, this new system can be used to extend the operating range of the FCPP by functioning as a pre-filter, as is shown in Figure 5.13.

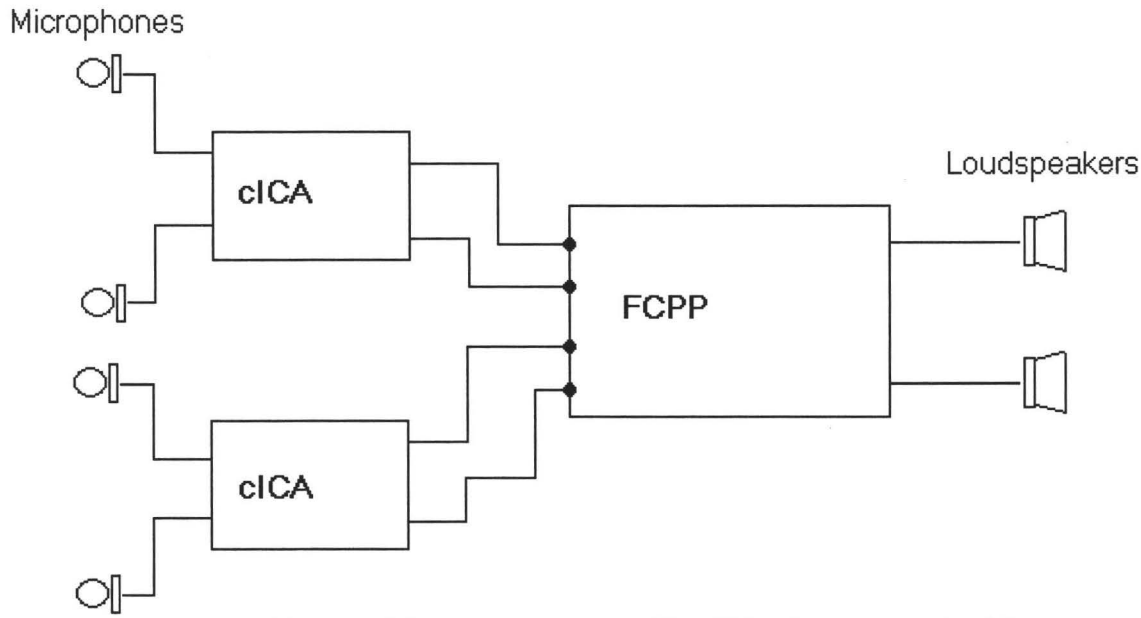

Figure 5.13 Diagram of the combined system. The cICA units on each side of the head (containing the front and back microphones), perform ICA on the two microphones in their respective units, and correlate the envelopes of the outputs.

A refinement of this idea may be needed in order to take into account the properties of real microphones. In such a case, the forumulation of (5.11) and the diagram of Figure 5.13, should be considered as operating in a channel-wise fashion.

In either case, the new formulation of cICA is a valuable addition to the FCPP. As will be shown in the following chapter, cICA can significantly improve the performance of the FCPP by extending the range of noise levels at which it can operate. The cICA enhanced version of the FCPP can therefore operate in much noisier environments than can the FCPP on its own.

Chapter 6

Objective and Subjective Evaluation

6.1 Evaluation

While the algorithms described in the previous chapters do substantially reduce the level of noise and hence enhance the desired target signal, it is important to be able to describe the degree of improvement that is achievable. Such evaluations form the basis for comparative studies between algorithms, as well as help pinpoint possible areas for improvement, or can be used for the purposes of quantitative optimization of system parameters. In the studies presented here, the focus is on establishing both comparisons between pre-existing algorithms, as well as describing the system's performance in the face of changing conditions in the acoustic environment. This will allow us to gauge how well the system behaves against different levels of reverberation, as well against varying levels of interference.

However, there are a large number of possible ways of objectively (i.e. computationally) evaluating the performance of these algorithms. As importantly, the amorphous nature of the concept of "perceptual quality" means that, by and large, these methods are not wholly accurate when it comes to predicting human acceptance of the output quality. Therefore, the more nebulous problem of evaluating the algorithms in

terms of perceptual quality will be reserved for the later section on subjective evaluation methods. This section will instead focus on the problem of evaluating signal fidelity, rather than quality.

For this thesis, the band-averaged SNR is used, in which the quality measure is an average of the signal-to-noise ratios of each individual frequency band $m = 1 \ldots M$. This quantity is in turn averaged over all time windows $n = 1 \ldots N$ for the segment in question, resulting in the following measure:

$$SNR = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} SNR_{nm} \tag{6.1}$$

The use of this measure has the benefit of simplicity as it is easy to compute as well as being intuitively clear in its meaning. In addition, the use of a uniform weighting in the averaging scheme of (6.1) ensures that the quality measure is not tied to any one signal model.

While there are other approaches to the quality estimation, many of which apply directly to speech, it must be remembered that the algorithms described in the preceeding chapters are general in nature; they are meant to enhance any acoustic input signal, not just speech. This is important, because one can expect a variety of possible non-speech target signals in the environment, ranging from music to other environmental sounds that may convey critical information to the user. The quality measure of (6.1) therefore conveys information about the algorithm's ability to extract the desired target sound, but without reference to any one signal model.

Experimental Setup

The experiments conducted in the course of this thesis compared Dong's Perceptual Binaural Speech Enhancement (PBSE), against the performance of the Fuzzy Cocktail Party Processor (FCPP) introduced in chapter 4 and the ICA enhanced FCPP introduced in Chapter 5. Each experiment used one target male speech signal directly in front of the listener, and three interfering talkers (two male and one female) located at 67, 180 and 270 degrees. The SNR was allowed to vary randomly for each trial, as were the start times of each interferer. All of the signals were generated using the R-HINT-E software developed by Karl Wiklund[101], which is a virtual acoustics platform designed to simulate the effects of room reverberation and spatial hearing. The software incorporates a database of pre-recorded HRTFs and room impulse responses which can be combined in order to create a virtual auditory scene.

The directional microphones used in the FCPP and ICA-FCPP were simulated by manipulating the gains of each of the original microphones according to the direction-of-arrival of the incident signal. For the forward-facing microphones this was accomplished using the formula of (6.2)

$$G(\theta) = 0.75 + 0.25 \cdot \cos(\theta) \tag{6.2}$$

This ensured a maximum directivity of 3 dB, which is a fairly conservative approximation of the current capabilities of directional microphones. In order to ensure a reasonable comparison between the algorithms, Dong's basic PBSE algorithm was modified so as to incorporate a pair of directional microphones instead of the original omni-directional microphones specified in [69] and [74]. This allowed for a comparison

of the basic behaviour of the algorithms themselves, while reducing the impact of different hardware choices.

The sampling rate for the initial front-end (common across all algorithms) was set to be 16 kHz, where the processing windows were 256 samples in length and with a 50% overlap. Each of the algorithms decomposed the input signal into 32 separate frequency bands using the gamma-tone filters described in Chapter 3 of this thesis. In order to estimate the SNR, the estimated target signals were also decomposed into 32 separate bands, but using window lengths of 2048 samples and a 50% overlap.

Directional PBSE Results

In order to establish a point of comparison for the other algorithms, the results for the directional PBSE system are presented below.

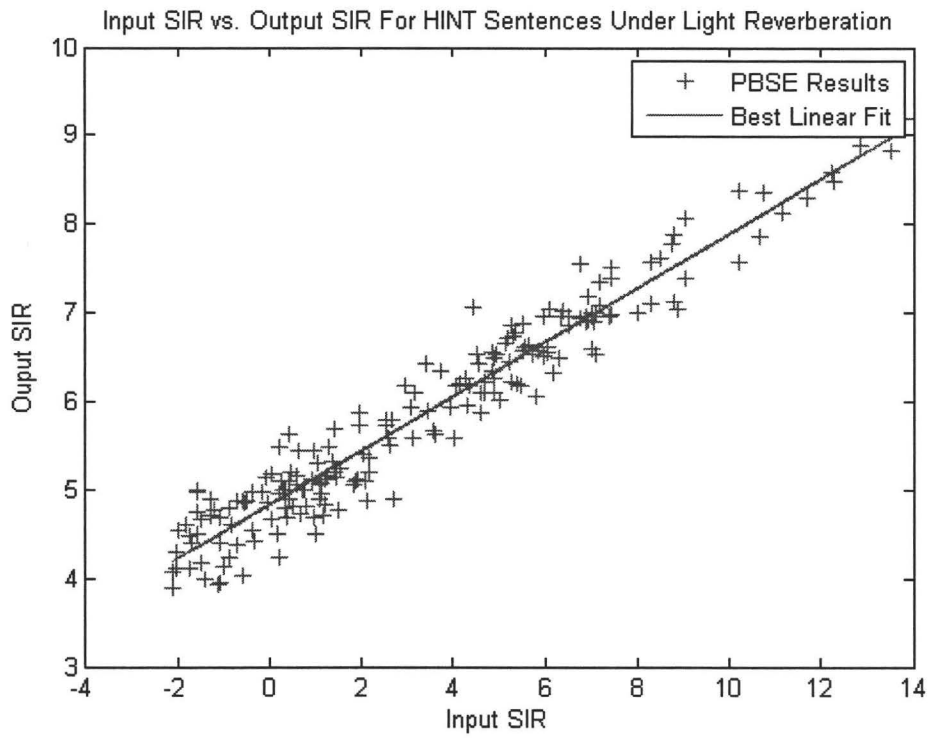

Figure 6.1. Results for the directional version of the PBSE under light reverberation.

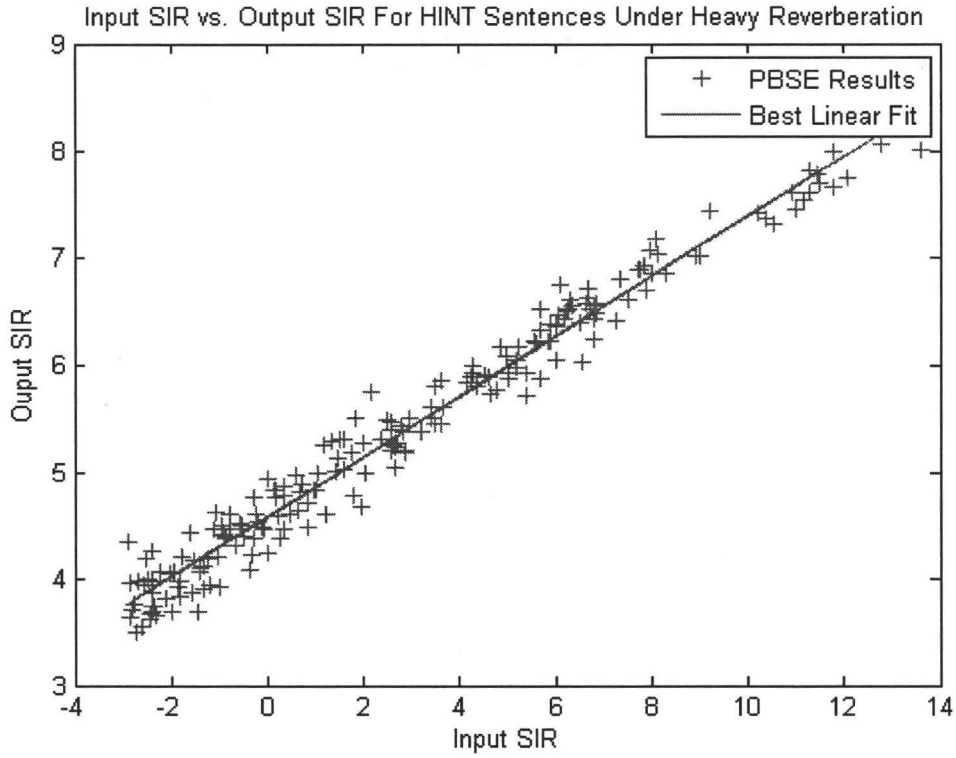

Figure 6.2. Results for the directional version of the PBSE under heavy reverberation.

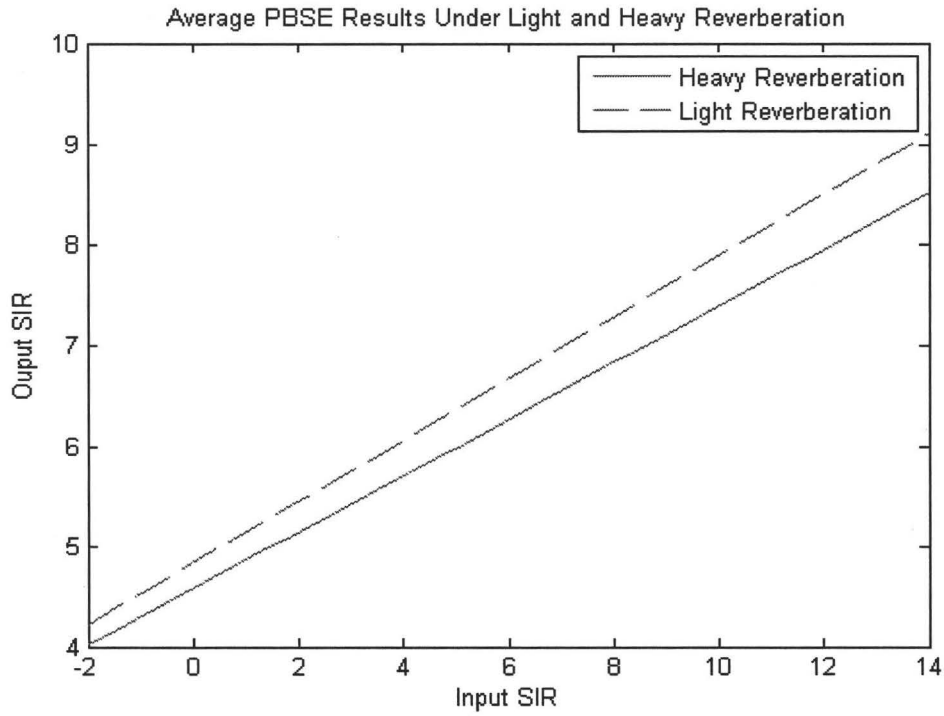

Figure 6.3. A Comparison of the PBSE performance for different environments.

It is clear from Figures 6.1-6.3 that the PBSE does reduce the effects of interference in noisy environments. However, the output signal is still contaminated by some interference at low-to-moderate SIR levels, and the effects of processing artifacts are audible in most samples. It is these artifacts that largely explain the reduction in performance at higher SIR levels, as the PBSE typically does a good job of eliminating interference in this range.

FCPP Results

The FCPP was also tested under the same conditions arranged for the PBSE, and substantially different results were obtained.

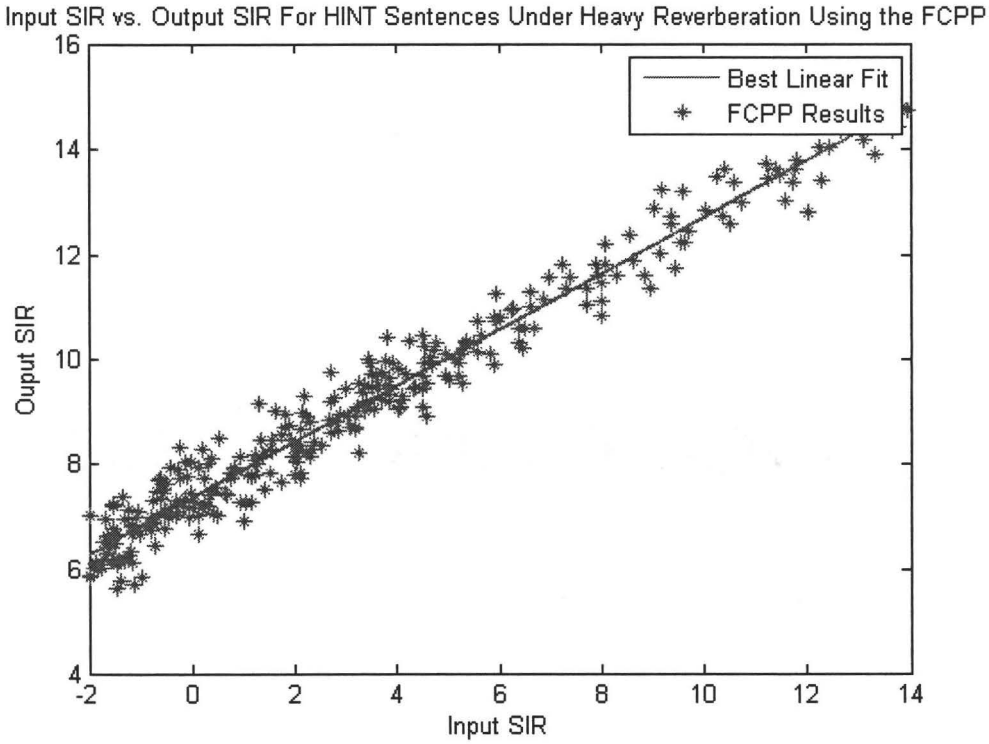

Figure 6.4 Performance of the FCPP under heavy reverberation.

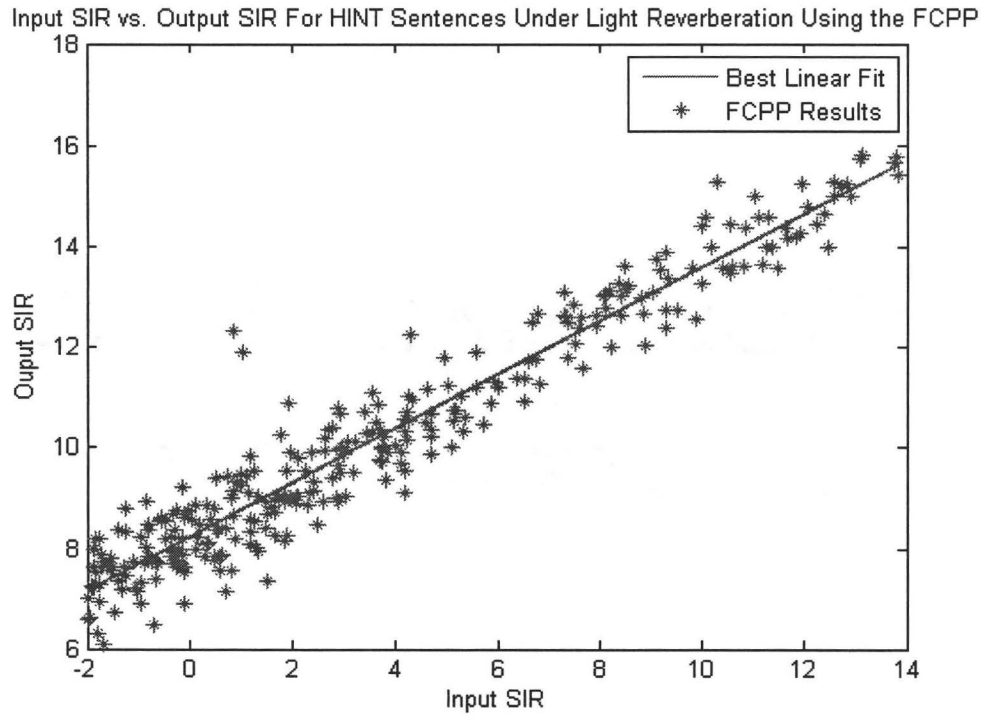

Figure 6.5. Performance of the FCPP under light reverberation.

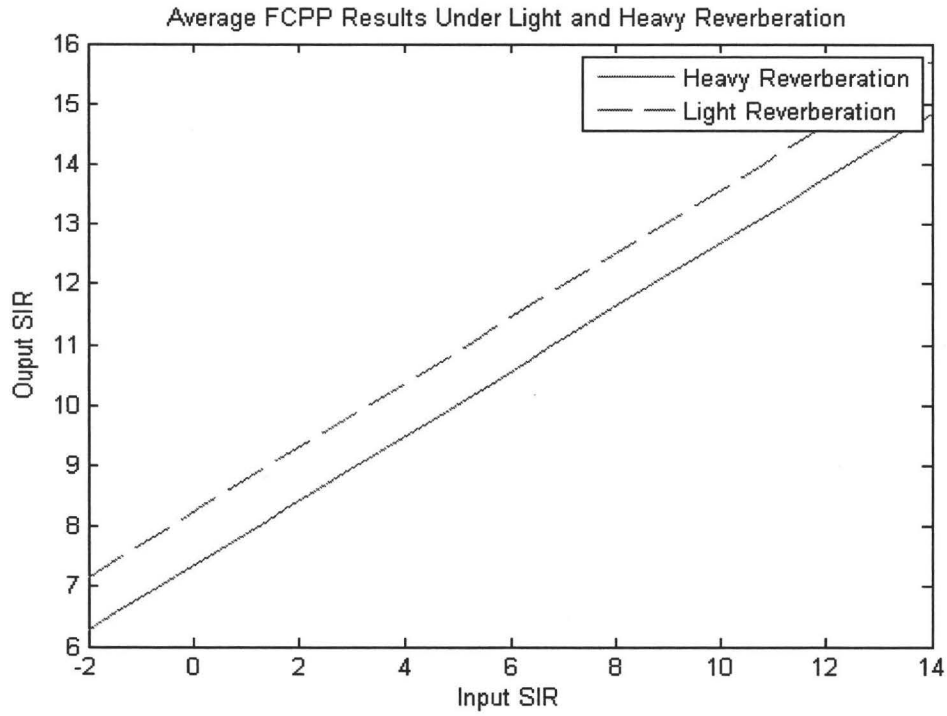

Figure 6.6. A comparison of the average FCPP performance in different environments

Comparison of FCPP and PBSE

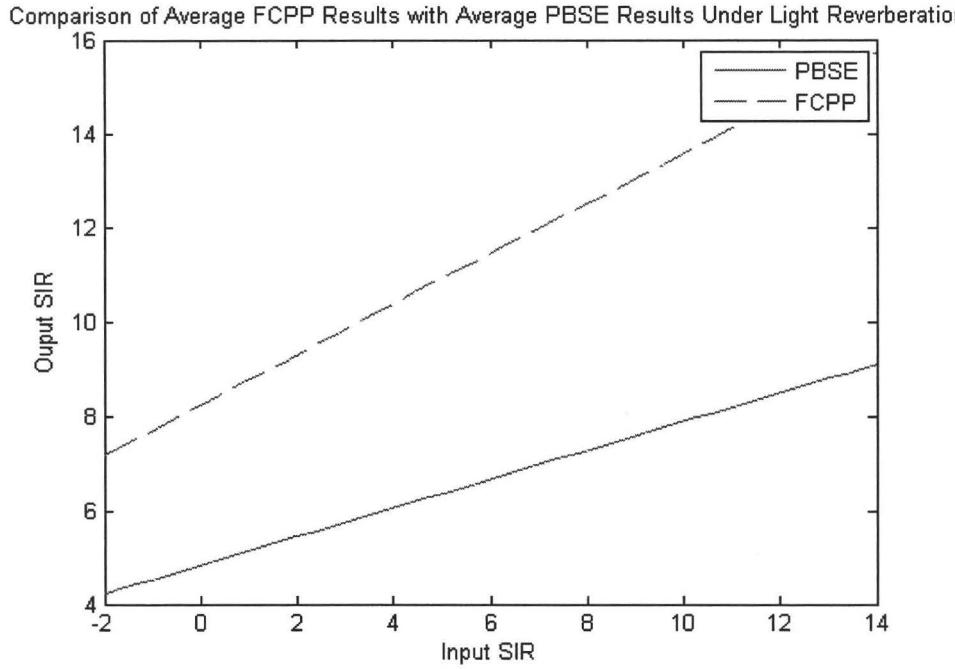

Figure 6.7 A comparison of the FCPP and the PBSE algorithms. For the sake of clarity, only the average performance is shown.

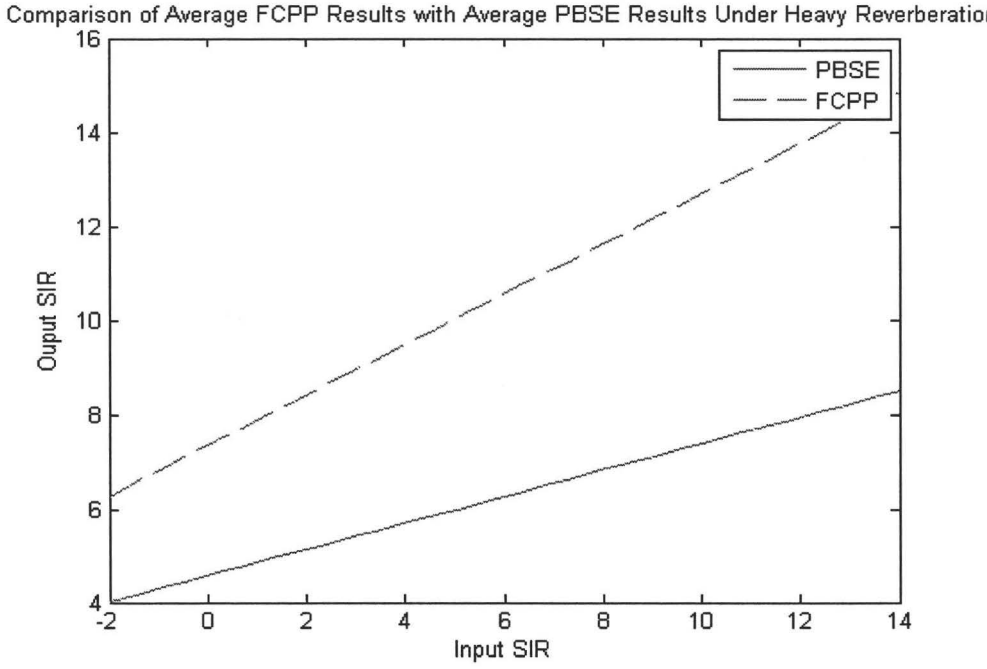

Figure 6.8. The same comparison as in Figure 6.7, but under heavy reverberation. There is a slight reduction in the performance of both algorithms.

FCPP with Coherent ICA

Adding coherent ICA to the algorithm can substantially improve the performance of the system as is shown below in Figure 6.9. This experiment implemented the formulation described in Chapter 5.

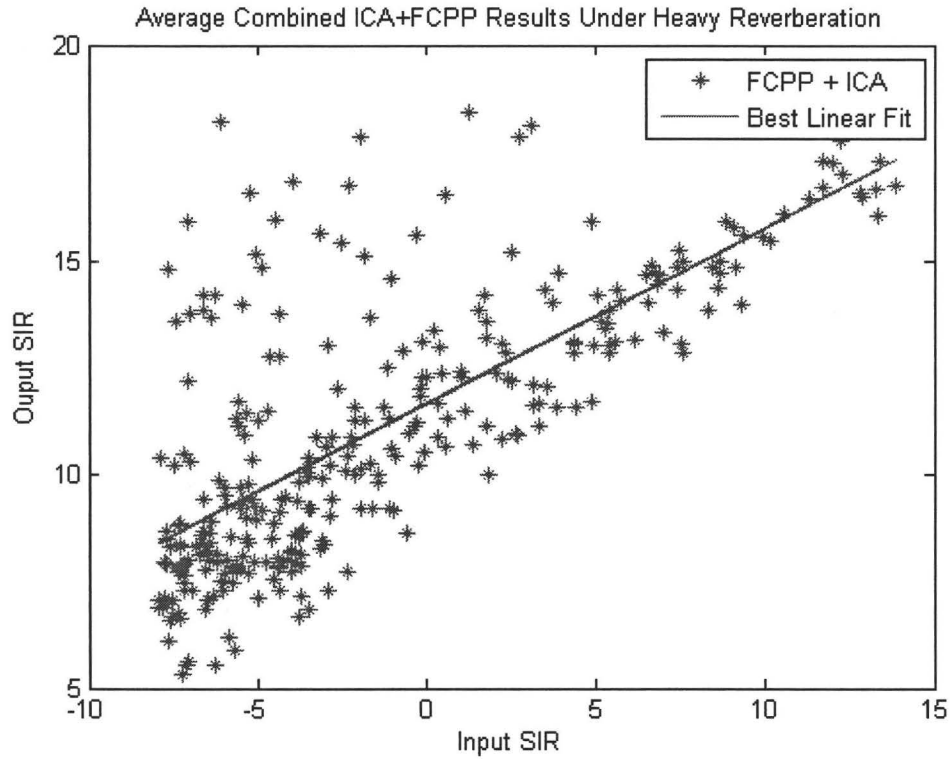

Figure 6.9. ICA+FCPP speech separation results using a common de-mixing matrix for both left and right ears.

It is worth noting that, in contrast to the case where the FCPP operates alone, these results display a greater variation about the line of best fit. It is likely that this arises from the fact that the interferers are positioned differently for each trial. Those trials with a comparatively greater amount of energy concentrated towards the rear of the subject should gain more from the cICA algorithm, than trials with more signal energy present on either side of the

subject. In the former case, there is a greater difference between the signals received at each sensor, which will allow for greater separation. In the latter case, the level of the interference arriving at either sensor will be about the same. Separating such a signal is harder, and may result in that signal being placed in with the "target" group.

Comparisons With Other Methods

Direct comparisons with other methods that have been reported in the literature is difficult undertaking because of the lack of suitable algorithms to compare the new system against. Earlier in this thesis, it was mentioned that of those working on this problem, the overwhelming majority have been developing algorithms based around a set of operational assumptions that are not realistic for the applications considered in this thesis. This deficiency in the literature was also recently mentioned by Wang[112], although without direct reference to any particular solution. Only Watts et al[111] has described a potential real-time solution, although useful comparisons are unavailable for reasons of commercial secrecy.

For these reasons, comparisons across the literature do not feature prominently in this thesis. However, a survey of the reported results is still instructive in that it shows that despite the much more generous assumptions common in this field, the new methods (FCPP and FCPP-ICA) still show superior results. Table 6.1, shown below, summarizes the average SNR gains of the three real-time processing algorithms discussed in this thesis: Dong's PBSE[74], as well the proposed FCPP and FCPP-ICA. These results reflect the algorithm's performance assuming a strongly reverberant room, and an SNR of 0 dB.

Algorithm	Avg. SNR Gain
PBSE	4.5 dB
FCPP	7.5 dB
FCPP-ICA	11 dB

Table 6.1. Comparison of the average performance of three real-time cocktail party processors.

Unfortunately, the data necessary to perform a similar comparison using the "Audience" system developed by Watts et al is unavailable. However, in [111], the authors described their results in terms of subjective scoring using the Mean Opinion Score (MOS)[57] protocol. Using this scheme, a fairly consistent MOS improvement of 0.75 is achieved when the Audience system is used as opposed to when the subjects hear the unprocessed sound. It should also be said that these results did not consider scenarios where the SNR was less than 0 dB.

A very similar protocol to MOS, called the Comparative Mean Opinion Score (CMOS)[57] is used later in this thesis. In this case, the difference in CMOS score between the basic FCPP and the unprocessed sound is ~1.5. Even given the slightly different methodology of the two protocols, the scaling in terms of SNR should be approximately similar. This means that the Audience system probably will not achieve an SNR improvement in excess of 4 dB for cocktail-like environments, at least for the kinds of configurations being considered in this thesis. As the Audience system is primarily concerned with cell-phone applications however, additional advantage may be taken using directional microphones that could improve its performance in such situations.

The following table (Table 6.2) displays the results from three non-real-time algorithms along with their respective operating assumptions:

Algorithm	Avg. SNR Gain
Wu and Wang[68] -non-real time. -anechoic.	1.2 dB
Hu and Wang[33] -non-real time. -anechoic.	4.4 dB
Roman et al[18] -reverberant room. -binaural. -room inverse filter learned in advance.	8.9 dB

Table 6.2. Average SNR gains reported for three non-real-time systems.

6.2 Subjective Evaluation

It has been noted by a number of authors that the signal-to-noise ratio does not completely reflect subjective definitions of speech quality. This is so, because the concept of quality encompasses a number of concepts such as intelligibility, signal distortion, and noise. Algorithm design very often involves trade-offs involving these and other criteria, and not all listeners weight all aspects of speech quality in the same way under all circumstances. For example, it is known that on average, listeners prefer greater levels of background noise over various signal distortion effects[57]. Because of this, and the fact that at present there is no fully agreed upon objective standard of measuring speech quality, it is necessary for studies of this subject to also include subjective, human-based trials in order to gauge the opinions of a typical user.

A simple and reliable approach to subjective testing of speech quality is the Comparison Mean Opinion Score, or CMOS test[57]. This test compares users' comparative ranking of two speech samples. The two samples use a common source signal, but are presented to the subject after processing by different algorithms. In the basic form of this test, the decision is a binary one; that is, the user is simply asked to express a preference for one sample over the other. For a more detailed analysis, especially where several algorithms are being compared to each other simultaneously, a seven level ranking system is also used, which is shown in Table 6.3. This scheme allows not only for an evaluation of whether a new processing scheme is better than another, but also by how much.

Voting	**Score**
A is much worse than B.	-3
A is worse than B.	-2
A is slightly worse than B.	-1
A and B are about the same.	0
A is slightly better than B.	1
A is better than B.	2
A is much better than B.	3

Table 6.3. Listening scale of the 7-level CMOS test.

Subjective Testing Procedure

To test the responses of human listeners to the output quality of the new algorithms described in this thesis, an implementation of the CMOS test was developed after consultation with Professors Ian Bruce and Laurel Trainor. The testing procedure that was created relies on MATLAB-based simulations both for the compilation of testing data, as well as the actual tests themselves.

Using the measured impulse responses from the R-HINT-E room simulator, 120 different speech scenarios were created using both HINT sentences, speech samples from film and television (Appendix A), as well as a number of musical pieces of varying styles (Appendix B). Each scenario was generated using a randomly chosen target speech sample, and four randomly placed interfering signals (which were also chosen randomly). The same scenario was generated under conditions of both light and heavy reverberation. Each scenario was processed by the various algorithms under test, and saved for later presentation to the subjects.

The actual test used a specially written MATLAB GUI-based program in order to present a double-blind list of examples to the subject, which is shown in Figure 6.10. For each trial, 80 of the 120 samples were randomly chosen for presentation, and then shuffled so that the list would be presented in random order. The user could play either sound by selecting the appropriate button (either A or B), and could do so as many times as he or she wished before making a decision about how to rank the comparison. Which processing algorithm corresponded to Button A, and which to Button B was also randomly assigned with each presentation in order to prevent any systematic bias in the experiment.

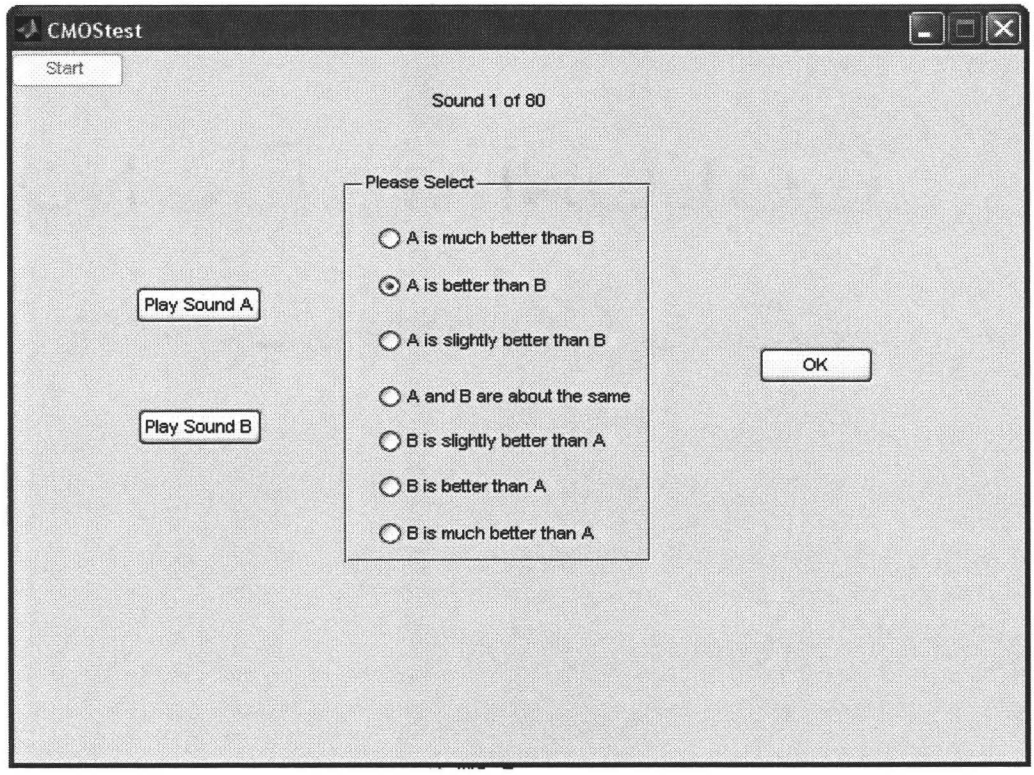

Figure 6.10. The user interface for the CMOS audio test.

For the human trials, ten volunteer subjects were used who were then broken into two groups in order to ensure that a sufficient range of possible performance comparisons was made. Six people (3 females and 3 males) were chosen for the first group, and four (2 males and 2 females) made up the second group. The comparison types used for each group are shown tin Table 6.3. Because of their importance, comparison types (1) and (4) are common to both groups of subjects.

Group 1	**Group 2**
1. FCPP vs. PBSE	1. FCPP vs. PBSE
2. FCPP vs. Unprocessed	4. FCPP + ICA vs. FCPP
3. PBSE vs. Unprocessed	5. FCPP + ICA vs. Unprocessed
4. FCPP + ICA vs. FCPP	6. FCPP + ICA vs. PBSE

Table 6.3. A listing of the comparison types made during the human trials.

In the following plots, the CMOS results of each group are shown in terms of the average score assigned to each comparison type (refer to Table 6.3 for a listing of the scores). These results are then further broken down in terms of the responses of the individual subjects.

CMOS Results

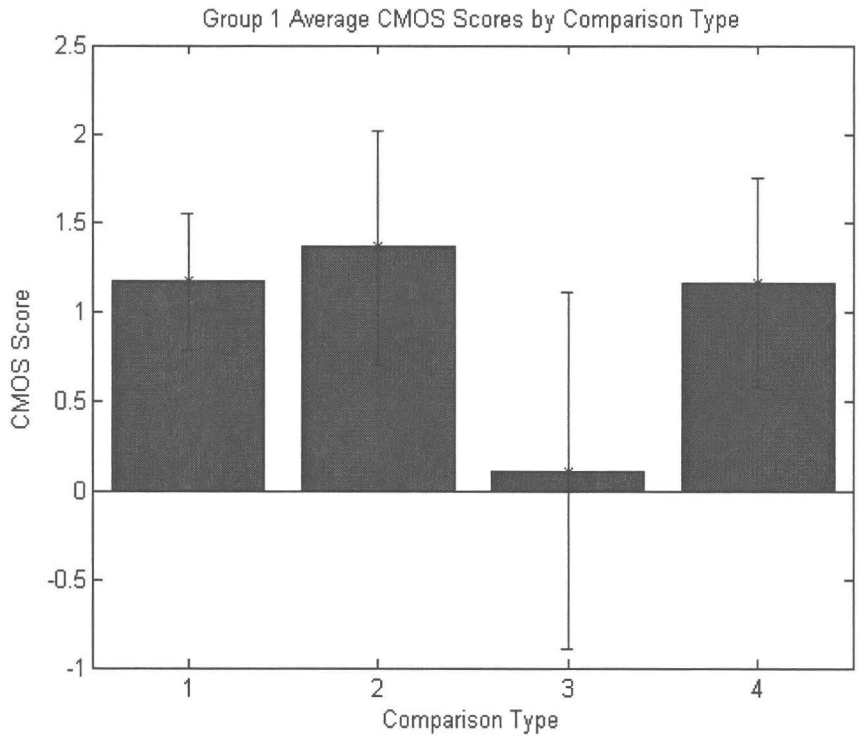

Figure 6.11 Group 1 average CMOS scores. The size of the error bar is equal to the standard deviation of the responses for that particular category.

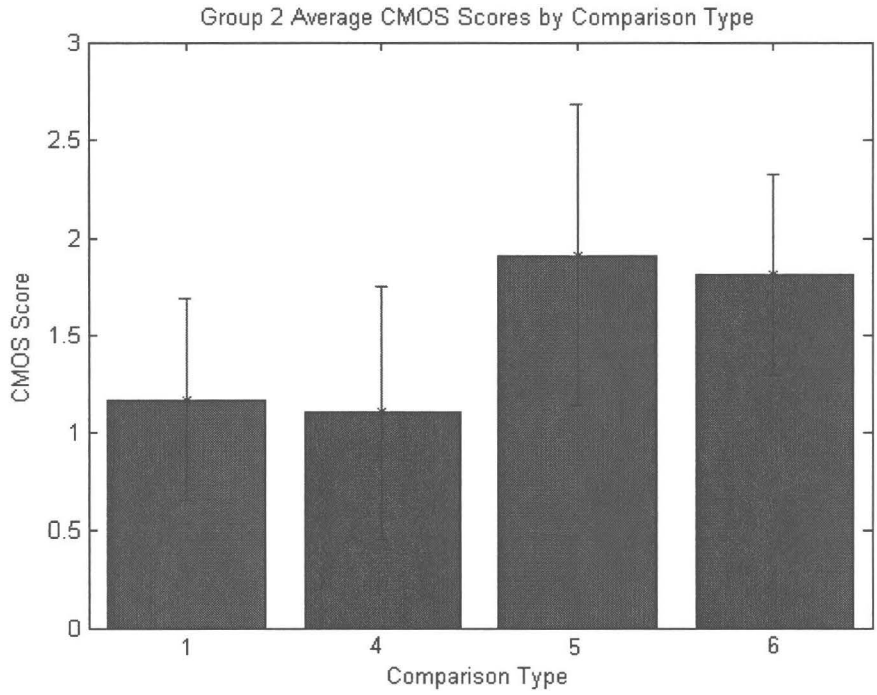

Figure 6.12 Group 2 average CMOS scores.

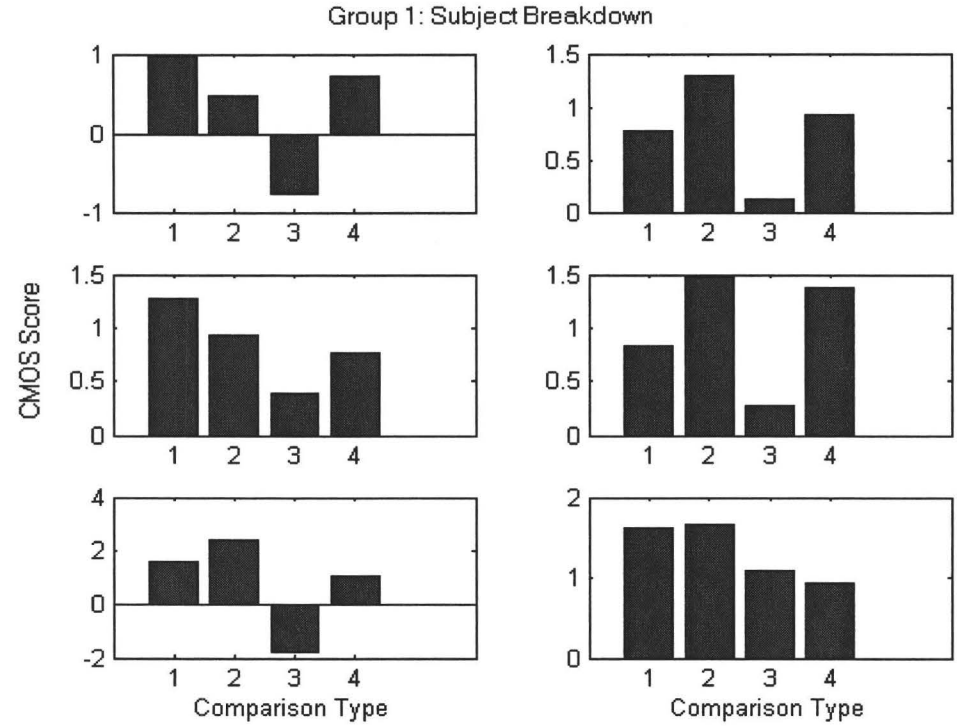

Figure 6.13. Group 1 results by subject.

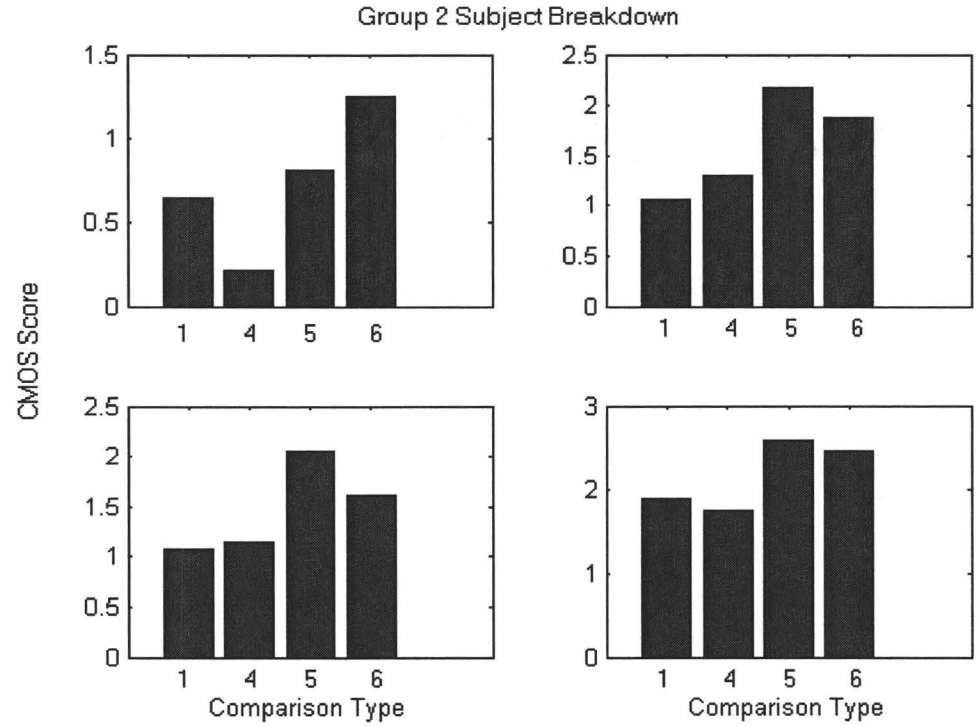

Figure 6.14. Group 2 results by subject.

Type I Comparison: FCPP vs. PBSE

This comparison was common throughout both groups of subjects, and in both groups, there was the consistent judgment that the FCPP ranked between "slightly better" and "better" than the modified PBSE algorithm. In terms of the actual CMOS rating, this translated into both groups giving an average opinion score of 1.2. As this result is essentially identical for both sets of subjects, it is reasonable to believe that this result is fairly representative of the algorithm's subjective performance given a mostly non-hearing impaired group of subjects. Subjects with hearing impairment should be considered separately, and it is uncertain at this time what their responses would be.

It should be noted that the first subject of group 2, while still ranking the FCPP better on average that the PBSE, gave significantly worse grades than the other subjects (a fact that carried on throughout the other comparisons made by this individual as well). It is possible that this can be attributed to the subject's being well outside of the typical age range of the other subjects, as he is a 72-year old male with some hearing loss. However, as this study also included a 50-year old female with some hearing loss, who scored this category substantially higher than the average, it is not clear if that is a sound conclusion.

Type II Comparison: FCPP vs. Unprocessed Sound

This comparison establishes an important baseline for the trial in that it tells us whether or not the new algorithm is really an improvement over doing nothing at all. This is especially important given a subject's opinions on the subjective speech quality depend both on the amount of noise, and on the amount of signal distortion, with the

latter being dominant. Therefore it is important to determine whether the quality of the output speech is acceptable, even when given a significant level of noise reduction.

From the actual tests, it was found that the average opinions on the output quality of the FCPP algorithm lie roughly mid-way between the "slightly better" and "better" categories. In terms of the CMOS score, the average rating is about 1.4. Although only the first group of subjects was tested on this comparison, their responses were consistent in their opinion that the FCPP improves the quality of the output sound.

Type III Comparison: PBSE vs. Unprocessed Sound

The rationale for this test was much the same as the first, although including both of these tests allowed for an extra point of comparison between the two systems. In this case, there is slightly more diversity in opinion regarding whether or not it is an improvement over the unprocessed condition. Two of the respondents identified the results as being worse than no processing, while three others concluded that the algorithm was only marginally better on average than the alternative. Only one subject out of the six ranked the PBSE as being "slightly better" on average.

Over all the relevant subjects, the average CMOS ranking for this comparison is only slightly above zero. As noted above, there was a significant diversity of opinions from the subjects. It is likely that this diversity might spring from differing definitions of "quality" on the part of the subjects. As can be seen from the plot in Figure 6.3, the PBSE does improve the output SNR by removing much of the interference. However, there are also very noticeable distortion artifacts present, especially in situations with high noise or reverberation levels. Different subjects may be emphasizing either the

noise reduction or the distortion effects in their assessment of speech quality, resulting in a disparity of opinions.

Type IV Comparison: ICA + FCPP vs. FCPP

In these trials, which were also common to both groups, the two new algorithms introduced in this thesis were compared to each other. The results showed that in general, the subjects ranked the ICA-enhanced FCPP as being "slightly better" than the basic FCPP algorithm. Again, there is a significant range of opinion on this comparison, as can be seen from the plots shown earlier.

In this case, both the FCPP and its ICA-enhanced counterpart are able to effect good interference rejection with very little speech distortion. Under the environmental conditions for these trials, the FCPP-ICA does have an edge in interference reduction, although in both cases, the target speech can be heard clearly. However, this test only includes the range of input SNRs for which the basic FCPP algorithm is capable of adequate operation (down to about -1 dB). The primary purpose of the FCPP-ICA algorithm is to operate under conditions beyond the range of the basic FCPP. From the subjective tests shown in Figure 6.9, it is clear that it does so, although it is difficult to make useful performance comparisons given the lack of algorithms to compare it with.

Type V Comparison: ICA + FCPP vs. Unprocessed Sound

The results for these tests indicate that the respondents ranked this case fairly highly in favour of the ICA-enhanced FCPP, with an average CMOS score of 1.9. These results are consistent with the earlier comparisons of the FCPP vs. the unprocessed

sound, and the FCPP vs. the ICA-enhanced FCPP. In this case, the respective CMOS scores were close to 1, so it is not surprising that the overall comparison of FCPP-ICA vs. unprocessed sound should be about double that.

Type VI Comparison: ICA + FCPP vs. PBSE

As with the previous comparison, these results strongly favour the ICA-enhanced FCPP. On average the subjects consistently ranked the new algorithm as being "better" than the PBSE, which translates to a CMOS score of about 1.8.

6.3 Discussion of Results

From the results of both the objective and subjective tests, it is clear that the FCPP algorithm is capable of producing significant gains in the quality of filtered speech. In the objective trials, which were carried out over many trials, and a range of SNRs, the results are clearly visible. In objective terms, the SNR improvement is about 7-8 dB depending on the specific characteristics of the acoustic environment. In addition, the strong correlation of the output SNR with the input SNR means that the approximate performance of the FCPP can be readily predicted for a given level of background interference.

Subjective trials consistently show that subjects find that using the FCPP to reduce interference produces a significant improvement in the quality of the resulting speech. When compared to the unfiltered state, signals filtered using the FCPP are consistently ranked high in terms of quality. As these results were obtained from large numbers of tests carried out over many types of acoustic conditions and interference

signals, it is reasonable to believe that they reflect the potential real-world performance of the FCPP.

The basic FCPP algorithm can also be extended through the addition of a cICA-based adaptive pre-filter. As can be seen in Figure 6.9, the addition of the cICA component has fulfilled its primary purpose, which was to extend the ability of the FCPP to operate in very noisy environments. The use of this extra filter noticeably improves speech quality, resulting in average SNR gains of about 3dB over the FCPP. However, the broad variance of the FCPP-ICA results mean that that figure is only approximate. It is not clear at this time how to account for this increase in variance over the much more tightly constrained results produced by the FCPP alone.

Consistent with these results, the subjective trials also show that the use of the ICA-enhanced FCPP results in audible performance improvements over using the FCPP by itself. Likewise, subjects found there to be a dramatic difference between the FCPP-ICA and the unprocessed state. Because these results are so consistent across all of the subjects, and because they represent many different random trials, it is possible to be confident of their ability to represent the system's performance in realistic environments.

The FCPP and its ICA-enhanced variant are both evolutionary improvements to Dong's PBSE algorithm. While using similar design concepts, the FCPP was developed in order to address specific weaknesses of the original work. Both objective and subjective comparisons of the FCPP and PBSE, indicate that this research plan has been successful. These results demonstrate that with respect to the metrics discussed here, the FCPP does indeed outperform the PBSE.

Of particular interest is the fact that the average CMOS quality rating for the PBSE vs. the unprocessed state is close to zero, indicating little perceived improvement. Based on the subjective results, it seems that this score arises from the fact that distortion artifacts are prominent in PBSE-processed sound, despite the reduced interference levels. Such distortion is greatly reduced in the case of the FCPP.

However, when considering these results, two issues should be kept in mind. The first, which to some extent has already been discussed, is the fact that different people will differ on how they rate the quality of the sound, based on personal tolerances to noise, distortion, acuteness of hearing, etc. This is notion is readily illustrated in Figure 6.13, where the subjects are asked to rate the quality of the PBSE output vs. the unfiltered input. It is also demonstrated though the scores of the two hearing-impaired subjects, who gave two markedly different sets of quality scores.

In addition, it should be noted that the SNR is an imperfect measure of speech quality. It was chosen for use in this thesis primarily because of its simplicity, as well as the fact that as a metric, it is one that is immediately familiar to engineers. However, as an objective measure, it can only be considered an approximate guide.

6.4 Conclusions

In both objective and subjective evaluations, the Fuzzy Cocktail Party Processor (FCPP) consistently outperformed its predecessor, the Perceptual Binaural Speech Enhancer (PBSE). A further improvement was made with the ICA-enhanced FCPP, although at a greater computational cost. These results however, only apply to MATLAB-based implementations of these algorithms. As of yet, neither the FCPP, nor

its enhanced version have been implemented on a deployable digital-signal processing platform. In addition, while fewer assumptions about the environment and processing capabilities have been made than elsewhere in the CASA literature, the fact remains that none of these algorithms have been instantiated in hardware or tested in real-time environments using actual devices. It is not yet wholly clear how well these new algorithms will perform in real environments.

In spite of this, it is possible to be confident that a clear way forward towards a deployable product has in fact been shown. In all cases, real room impulse responses have been used, which correctly reflect the spatial nature of sound. Similarly, the computational burden of the new algorithms is not unrealistic for digital signal processors currently being sold or for those that may become available in the near future.

In both its results, and in its formulation, the FCPP is a significant step forwards and is a meaningful contribution to the literature on computational auditory scene analysis. The unique approach to data fusion involving fuzzy logic is especially novel in that unlike many other CASA algorithms, it makes the unreliability of the acoustic cues a central feature in the design. Doing so better represents the environment, and ultimately results in better performance. Further, improvements to the control and adaptation mechanisms also include concepts that have not before been implemented, either in CASA systems in general, or in CASA systems intended for real-time operation.

A further contribution has been realized through the use of the coherent ICA algorithm as a pre-filter prior to the application of the FCPP. While this work has drawn from other discoveries, it is unique in both its application and its formulation. Most importantly, the contribution of this work is significant because of the results that have

been so far obtained. The combined FCPP-ICA has been shown to be extremely effective in combating spatial interference, even when operating in particularly adverse conditions.

Chapter 7

Summary and Future Work

7.1 Summary

This thesis has addressed the so-called "machine cocktail-party problem", and has presented a solution that is appropriate for certain types of wearable devices. Throughout this work, the central goal has been to address practical solutions that focus on resolving the problems associated with real-time operation, and with the inherent ambiguity of the acoustic cues in the presence of noise and reverberation. This stands in contrast to much of the other work in the literature, which has focused on non-real-time processing, and on certain very constrained problems that are not appropriate when considering the operation of more general-purpose device.

By and large, this thesis has been successful in the goals outlined in the first chapter. In particular, this thesis has made several key contributions to the field, which include:

1) A hierarchical grouping of acoustic cues based on robustness The identity of the segments that have been grouped are then constrained based on the average behaviour of the less reliable cues.

2) A novel approach to decision and data fusion rules that is rooted in fuzzy logic. This has allowed for a system that is substantially more robust to reverberation, as well as a reduction in musical noise.

3) A novel approach to reducing the problems associated with front-back ambiguity has been proposed. This new approach combines the use of directional microphones with the well-known spectral subtraction algorithm.

4) A new SNR adaptive control mechanism was introduced in order to improve the perceptual performance in especially difficult environments.

5) A novel approach to mitigating front-back confusion using coherent independent component analysis was proposed. This system has also substantially increased the system performance.

6) The method of coherent independent components analysis has been applied in a novel architecture that significantly reduces acoustic interference.

7) Of particular importance and novelty is that in contrast to other approaches, this method is meant to work in an on-line (and real-time) fashion.

The ideas that were presented in this thesis have also been tested using both objective and subjective metrics. It was found that the new algorithms offered a substantial gain in output SNR for cocktail-like scenarios, which included spatially distributed noise sources, and realistic reverberation effects. At low SNRs, the new algorithms offered an SNR improvement of around 6 dB in the case of the FCPP, and an 8-10 dB improvement for the ICA-enhanced version. These results compare favourably with the PBSE.

In addition to the use of the objective metric discussed above, human trials were also conducted in order to ensure that the processed speech offered perceptual improvements over the other options. These trials show that the subjects are consistent in their opinions that the new methods do indeed improve the output speech quality in realistic environments.

7.2 Future Work

What has been described in this thesis has credibly demonstrated the feasibility of using CASA methods to solve the cocktail-party problem for real-

time applications. However, there remain areas for possible future improvement. These include:

1) Temporal fusion rules. The rules that have currently been implemented are very simple and do not make good use of the information from the previous frames. Using this information may allow for more reliable mask estimation

2) Spectral Subtraction and noise-level estimation. Currently, only very simple approaches to these problems have been considered in deference to the over-riding need for minimizing the computational complexity. However, it is not clear if there are better methods to use than what has been implemented. Further exploration of this area could yield some needed improvements.

3) New configurations and hybrid systems. This thesis only considered a binaural arrangement of sensors. Future work may include other sensor geometries, or integration with other sensor systems and signal processing strategies.

4) The work in this thesis has been fairly general in terms of its reference to the applications. Moving to more specific applications may require further research in terms of feasibility and in terms of configuration.

Reducing Computational Complexity

Of particular concern throughout this thesis has been the question of managing the computational complexity of the system so as to be able to develop a feasible implementation. As it stands now, the algorithm that has been described is able to meet the stated real-time goals for the current generation of high-performance DSPs. However, in many applications, the power consumption of these devices is larger than is desirable, meaning that a less computationally demanding implementation should be sought.

To that end, several avenues of investigation can be pursued. In particular, given that the lion's share of the system's computational resources are devoted to the earliest stages of processing, meaning the cochlear filtering and FFT-based cue computation, rather than the actual data fusion and mask estimation components, it stands to reason that more effort should be focused on improving the former components instead of the latter. For such work, several possibilities present themselves to the engineer:

1)Replace the cochlear filters with less computationally demanding filters that possess similar properties. This is currently an active area of research[107], in which technical performance criteria such as delay and complexity must be balanced against the resulting perceptual qualities[24].

2)Subsampled cochlear filter banks. Most filterbank structures make use of some level of subsampling, although this is not the case with the gammatone filterbanks used in this thesis. Recently however, it was proposed that at least some level of subsampling is possible for at least the lower frequency ranges[108]. This subsampling strategy could be used to reduce the length of some of the FFTs, and so result in some computational savings.

3)Use of non-radix 2 FFTs. Depending on the window length that is being used, a radix-4, or radix-8 transform may be preferable[109]. Typical savings, given equal length FFTs, are on the order of about 10-15%.

4)Dedicated hardware acceleration. Both the gammatone filters and the FFTs are well-understood algorithms that are readily implementable in hardware[110]. Using such structures would reduce the burden on the processor core, as well as open up greater possibilities for parallel processing. Work by Watts[111] has very recently demonstrated the feasibility of this approach, as well as the significant savings that are achievable. As a result, this approach should be given particular priority.

Future Research Directions: Features of Features

Some suggestion has been made that an implementation of the FCPP incorporating a "features of features" approach may result in better performance. Such an idea is workable in that the derivation of higher-order features can increase the sparsity of the representation, and so therefore ease the task of separation. It is also well-founded in the neuroscience literature, as it follows naturally from the idea of spectro-temporal receptive fields[50][28]. Therefore, interposing a second layer of filters that extract higher-order features between the gammatone filter bank and the separation/grouping algorithm, is an approach that is worth investigating.

The primary value of this approach will be in regards to either neurobiological modelling, or off-line separation algorithms. The additional latency added by the second

layer of filters, as well the increased processing and memory requirements pose a serious obstacle to developing real-time and embedded applications based on this concept. In addition, it is not yet clear how to best handle the problem of signal reconstruction for such a dual-layer filtering system.

New Applications: The Smart Helmet

The rule-based FCPP system has so far been shown to be an effective solution to the problem of binaural speech enhancement. However, given its ability to perform high quality speech enhancement with very few sensors and a minimum of computational power, the basic concepts of the FCPP are also potentially applicable to a wide range of applications relating to noise control, surveillance and other problems in acoustic signal processing. In particular, we were recently approached by the Department of National Defence who are interested in applying the concepts behind the FCPP in order to develop noise control solutions for soldiers in the field.

The goal of this project would be to develop a helmet-mounted acoustic signal processing system. This system is to be capable of passing/enhancing speech signals of interest, reducing low-frequency noise (engines, etc) and passing attenuated high frequency transients that take the form of impulsive noise (such as gunshots). In addition, the helmet is to estimate the direction of arrival of such high frequency transients in order to assist the wearer in identifying potential threats. As this investigation is only its earliest and most speculative phase, it is difficult to consider what its precise configuration will be. However, it is already believed that it will combine elements of our fuzzy rule-based processor as well as advanced techniques in array processing.

Current Implementations

A critical future phase in the development of these ideas is to implement them on a standard DSP platform, such as those provided by Texas Instruments or Analog Devices. While it is not completely known whether or not the current generation of processors handle the demands of the algorithms described in this thesis, the examples of previous authors are encouraging. Dong, whose work laid so much of the foundation for this current project, demonstrated that her CASA-system could be implemented in real-time on the TMS 320C6713 processor. Because the new system is only somewhat more computationally intensive that its predecessor, it is reasonable to expect that it can be implemented with a minimum of modifications.

In addition, other CASA systems that have been demonstrated in the literature bear out this optimism. Pedersen's work[87], for example, has been demonstrated in hardware using an acoustic mannequin, which has borne out his contentions on the use of closely spaced directional microphones for hybrid ICA-CASA applications. Pedersen's work did not entail demonstrating the practicality of processing, although it did effectively demonstrate the proposed sensor apparatus. Similar work by Mori et al[102], have also demonstrated the feasibility of hybrid ICA-CASA system, via a real-time implementation of a two-stage speech separation model in which ICA forms the front end to a binary mask estimation algorithm.

While neither of these systems involve all of the complexity of the work completed in this thesis, they do demonstrate that the underlying concepts that have been discussed are indeed feasible. To this end, future work on this project will focus on implementing the FCPP and the ICA-FCPP in hardware. With respect to this challenge

some important hurdles will have to be overcome. The most important of these is the fact that unlike all of the systems mentioned previously, the new algorithms make use of more than two data streams. This adds significantly to the computational burden, and may pose a challenge in terms of the number of available input/output ports for the microprocessor. A further problem is the fact that the capabilities of wearable directional microphones have not been fully addressed, especially in relation to how such capabilities will impact the proposed system's performance.

7.3 Conclusion

In spite of these difficulties, it is considered, with reasonable justification, that the FCPP and the ICA-FCPP represent valuable contributions to the field. These algorithms are not only effective in reducing environmental noise and for enhancing speech quality, they are also practical in terms of their implementability, and in terms of the kinds of environments in which they can operate. It is hoped therefore, the succeeding stages of this work will bring these ideas to fruition in terms of a commercializable product.

Appendix A
Speech Samples Used in Testing

HINT Sentences

They head a funny noise.
He found his brother hiding
The book tells a story
The milk is by the front door.
She lost her credit card.
The team is playing well.
The little boy left home.
They're going out tonight.
A cat jumped over the fence.
He wore his yellow shirt.
The boy did a handstand.
The young people are dancing.
The shirts are in the closet.
They watched a scary movie.
The milk is in the pitcher.
The truck drove up the road.
The tall man tied his shoes.
A letter fell on the floor.
The dog growled at the neighbours.
A tree fell on the house.
Her husband brought some flowers.
The children washed the plates.
They went on vacation.
The mailman shut the gate.
The dishcloth is soaking wet.
She spoke to her eldest son.
The oven door was open.
He broke his leg again.
The chicken laid some eggs.
They met some friends at dinner.
The man called the police.
The cat drank from the saucer.
They took some food outside.
The matches were on a shelf.
The kitchen sink was empty.
The candy shop was empty.
The fruit was on the ground.
Don't ask me to carry an oily rag like that.
She had your dark suit in greasy wash water all year.

Samples from Film and Television

"Well, I got to hand it you George; you certainly have a talent for trivializing the momentous and complicating the obvious". (Gettysburg).

"Let Achilles fight for honour, let Agamemnon fight for power, and let the gods decide". (Troy).

"I made you coffee...that ought to help you cope with the injustice of the world a little". (Dark Angel).

"Master at arms, take that man below and clap him in irons". (Master and Commander: Far Side of the World).

"I hate it when you talk of the service in this way, it makes me feel so very low". (Master and Commander: Far Side of the World).

"But on the net Mulder, he can find out practically anything about you". (X-Files).

"Well, no radiation so far. I'm sure you're glad to hear that". (Star Trek: The Next Generation).

"I'm sure you've suffered mightily in my absence". (Battlestar Galactica).

"DRADIS contact. Single bogey and it's nearly on top of us. No transponders, no recognition codes. It has to be a Cylon raider". (Battlestar Galactica).

"Gentleman, you will always remember this as the day that you almost caught Captain Jack Sparrow". (Pirates of the Caribbean).

"If you spring me from this cell, I swear on pain of death, that I shall take you to the Black Pearl". (Pirates of the Caribbean).

Appendix B
Music Samples Used in Testing

Bortnyansky, D. "O Lord Who Shall Dwell on the Holy Hill?". Perf. Monks Choir Kyiv-Pechersk Monastery. Anthology of Sacred Music Vol. I. Compact disc. Origen Music, 2002.

Gjallarhorn. "Kokkovirsi" Rimfaxe. Compact disc. Northside Records, 2006.

Hedningarna. "Alkusanat". Karelia Visa. Compact disc. Northside Records, 1999.

Kanno, Yoko. "Home Stay". Stand Alone Complex OST. Compact disc. Bandai, 2004.

Kanno, Yoko. "Rize". Stand Alone Complex OST. Compact disc. Bandai, 2004.

Kanno, Yoko. "Tank!". Cowboy Bebop. Compact disc. Victor Entertainment, 2001.

Ranarim. "Hem Igen". Morgonstjärna. Compact disc. Northside Records, 2006.

Russell, Tom. "Sinatra Played Juarez". Borderland. Compact disc. Hightone Records, 2001.

Vas. "Mandara". Feast of Silence. Compact disc. Narada, 2004.

Zevon, Warren. "Lawyers, Guns and Money". A Quiet, Normal Life: The Best of Warren Zevon. Compact disc. 1990.

References

[1] Cherry, E. Colin. "Some Experiments on the Recognition of Speech With One and With Two Ears". Journal of the Acoustic Society of America. 25 (1953): 975-979.

[2] Wang, Deilang and Brown, Guy J. "Fundamentals of Computational Auditory Scene Analysis" Ed. Deliang Wang and Guy J. Brown. Computational Auditory Scene Anlaysis. Piscataway, NJ: IEEE Press, 2006. 1-44.

[3] Haykin, Simon and Zhe Chen. "The Machine Cocktail Party Problem". Ed. Simon Haykin et al. New Directions in Statistical Signal Processing: From Systems to Brains. Cambridge, MA: MIT Press, 2006. 51-73.

[4] Haykin, Simon and Zhe Chen. "The Cocktail Party Problem". Neural Computation. 17 (2005): 1875-1902.

[5] Quateieri, Thomas F. Discrete-Time Speech Signal Processing. Upper Saddle River, NJ: Prentice-Hall, 2002.

[6] Hansler, Eberhard and Gerhard Schmidt. Acoustic Echo and Noise Control: A Practical Approach. Hoboken, NJ: Wiley, 2004.

[7] Brown, Guy J. and Palomaki, Kalle J. "Reverberation". Ed. Deliang Wang and Guy J. Brown. Computational Auditory Scene Analysis. Piscataway, NJ: IEEE Press, 2006. 209-250.

[8] Eliasmith, Chris. Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems. Cambridge, MA: MIT Press, 2003.

[9] Algazi, V.R. et al. Structural Composition and Decomposition of HRTFs. IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics. New Paltz, NY. Oct 21-24, 2001.

[10] Zotkin, D.N. et al. Virtual Audio Customization Using Visual Matching of Ear Parameters. IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics. New Paltz, NY. Oct 21-24, 2001.

[11] Van Trees, Harry L. Optimum Array Processing. Hoboken, NJ: Wiley, 2002.

[12] Haykin, Simon. Neural Networks: A Comprehensive Foundation. Upper Saddle River, NJ: Prentice-Hall, 1999.

[13] Bell, A.J. and Sejnowski, T.J. "An Information-maximization approach to Blind Source Separation". Neural Computation. 16 (1995): 1129-1159.

[14] Torkkola, Kari. "Blind Separation of Delayed and Convolved Sources". Ed. Simon Haykin. Unsupervised Adaptive Filtering. Hoboken, NJ: Wiley, 2000.

[15] Lambert, Russell and Nikias, Chrysostomos L. "Blind Deconvolution of Multipath Mixtures". Ed. Simon Haykin. Unsupervised Adaptive Filtering. Hoboken, NJ: Wiley, 2000

[16] Bregman, Albert. Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: MIT Press, 1990.

[17] Hu, G. and Wang, Deliang. "Auditory Segmentation based on Onset and Offset Analysis." IEEE Transactions on Audio, Speech and Language Processing. 15 (2007): 396-405.

[18] Roman, Nicoleta et al. "Binaural Segregation in Multisource Reverberant Environments." Journal of the Acoustical Society of America. 120 (2006): 4040:-4051.

[19] Gelfand, Stanley A. Hearing: An Introduction to Psychological and Physiological Acoustics. New York: Marcel Dekker, 2004.

[20] Geisler, C. Daniel. From Sound to Synapse. Oxford, UK: Oxford, 1998.

[21] Hyvarinen, Aapo et al. Independent Component Analysis. Hoboken, NJ: Wiley, 2001.

[22] Fettiplace, Robert and Hackney, Carole M. "The Sensory and Motor Roles of Auditory Hair Cells". Nature Reviews: Neuroscience. 7 (2006): 19-29.

[23] Smith, Evan C. and Lewicki, Michael. "Learning Efficient Codes of Natural Sounds Yields Cochlear Filter Properties." Advances in Neural Information Processing Systems 17. Cambridge, MA: MIT Press, 2004.

[24] E. Smith and L. Holt, A theoretical model of cochlear processing improves spectrally-degraded speech perception, Annual Meeting of the Acoustical Society of America, Salt Lake City, Utah. June 4-8, 2006.

[25] Cooke, Martin. "A Glimpsing Model of Speech Perception in Noise." Journal of the Acoustic Society of America. 119 (2006): 1562-1573.

[26] Rickard, Scott and Ozgur, Yilmaz. "On the Approximate W-Disjoint Orthogonality of Speech". ICASSP 2002.

[27] Jourjine, Alexander, et al. "Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources From 2 Mixtures". ICASSP 2000.

[28] Dayan, Peter and Abbot, L.F. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. Cambridge, MA: MIT, 2001

[29] Bregman, Albert S. and Rudnicky, A. "Auditory Segregation: Stream or Streams?" Journal of Experimental Psychology: Human Perception and Performance. 1 (1975) 263-267.

[30] Bregman, Albert S. "Auditory Streaming: Competition Among Alternative Organizations" Psychology: Human Perception and Performance. 4 (1978) 380-387.

[31] Liberman, A. M. "On Finding That Speech is Special". American Psychologist. 37 (1982): 148-167.

[32] Fowler, C. A. and Rosenblum, L. D. The Perception of Phonetic Gestures. Conference on Modularity and the Motor Theory of Speech Perception. June 5-8, 1988. New Haven, Conn.

[33] Guoning Wu and Deliang Wang. "Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation." IEEE Transactions on Neural Networks. 15 (2004): 1135-1150.

[34] Shilling, R. D. and Shinn-Cunningham, Barbara. Ed. Kay Stanney. "Virtual Auditory Displays". Handbook of Virtual Environments Technology. Mahwah, NJ: Erlbaum, 2002.

[35] Kopco, N. and Shinn-Cunninghman, Barbara. Auditory Localization in Rooms: Acoustic Analysis and Behaviour. 32nd International Acoustical Conference, 2002. September 10-12, Zvolen, Slovakia.

[36] Wightman, F. L. and Kistler, D. J. "Factors Affecting the Relative Salience of Sound Localization Cues". Ed. R. H. Gilkey and T. R. Anderson, Binaural and Spatial Hearing in Real and Virtual Environments. Mahwah, NJ: Lawrence Erlbaum, 1997. 1-23.

[37] Broadbent, D.E. and Ladefoged, Peter. "On the Fusion of Sounds Reaching Different Sense Organs". Journal of the Acoustical Society of America. 29 (1957): 708-710.

[38] Darwin, C. J. and Sutherland, N. S. "Grouping Frequency Components of Vowels: When is a Harmonic not a Harmonic?". Quarterly Journal of Experimental Psychology. 36A (1984): 193-208.

[39] Darwin, C.J. "Perceptual Grouping of Speech Components Differing in Fundamental Frequency and Onset-Time". Quarterly Journal of Experimental Psychology. 33A (1981): 185-207.

[40] Gelfand, Stanley and Silman S. "Effects of Small Room Reverberation Upon the Recognition of Some Consonant Features". Journal of the Acoustic Society of America. 66 (1979): 22-29.

[41] McAdams, Stephen. Spectral Fusion, Spectral Parsing and the Formation of Auditory Images. Ph.D. Thesis, Stanford University, 1984.

[42] Shannon, Robert, V. et al. "Speech Recognition with Primarily Temporal Cues". Science 270 (1995): 303-304.

[43] Cooke, Martin. "Making Sense of Everyday Speech: A Glimpsing Account". Ed. Pierre Divenyi. Speech Separation by Humans and Machines. Norwell, MA: Kluwer, 2005. 305-314.

[44] Devore, Sasha and Shinn-Cunningham. Perceptual Consequences of Including Reverberation in Spatial Auditory Displays. International Conference on Acoustic Displays, 2003. July 6-9, Boston, MA.

[45] Deutsch, Diana. "Two-channel Listening to Musical Scales". Journal of the Acoustical Society of America. 55 (1975): 1156-1160.

[46] Palomaki, Kalle, J. et al. "A Binaural Processor for Missing Data Speech Recognition in the Presence of Noise and Small-Room Reverberation." Speech Communication. 43 (2004): 361-378.

[47] Hartmann, W.M. "Localization of Sound in Rooms". Journal of the Acoustical Society of America. 74 (1983): 1380-1391.

[48] Wallach, H. W. et al. "The Precedence Effect in Sound Localization". American Journal of Psychology. 62 (1949): 315-337.

[49] Slaney, Malcom. An Efficient Implementation of the Patterson-Holdsworth Auditory Filter. Technical Report 45, Apple Computer.

[50] Chi, Taishi, et al. "Spectrotemporal Modulations and Speech Intelligibility." Journal of the Acoustic Society of America. 106 (1999): 2719-2732.

[51] Roweis, Sam. T. "One Microphone Source Separation". Advances in Neural Information Processing Systems 13. Cambridge, MA: MIT Press, 2000. 793-799.

[52] Cooke, Martin et al. "Robust Automatic Speech Recognition Speech Recognition with Missing and Unreliable Acoustic Data". Speeech Communication. 34 (2001) 267-285.

[53] Wang, Deliang. "On Ideal Binary Mask as the Computational Goal of Auditory Scene Analysis." Ed. Pierre Divenyi. Speech Separation by Humans and Machines. Norwell, MA: Kluwer, 2005. 181-197.

[54] Roman, Nicoleta, et al. "Speech Segregation Based on Sound Localization". Journal of the Acoustical Society of America. 114 (2003): 2236-2252.

[55] Brungart, David et al. "Isolating the Energetic Component of Speech-on-Speech Masking with an Ideal Binary Time-Frequency Mask". Journal of the Acoustical Society of America. 120 (2006): 4007-4018.

[56] Brungart, David and Simpson, Brian. "Effect of Target-Masker Similarity on Across-Ear Interference in a Dichotic Cocktail Listening Task". Journal of the Acoustic Society of America. 122 (2007): 1724-1734.

[57] Dreiseitel, Pia and Schmidt, Gerhard. "Evaluation of Algorithms for Speech Enhancement". Ed. Jens Blauert. Communication Acoustics. Berlin: Springer-Verlag, 2005.

[58] Haykin, Simon. Communication Systems. New York: Wiley, 2001.

[59] Jelena, Kovacevic and Vetterli, Martin. Wavelets and Subband Coding. Englewood Cliffs, CA: Prentice-Hall, 1995.

[60] Li, Yipeng and Wang, Deliang. "On the Optimality of Ideal Binary Time-Frequency Masks". Speech Communication. *In Press*.

[61] Barker, Jon et al. Soft Decisions in Missing Data Techniques for Robust Automatic Speech Recognition. 6th International Conference on Spoken Language Processing, 2000. October 16-20. Bejing, China.

[62] Roweis, Sam T. "Factorial Models and Re-filtering for Speech Separation and Denoising". Eurospeech. 7 (2003): 1009-1012.

[63] Reddy, Aarthi M. and Bhiksha, Raj. "Soft Mask Methods for Single-Channel Speaker Separation". IEEE Transactions on Audio, Speech and Language Processing. 15 (2007): 1766-1776.

[64] Abramson, Ari and Cohen, Israel. "Single-Sensor Audio Separation Using Classification and Estimation Approach and GARCH Modelling". IEEE Transactions on Speech and Language Processing. 16 (2008): 1528-1540.

[65] Ozerov, Alexy et al. "Adaptation of Bayesian Models for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs". IEEE Transactions on Audio, Speech and Language Processing. 15 (2007): 1564-1578.

[66] Tabrikian, J. et al. Speech Enhancement by Harmonic Modelling via map pitch tracking. ICASSP 2002, May 13-17, 2002. Orlando, Florida.

[67] Zhang, Wenyao, et al. Pitch Estimation Based on Circular AMDF. ICASSP, 2002, May 13-17, 2002. Orlando, FL.

[68] Mingyang, Wu et al. "A Multipitch Tracking Algorithm for Noisy Speech". IEEE Transactions on Speech and Audio Processing. 11 (2003): 229-236.

[69] Rong, Dong. Perceptual Binaural Speech Enhancement in Noisy Environments. M.A.Sc thesis, McMaster University, 2004.

[70] Supper, Ben et al. "An Auditory Onset Detection Algorithm For Improved Automatic Source Localization." IEEE Transactions on Audio, Speech, and Language Processing. 14 (2006): 1008-1017.

[71] Wang, Deliang. "Feature-Based Speech Segregation". Ed. Deliang Wang and Guy J. Brown. Computational Auditory Scene Analysis. Piscataway, NJ: IEEE Press, 2006. 81-114.

[72] Eberle, Geoff, et al. "Localization of Amplitude-Modulated High Frequency Noise." Journal of the Acoustical Society of America. 107 (2000): 3568-3571.

[73] Grothe, Benedikt. "Sensory Systems: New Roles for Synaptic Inhibition in Sound Localization." Nature Reviews Neuroscience. 3 (2002): 803-812.

[74] Dong, Rong. Technical Report on Adaptive Hearing System Project. *Unpublished Technical Report*, McMaster University, 2006.

[75] Roweis, Sam T. "Automatic Speech Processing by Inference in Generative Models". Ed. Pierre Divenyi. Speech Separation by Humans and Machines. Norwell, MA: Kluwer, 2005. 97-133.

[76] Fishbach, et al. "Auditory Edge Detection: A Neural Model for Physiological and Psychoacoustical Responses to Amplitude Transients". Journal of Neurophysiology. 85 (2001): 2303-2323.

[77] Shao, Yang, et al. "A Computational Auditory Scene Analysis System for Speech Segregation and Robust Speech Recognition". Computer Speech and Language. *In Press*.

[78] Rowe, Daniel. Multivariate Bayesian Statistics: Models for Source Separation and Unmixing. Boca Raton, FL: CRC, 2003.

[79] Pedrycz, Witold and Fernando Gomide. An Introduction to Fuzzy Sets: Analysis and Design. Cambridge, MA: MIT Press, 1998.

[80] Terano, Toshiro et al. Fuzzy Systems Theory and its Applications. San Diego, CA: Academic Press, 1992.

[81] Harris, Chris et al. Adaptive Modelling, Estimation and Fusion from Data: A Neurofuzzy Approach. Berlin: Springer-Verlag, 2002.

[82] Rouat, Jean et al. "A Pitch Determination and Voiced/Unvoiced Decision Algorithm for Noisy Speech". Speech Communication. 21 (1997): 191-207.

[83] Diethorn, Eric J. "Subband Noise Reduction Methods for Speech Enhancement". Ed. Yiteng Han and Jacob Benesty. Audio Signal Processing For Next Generation Multimedia Communication Systems. Norwell, MA: Kluwer, 2004. 91-118.

[84] Nauch, Detlef et al. Foundations of Neurofuzzy Systems. Chichester, UK: Wiley, 1997.

[85] Ito, et al. "Moving-source Separation Using Directional Microphones". Proceedings of the 2nd IEEE International Symposium on Signal Processing and Informtation Technology. Piscataway, NJ: 2002. 523-526.

[86] Celik, Abdullah et al. "Gradient Flow Indpendent Component Analysis in Micropower VLSI". Advances in Neural Information Processing Systems 19. Cambridge, MA: MIT Press, 2006. 187-194.

[87] Pedersen, Syskind et al. "Two-Microphone Separation of Speech Mixtures". IEEE Transactions on Neural Networks. 19 (2008): 475-491.

[88] Bell, Tony and Sejnowski, Terry. "An Information-Maximization Approach to Blind Separation and Blind Deconvolution". Neural Computation. 7 (1995): 1129-1195.

[89] Becker, S. and Hinton J. "A Self-Organizing Neural Network that Discovers Surfaces in Random-dot Stereograms". Nature. 355 (1992): 161-167.

[90] Kan, Kevin. Coherent Independent Components Analysis. M.A.Sc thesis, McMaster University, 2007.

[91] Haykin, Simon and Kan, Kevin. Coherent ICA: Implications for Auditory Processing. 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. Oct 21-24, 2007. New Paltz, NY.

[92] Cherubini, Umberto. Copula Methods in Finance. Hoboken, NJ: Wiley, 2004.

[93] Nelsen, Roger B. An Introduction to Copulas. New York: Springer, 1999.

[94] Ma, Jian and Sun, Zenqi. Copula Component Analysis. ICA 2007. September 9-12, 2007. London, UK.

[95] Cichocki, Andrzej and Unbehauen, Rolf. "Robust Neural Networks with On-Line Learning for Blind Identification and Blind Separation of Sources". IEEE Tranactions on Circuits and Systems -1: Fundamental Theory and Applications. 43 (1996): 894-906.

[96] Douglas, Scott C. and Amari, Shun-ichi. "Natural Gradient Adaptation". Ed. Simon Haykin. Unsupervised Adaptive Filtering. Hoboken, NJ: Wiley, 2000.

[97] Fyfe, Colin. Hebbian Learning and Negative Feedback Networks. Berlin: Springer-Verlag, 2005.

[98] Ricketts, Todd and Mueller, Gustav. "Making Sense of Directional Microphone Hearing Aids". American Journal of Audiology. 8 (1999): 117-127.

[99] Compton-Conley, Cynthia L. et al. "Performance of Directional Microphones for Hearing Aids: Real-World vs. Simulation". Journal of the American Academy of Audiology. 15 (2004): 440-455.

[100] Wiklund, Karl. R-HINT-E: A Realistic Hearing in Noise Test Environment. M.A.Sc. Thesis, McMaster University, 2003.

[101] Wiklund, Karl, et al. R-HINT-E: A Realistic Hearing in Noise Test Environment. ICASSP 2004, May 17-21, 2004. Montreal, QC.

[102] Mori, Yoshimoto et al. "Blind Separation of Acoustic Signals Combing SIMO-Model-Based Independent Component Analysis and Binary Masking". EURASIP Journal on Applied Signal Processing. Volume 2006. 1-17.

[103] Stark, Henry and Woods, John W. Probability and Random Processes with Applications to Signal Processing. Upper Saddle River, NJ: Prentice-Hall, 2002.

[104] Bar-Shalom, Yakov et al. Estimation With Applications to Tracking and Navigation. New York: Wiley, 2001.

[105] Knill, David C. and Pouget, Alexandre. "The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation". Trends in Neurosciences. 27 (2004): 712-719.

[106] Lambert, Russell and Nikias, Chrysostomos L. "Blind Deconvolution of Multipath Mixtures." Ed. Simon Haykin. Unsupervised Adaptive Filtering. Hoboken, NJ: Wiley, 2000.

[107] Lollman, H. W. and Vary, P. "Low-delay Filter Banks for Speech and Audio Processing". Ed. Hansler E. and Schmidt, G. Speech and Audio Processing in Adverse Environments. Berlin: Springer-Verlag, 2009.

[108] Kit, Wong Chun. A Decimated Electronic Cochlea on a Reconfigurable Platform. Master's Thesis, Chinese University of Hong Kong, 2006.

[109] Chassaing, Rulph. Digital Signal Processing and Applications with the C6713 and C6416 DSK. Hoboken, NJ: Wiley, 2005.

[110] Meyer-Baese, Uwe. Digital Signal Processing with Field Programmable Gate Arrays. Berlin: Springer-Verlag, 2009.

[111] Watts, Lloyd et al. "Voice Processors Based on the Human Hearing System". IEEE Micro. 29 (2009):54-63.

[112] Wang, Deliang. "Computational Auditory Scene Analysis and its Potential Application to Hearing Aids". IHCON 2008.

[113] Gerstner, Wulfram and Kistler, Werner. Spiking Neuron Models. Cambridge, UK: Cambridge University Press, 2002.