A Psychoacoustic Engineering Approach to Machine Sound Source Separation in Reverberant Environments

Christopher Hummersone

Submitted for the Degree of Doctor of Philosophy

Department of Music & Sound Recording
Faculty of Arts & Human Sciences
University of Surrey

February 2011

Abstract

Reverberation continues to present a major problem for sound source separation algorithms, due to its corruption of many of the acoustical cues on which these algorithms rely. However, humans demonstrate a remarkable robustness to reverberation and many psychophysical and perceptual mechanisms are well documented. This thesis therefore considers the research question: can the reverberation–performance of existing psychoacoustic engineering approaches to machine source separation be improved? The precedence effect is a perceptual mechanism that aids our ability to localise sounds in reverberant environments. Despite this, relatively little work has been done on incorporating the precedence effect into automated sound source separation. Consequently, a study was conducted that compared several computational precedence models and their impact on the performance of a baseline separation algorithm. The algorithm included a precedence model, which was replaced with the other precedence models during the investigation. The models were tested using a novel metric in a range of reverberant rooms and with a range of other mixture parameters. The metric, termed Ideal Binary Mask Ratio, is shown to be robust to the effects of reverberation and facilitates meaningful and direct comparison between algorithms across different acoustic conditions. Large differences between the performances of the models were observed. The results showed that a separation algorithm incorporating a model based on interaural coherence produces the greatest performance gain over the baseline algorithm. The results from the study also indicated that it may be necessary to adapt the precedence model to the acoustic conditions in which the model is utilised. This effect is analogous to the perceptual Clifton effect, which is a dynamic component of the precedence effect that appears to adapt precedence to a given acoustic environment in order to maximise its effectiveness. However, no work has been carried out on adapting a precedence model to the acoustic conditions under test. Specifically, although the necessity for such a component has been suggested in the literature, neither its necessity nor benefit has been formally validated. Consequently, a further study was conducted in which parameters of each of the previously compared precedence models were varied in each room in order to identify if, and to what extent, the separation performance varied with these parameters. The results showed that the reverberation–performance of existing psychoacoustic engineering approaches to machine source separation can be improved and can yield significant gains in separation performance.

Contents

List of Figures

List of Tables

List of Equations

Publications Arising from this Thesis

Conference papers

Hummersone C., Mason, R., & Brookes, T. (2010), A comparison of computational precedence models for source separation in reverberant environments, in *Proceedings of the 128th Audio Engineering Society Convention*, London, paper 7981. (Chapter 6)

Refereed Journal Papers

Hummersone, C., Mason, R. & Brookes, T. (2010), Dynamic precedence effect modelling for source separation in reverberant environments, *IEEE Transactions on Audio, Speech and Language Processing* 18, 7, 1867–1871. (based on Chapter 7)

Hummersone, C., Mason, R. & Brookes, T. (2011), Ideal Binary Mask Ratio: a novel metric for assessing binary–mask–based sound source separation algorithms, *IEEE Transactions on Audio, Speech, and Language Processing* (accepted). (Chapter 5)

Software

The separation algorithm and precedence models described in Chapters 6 and 7 have been packaged into a MATLAB® GUI and made available from:
`http://www.surrey.ac.uk/msr/people/chris_hummersone/`

Binaural Room Impulse Responses

The Binaural Room Impulse Responses (BRIRs) employed in this thesis (described in Chapter 5) are packaged with the MATLAB® GUI, but have also been made available separately, and at a higher sampling frequency, from:
`http://www.surrey.ac.uk/msr/people/chris_hummersone/`

Acknowledgements

First and foremost, I would like to thank my supervisors: Tim Brookes and Russell Mason, for their invaluable guidance, support and hard work throughout the project and in preparing this document. Thanks to my colleagues, past and present, for putting up with me, eating curry with me, drinking beer with me and occasionally discussing perplexing matters from a diverse range of topics: Kathy Beresford, Rob 'Bob' Conetta, Laurent 'Frenchie' Simon, William 'Willy' Evans, Paulo Marins, Chungeun 'Ryan' Kim, Daisuke Koya and Giuseppe Aliberti. Thanks to Martin Dewhirst for offering me advice and guidance on programming in MATLAB® and C. Last, but never least, thanks to Laurent and to Joshua Brooks for their assistance in capturing the Binaural Room Impulse Responses.

I would like to thank the following people who have, directly and indirectly, provided me with invaluable code that I have used throughout my PhD:

- Dr. Richard O. Duda from San Jose State University, California, for his assistance with implementing the Lindemann model.
- Guy Brown from Sheffield University, UK, for his freely available MATLAB® Gammatone, ERB and cross-correlogram code, available from:
  `http://www.casabook.org`
- Ning Ma for his MATLAB® C/Mex implementation of the Gammatone filter, available from:
  `http://www.dcs.shef.ac.uk/~ning/resources/gammatone/`
- Malcolm Slaney for his MATLAB® "Auditory toolbox" (the Meddis Hair Cell function in particular), available from:
  `http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/`

Chapter

1 Introduction

Sound is generated by a compression and rarefaction of air caused by physical vibrations of the sound source. In real life we often hear many sounds from numerous sources often located in different spatial locations. However, the resulting motion induced on each ear drum can, at any point in time, be measured as a single value indicating its displacement; the motion arises from a summation or mixture of all of these constituent sound sources. Yet, as many observers have documented, humans have an uncanny ability to segregate this mixture into its numerous components. In 1863 (translated 1885), Helmholtz writes:

> In the interior of a ball-room ... we have a number of musical instruments in action, speaking men and women, rustling garments, gliding feet, clinking glasses and so on ... a tumbled entanglement of the most different kinds of motion, complicated beyond conception. And yet, ... the ear is able to distinguish all the separate constituent parts of this confused whole ...
>
> (Helmholtz 1885)

Helmholtz is describing what, some ninety years later, Cherry (1957) termed "The Cocktail Party Effect". In the decades after Cherry coined this term there was to follow a host of psychophysical research, culminating in 1990 with Albert Bregman writing his seminal book *Auditory Scene Analysis* (Bregman 1990)—the first publication to give a comprehensive account of how the brain performs this *cocktail party* processing. In the time following this, engineers from a variety of fields became interested in realising Computational Auditory Scene Analysis (CASA) as they realised the possible uses for such a technology.

1.1 What is Auditory Scene Analysis?

Bregman (1990) points out that the auditory system has a task that is equal to that of vision: it must process complex sensory data and create a mental representation of the world around us. A crucial part of this is deciding which parts of the data are telling us about the same environmental object or event. Clearly, without this ability our perception of the environment around us would be nonsensical. To that end, Bregman

(1990) states that the goal of Auditory Scene Analysis (ASA) is "the recovery of separate descriptions of each separate thing in the environment". As Bregman (1990) points out, it is all too easy to underestimate the complexity of this task, and he provides the following analogy. You are playing a game with a friend at the edge of a lake. You are asked to dig a channel, a few feet long and several inches wide, from the edge of the lake to inland. Your friend does the same. A handkerchief is fastened at the end of each channel such that it can move sympathetically with the water. With you being able to look only at the handkerchiefs, your friend asks you a series of question regarding the lake: how many boats are there? Which one is closer? Which one is the most powerful? Is the wind blowing? Has a large object been dropped in the water? Now consider that the handkerchiefs are your ear drums, the channels are your ear canals and the lake is the air that surrounds your head. These rather difficult–sounding questions are not entirely dissimilar to the kind of questions that are asked of the auditory system during scene analysis. How many people are talking? Who is closer? Who is louder? What is the source of that background noise? Answering these types of questions is the purpose of ASA (Bregman 1990).

1.2 What is Computational Auditory Scene Analysis?

CASA has been defined in the following way:

> It is the field of computational study that aims to achieve human performance in ASA by using one or two microphone recordings of the acoustic scene.
>
> (Wang & Brown 2006)

One important observation to make about CASA is how it differs from other sound source separation techniques such as beamforming and Blind Source Separation (BSS) using Independent Component Analysis (ICA). Beamforming uses spatial filtering to achieve sound source separation: sources coming from a specific direction are enhanced whilst interfering sources from other directions are reduced. Essentially this involves many microphones, one pointing at the desired sound source whilst others are used to cancel out the interfering sound sources. Hence, for n interfering sound sources, $n + 1$ microphones will be required to enhance the desired sound source. This approach can therefore be quite impractical (Wang & Brown 2006).

BSS using ICA is similar to beamforming but combines adaptive filtering and machine learning techniques. The separation is formulated as a problem of calculating a demixing matrix (A^{-1}); the mixture signal $x(t)$, which is a mixture of signals recorded by different microphones, is modelled as a product of A (the mixing matrix) and a vector of unmixed, statistically independent signals $x_m(t)$ such that $x(t) = Ax_m(t)$.

There are, however, several constraints to this approach. Firstly, for the process to work several assumptions must be made resulting in the scope being somewhat limited as a result. As with beamforming, one such constraint is that $n + 1$ microphones are required, although work is being done to reduce this constraint (Lee et al. 1999; Winter et al. 2004). However, a more serious limitation is that A needs to be spatially stationary for a period of time in order for the parameters to be calculated. This is particularly problematic since many sources are not spatially stationary. Lastly, the sources must occupy different spatial locations in order for them to be separable (Wang & Brown 2006).

CASA differs from these approaches by being fundamentally linked to how the auditory system performs source separation. As a result, all CASA models incorporate a level of auditory modelling. Some of these processing techniques will be discussed in Section 3.1. Furthermore, CASA approaches tend to use only one or two microphones (standard microphone or binaural recordings) which makes it more practical to implement (Wang & Brown 2006). Consequently, the research described in this thesis was limited to binaurally captured signals; since CASA is intended to model perceptual (and physiological) processes, this maximises the perceptual relevance of the research.

One further distinction must be made. Many sound source separation algorithms are inspired by ASA theories, but do not strictly adhere to its principles. These algorithms may adopt some CASA techniques. Although these algorithms could contribute to, or form part of, a CASA model, the term "CASA model" is reserved for models that exclusively use perceptual mechanisms and acoustic features known to be used during ASA. The title of this thesis reflects the former: "a psychoacoustic engineering to machine sound source separation..." describes any algorithm that uses a psychoacoustically–inspired approach to sound source separation without necessarily conforming to ASA principles. However, since CASA provides a common reference point for these psychoacoustic engineering approaches, it is useful to centre subsequent discussions on CASA.

1.3 What is the Goal of CASA?

CASA has been defined above but, as pointed out by Marr (1982), an important consideration for any complex information processing system is its goal. So what is the goal of CASA? It was stated in Section 1.1 that Bregman (1990) defines the goal of ASA to be "the recovery of separate descriptions of each separate thing in the environment". However, this goal can not be directly transferred to CASA since it is too vague and has to be adapted to make it more computationally relevant. This led to Wang (2005) proposing that the goal of CASA should be to estimate the Ideal Binary Mask (IBM).

To calculate this, a time–frequency signal representation such as a spectrogram is first divided into discrete units. These discrete units, or Time–Frequency (T–F) units, are simply a subdivision of the time–frequency representation specified by a given time frame and filterbank channel. The ideal T–F mask $\mathbf{m}(i,l)$ is a binary matrix such that each T–F unit is set to one in frequency channel i and frame l when the ratio of the target source energy $\acute{\mathbf{u}}_t$ to total interference energy $\acute{\mathbf{u}}_i$ exceeds a threshold value, and zero otherwise:

$$\mathbf{m}_{\text{ibm}}(i,l) = \begin{cases} 1 & \text{if } 10\log_{10}\left(\dfrac{\acute{\mathbf{u}}_t(i,l)}{\acute{\mathbf{u}}_i(i,l)}\right) > \Theta_{\text{ibm}} \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

where Θ_{ibm} is the threshold and is usually set to zero, equating to a 0 dB criterion (Wang 2005). This concept is based on the the psychoacoustical phenomenon of auditory masking, whereby stronger energy within a critical band masks weaker energy (Moore 2004; Roman et al. 2003).

1.4 Applications of CASA

It is important to consider at this point why CASA would be worthy of research. According to Wang & Brown (2006) there are numerous applications for research into CASA. The following is a list of some of these applications. They justify the academic interest in CASA.

Audio Information Retrieval There is a large amount of audio available in both private and public archives. A useful facility and a key research interest is the ability to search these archives. However, these recordings usually consist of a mixture of sounds and hence separating them is necessary before information can be extracted (Wang & Brown 2006).

Auditory Scene Reconstruction Following the separation of acoustic components, it could be possible to reconstruct the auditory scene with the component sources placed in different spatial locations (Wang & Brown 2006).

Automatic Music Transcription The aim of automatic music transcription is to convert a musical recording into a symbolic (note–based) representation. Clearly, to transcribe multiple instruments first requires each instrument to be separated. Automatic music transcription would also allow the transcription of ethnic music that often has no written form (Wang & Brown 2006).

Communications In February 2008, Audience Inc. announced the release of a voice processing chip for mobile phones based on CASA technology, which actively

extracts the voice in the conversation. Their motivation for the chip is general consumer dissatisfaction with the quality of mobile phone calls, highlighted by a recent audit (Ditech Networks 2008) which showed that, on average, 39% of calls fall below the acceptable audio quality level (a Mean Opinion Score of 2.5 out of 5). The chip apparently achieves a noise suppression factor of 25 dB (Audience Inc. 2008).

Contribution to Hearing Science CASA research can contribute to hearing science by suggesting mechanisms that could aid our understanding of how the auditory system performs ASA (Wang & Brown 2006).

Hearing Prostheses A big problem for hearing aids is that they amplify both speech and noise. This means that listeners using a hearing aid often have trouble understanding speech in noisy environments. CASA may be able to provide a level of noise robustness that could greatly improve the level of speech intelligibility for hearing impaired listeners in noisy environments (Wang & Brown 2006).

Robust Automatic Speech Recognition (ASR) Although ASR has made much progress in recent years, performance of these systems degrades significantly when mixed with acoustic interference (see for example Yang et al. 2007). CASA systems could be integrated into ASR by providing a front-end that handles acoustic interference, thus making ASR more robust (Wang & Brown 2006).

1.5 About this Thesis

Considering the above list of applications for CASA, it may be fair to say that reverberation will be present in many of these scenarios. For example, audio recordings often contain some form of reverberation, from artificial reverberation applied to a lead vocal to a classical recording made in a large concert hall. Speech is also likely to be encountered with reverberation, arising from the room or environment in which the speaker may be located. Yet reverberation presents a major problem for traditional CASA systems that are not specifically designed to handle it (Brown & Palomäki 2006). This is because reverberation blurs many of the cues that CASA systems rely on for separation. However, it is well documented that humans demonstrate a remarkable level of robustness to reverberation. Hence, there must be some additional processing or technique(s) that these traditional CASA systems are missing, causing them to fall short of human performance and robustness.

1.6 Research Questions

This thesis aims to answer the following research question:

> Can the reverberation–performance of existing psychoacoustic engineering approaches to machine source separation be improved?

From this question, several sub-questions arise:

1. What are the problems posed by reverberation to human auditory perception in general?
2. What are the problems posed by reverberation to machine listening in general?
3. What are the human solutions to reverberation?
4. What are the machine listening solutions to reverberation, in particular in terms of source separation? How do machine listening solutions relate to human solutions?
5. Which reverberant source separation solution has most scope for improvement?
6. How should the performance of different approaches to the chosen solution be evaluated? What signals? What metrics?
7. Which approaches work best and are there any lessons to be learned for future development?
8. Can performance be further improved?
9. Are the results generalisable?

These questions will be answered throughout this thesis in order to answer the main research question. Specifically, Questions 1–5 will be answered in Chapter 4, Question 6 will be answered in Chapter 5, Question 7 will be answered in Chapter 6 and Questions 8 and 9 will be answered in Chapter 7. Finally, the main research question will be answered in Chapter 8. However, before these questions can be answered, some background information on human auditory perception, especially ASA, and CASA is necessary and this is provided in Chapters 2 and 3 respectively. For the reader's benefit, a list of acronyms is included on page 144 and mathematical symbols are listed on page 146.

Chapter

2 Auditory Scene Analysis

As discussed in Section 1.6, before the specifics of human auditory perception in reverberation can be discussed, it is first necessary to have an understanding of its underlying mechanisms, particularly with regard to source separation. Therefore the aim of this chapter is to establish the physiological and perceptual mechanisms behind Auditory Scene Analysis (ASA), the process through which humans separate mixtures of sounds. This will be achieved in two steps: firstly, the physiological mechanisms of the peripheral auditory system will be discussed (Section 2.1). This is important for two reasons: firstly, the peripheral processing is arguably an integral part of ASA, since it is the output of this system upon which ASA is performed. Secondly, for this reason, modelling auditory processing is a key component for any psychoacoustically–inspired machine sound source separation algorithm. The second step will be to establish the stages of ASA (Section 2.2). Bregman (1990) lists two stages to ASA: segmentation (Section 2.3), whereby the sound arriving at the ear is broken down into local time–frequency regions, and grouping (Sections 2.4–2.5), whereby these time–frequency regions are recombined such that each combination is likely to have arisen from the same sound source.

2.1 Human Auditory Processing

In his book, Bregman (1990) presumes auditory processing to have already taken place and hence discusses ASA processes in terms of the output of the peripheral auditory system. Presenting the peripheral system here provides a context in which ASA can be discussed through the following sections. Furthermore, modelling these processes is an integral part of CASA, as will be shown in Chapter 3.

The ear can be loosely divided into three sections: outer, middle and inner ear. The inner ear is then connected via the auditory nerve to the brain (Pickles 2008). These sections are explained below (taken from Pickles 2008). See Figure 2.1 for a diagram of the ear.

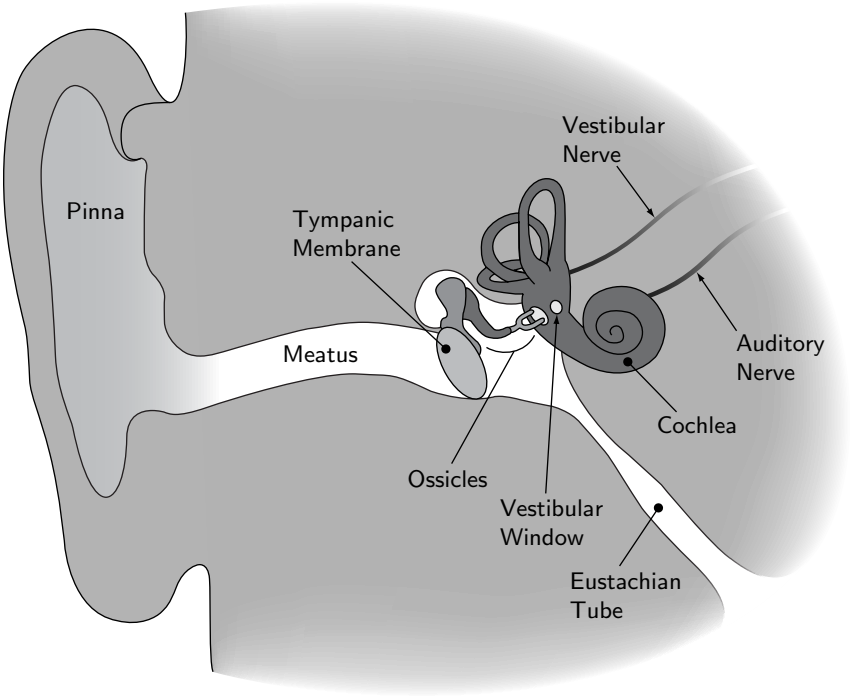


Figure 2.1: The human ear.

Outer Ear

The outer ear is made up of the pinna (the external part), the ear canal (meatus) and the eardrum (tympanic membrane). Because of the position of the pinnae on opposite sides of the head, three important cues are introduced to assist in localising sound: Interaural Time Difference (ITD), Interaural Level Difference (ILD) and spectral changes. For sounds not on the median-sagittal plane (the plane running from head to toe that bisects the left and right sides of the body), an incident sound will introduce an ITD and a frequency–dependent ILD. This is caused by the difference in path lengths between the sound source and ears. The path length difference introduces the ITD; ILD occurs at higher frequencies (above about 1500 Hz) and arises from the baffling effect of the head as the sound propagates to the opposing ear. These cues help to judge the azimuth and elevation of the sound relative to the listener. If the sound is on the median-sagittal plane, spectral changes caused by the head may help to judge the elevation/direction of the sound. Thereafter, the sound travels down the ear canal and causes the eardrum to vibrate.

Middle Ear

The purpose of the middle ear is to match the impedance of the air to the impedance of the cochlear fluids of the inner ear. This is done by the 3 small bones (ossicles: malleus, incus, and stapes) of the middle ear that act as a lever and transmit the vibrations of the eardrum to the oval (vestibular) window of the cochlea.

Inner Ear

The cochlea is found in the inner ear. It is a long and coiled tube, divided length–ways by two membranes: the Reissner's membrane and the basilar membrane. The basilar membrane varies in mass and compliance along its length. This results in the membrane having different resonant frequencies in different regions. For a sinusoidal stimulus at a given frequency there will be a travelling wave induced by the cochlear fluids in the membrane at the same frequency. The resonance of the basilar membrane introduces a timing code that relates to the firing rate of neurones in the auditory system. Also, because the membrane will resonate strongly at the point that has a resonant frequency equivalent to the stimulus frequency, a place code will also be introduced corresponding to that point on the membrane. However, there remains some controversy over the exact nature of cochlear mechanics: a passive process is widely acknowledged, but an additional active process, perhaps initiated by the outer hair cells (at low and medium stimulus levels), is still controversial. It is believed that the active process accounts for the sharp tuning whereas the passive process is insensitive and broadly tuned.

Regardless of the exact nature of the mechanics, the movement of the basilar membrane is transmitted to the Inner Hair Cells (IHCs) that subsequently convert this movement to neural activity. However, the exact nature of this transmission medium also remains unknown. The IHCs initiate action potentials in the spiral ganglion cells, the axons of which form the auditory nerve. The auditory nerve transmits a series of spikes from whose timing, density and place of origin a half-wave rectified and compressed version of the stimulus could perhaps be reconstructed, since action potentials are only initiated by the hairs moving in one direction.

The Auditory Nerve

Pickles (2008) states that responses from the auditory nerve reveal a number of important properties:

- The nerve exhibits similar frequency selectivity to that of the basilar membrane.
- Due to the limited firing rate of the auditory nerve, for low frequency stimuli the nerve appears to be phase–locking whereby it responds directly to instantaneous displacement of the basilar membrane. At higher frequencies the nerve appears to be envelope–locking.
- The nerve fires spontaneously when no stimulus is present, a kind of noise floor.
- The firing rate and stimulus level are correlated by a sigmoidal (s-curve) function and hence the nerve response appears compressed and will saturate at high stimulus levels.

- The nerve adapts to steady stimuli: a higher firing rate is apparent at stimulus onset which then drops to a steady state. After the offset the firing rate drops below the spontaneous level.

Beyond the auditory nerve

The auditory system terminates in the auditory cortex, a section of the brain dedicated to auditory processing. The auditory nerve response has to pass through four neural structures before reaching the auditory cortex: cochlear nucleus, superior olive, inferior colliculus and medial geniculate nucleus. The neurons in these higher centres appear to look for particular perceptual cues, e.g. ILD, ITD, Amplitude Modulation (AM), Frequency Modulation (FM) and periodicity (Pickles 2008). However, relatively little is known about how these higher centres perform ASA on a physiological level, although some knowledge has come from psychophysical studies such as those carried out by Bregman (Wang & Brown 2006). These findings will be presented later in the chapter.

Centrifugal Pathways

So far, this Section has discussed so-called *bottom–up* processing whereby sounds incident at the ear are passed directly from lower level to higher level processing stages. However, Pickles (2008) states that the auditory system is also capable of *top–down* processing whereby higher level sensory data is used in a type of feedback circuit to affect lower level responses. This feedback occurs from the auditory cortex to the outer hair cells through centrifugal pathways and:

- helps to enhance responses to sounds that may be of particular interest
- helps protect the cochlea against damage due to high sound levels
- aids in the detection of signals amidst noise
- adjusts the dynamic range of the hearing system
- aids selective attention

Such a mechanism may have a significant effect on the way ASA is carried out on a physiological level, although the exact nature of this link remains unclear.

Summary

The output of the peripheral auditory system provides the data with which higher centres of the brain are able to perform ASA. To summarise, ASA is performed on neural activity that represents the sound arriving at the ear. The neural activity is directly related to the sound, but has the following characteristics, which are crucial to the operation of ASA:

- The sound is separated in frequency by the cochlea across many bands
- The neural responses appear to be half–wave rectified
- The dynamic range of the stimulus is compressed by the auditory nerve
- Onsets are exaggerated due to adaptation in the auditory nerve
- Cues such as amplitude, AM, FM and directional cues (ITD and ILD) are encoded in the auditory nerve and auditory cortex
- A feedback system may provide a physiological mechanism to assist ASA by adjusting the lower level responses of the auditory system based on higher level data

2.2 ASA

As discussed in Section 1.1, Bregman (1990) states that the goal of ASA is "the recovery of separate descriptions of each separate thing in the environment". This goal has important consequences for how ASA is performed. Clearly ASA does not intend on separating each sound since "each separate thing" may in fact be made up of numerous sounds, e.g. footsteps. This leads to the conclusion that ASA can be considered as a two–stage process: firstly the acoustic mixture arriving at the ear must be segmented: broken down into a collection of local T–F regions in a process called *segmentation*[1]. Secondly, these segments must be recombined both simultaneously and sequentially into collections, or streams, that are likely to have arisen from the same environmental sources—this is called *grouping*. Bregman also points out that these stages are not mutually exclusive but often work together to solve the ASA problem. One more distinction must also be made; there are two types of grouping: *primitive* and *schema–based*. Primitive grouping can be considered a bottom–up process whereby sounds, especially ecological sounds, are segregated based on their intrinsic structure. Schema–based grouping relies on learned pattern recognition; sounds are grouped based on these patterns. This can be considered a top–down process and is particularly relevant in terms of recognising speech (Bregman 1990). Segmentation and grouping are discussed further in the next two sections. It should be noted that in terms of CASA systems, what Bregman refers to as 'segmentation' is actually a three–stage computational process involving a peripheral analysis to simulate ear physiology followed by feature extraction and then segmentation into some intermediate representation.

[1] Bregman borrows the term "segmentation" from video engineering; a common task in this field is to segment an image into its constituent objects.

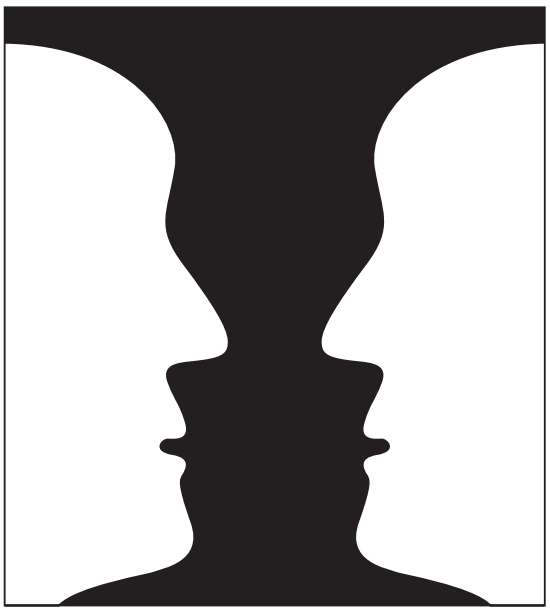


Figure 2.2: The principle of exclusive allocation. Do you see a vase or two faces? According to the Gestalt psychologists, the black/white edge is exclusively allocated so that *either* two faces or a vase are perceived. Adapted from (Bregman 1990).

2.3 Segmentation

As discussed in the previous section, segmentation is the first stage of ASA, whereby an acoustic mixture arriving at the ear is broken down in to local time–frequency regions. A segment is a fundamental building block of a stream (see Section 2.4) and provides an intermediate stage between the peripheral processing of the auditory system and the grouping that takes place in higher stages. These regions are local in terms of belonging to a particular moment in time or to a particular frequency interval. They are described in terms of several properties, including but not limited to: AM, FM, Fundamental Frequency ($F0$), ITD and ILD. Furthermore, each segment is *exclusively allocated* the sound energy received at the ear. This 'principle of exclusive allocation' in audition is analogous to that of vision, as originally proposed by the Gestalt psychologists, whereby a sensory element (in this case a segment) can not be used in more than one description of an object at a time (see Figure 2.2), although Bregman admits that there are exceptions to this rule (Bregman 1990).

Additionally, there is physiological evidence that supports these analyses in higher stages of the auditory cortex. Brown & Cooke (1994a) describe the creation of "computational maps", a term taken from neurophysiology to describe a set of higher stage neurones that are sensitive to a range of parameters such as such as intensity (Suga & Manabe 1982), FM (Shamma et al. 1992), AM (Schreiner & Langner 1988) and spatial location (King & Hutchings 1987). The computational maps that arise from these neurones are two-dimensional, with a previously described parameter on one axis

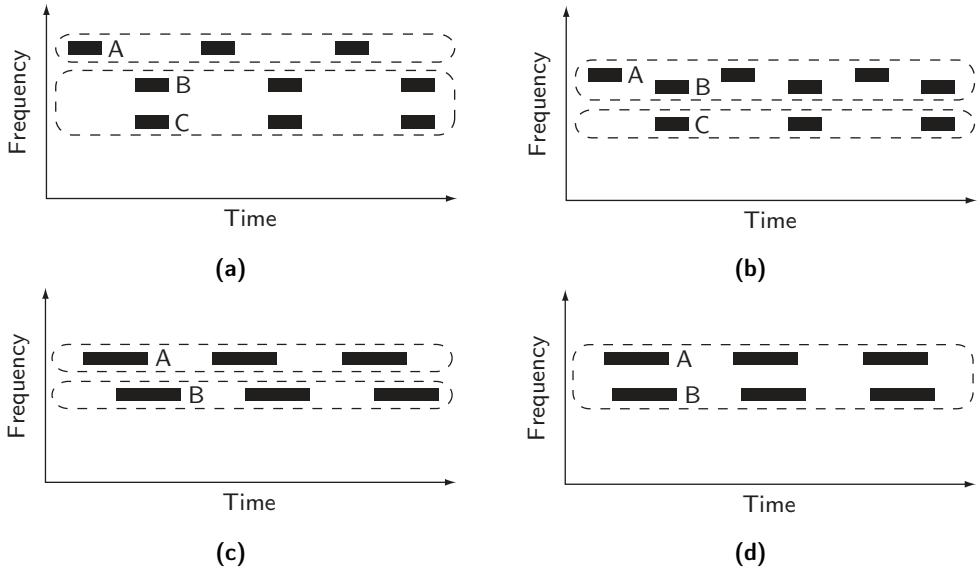


Figure 2.3: Simultaneous grouping of pure tone patterns; adapted from Wang & Brown (2006) (see Section 2.4.1 for details).

and frequency on an orthogonal axis.

2.4 Grouping

As discussed in Section 2.2, grouping can be considered as the second stage in ASA. Grouping of segments into streams occurs both simultaneously and sequentially (Bregman 1990). These two forms are discussed below (except where noted, the information is taken from (Bregman 1990)). A discussion of primitive and schema–based grouping is given in Section 2.5.

2.4.1 Simultaneous Grouping

Simultaneous grouping aims to group segments that occur at the same moment in time. Simultaneous grouping can be demonstrated using simple pure tone stimuli. Consider an alternating pattern of three tones. One tone, A, alternates with two simultaneous tones, B and C (see Figure 2.3). If the onsets of B and C are concurrent, and the offsets of B and C are concurrent, then the two tones will be heard as one complex tone (BC) and A will be heard as another stream (Panel (a)). However, if the frequency of A is made similar to that of B then B will be treated as a repetition of A and is less likely to be heard as part of BC (Panel (b)). B and C can also be separated if their onset times are different. For example, for two 250 ms tones presented with an overlap of 50% (Panel (c)), the two tones would be clearly separated into individual streams. When the overlap was increased to 88% the two tones were fused into a single complex tone (Panel (d)). The separation of the tones also became clearer as the frequency difference

was increased. Bregman also discusses other cues that the auditory system may utilise in order to achieve simultaneous grouping. These are listed below.

Spatial Location

Bregman argues that spatial location is one of the most important cues for simultaneous grouping. Bregman also points out that spatial location alone is not enough to achieve grouping and that comparisons between the ears are frequency–specific. Kubovy (1981) argues that spatial location is not indispensable[2] (rather that time and frequency are) and that two identical sounds at different spatial locations will be fused and perceived as coming from an intermediate direction. However, whilst Bregman acknowledges that a difference in spatial location alone can not cause two simultaneous tones to be segregated, he argues that segregation of otherwise identical spatially–separate tones may occur under more complex circumstances. For example, he describes an informal experiment in which himself and a colleague were replayed two auralised complex tones over headphones. One complex tone was simulated at -45° with frequency components at 200, 400, 600 and 800 Hz. The other complex tone was simulated at 45° and had components at 300, 600, 900 and 1200 Hz. Each component had equal intensity. Note that both stimuli had a common component of 600 Hz. The two complex tones were replayed at irregular intervals but in such a way as to always overlap. If the sounds were on at the same time the 600 Hz component would have identical intensity and phase in each ear. If spatial location was a truly indispensable attribute, then, according to Kubovy, the 600 Hz components should have been fused and perceived in-between the complex tones. However, neither participant found this to be the case and instead found the 600 Hz component to behave independently and in the same way as its neighbouring components[3].

More recent research has shown that ITD is only a weak cue for simultaneous grouping (Culling & Summerfield 1995; Drennan et al. 2003; Edmonds & Culling 2005). Edmonds & Culling (2005) performed three experiments measuring the Speech Reception Threshold (SRT) of target speech mixed with a masker in range of spatial configurations. The target and masker were each split into two frequency bands at a splitting frequency (two were used: 750 and 1500 Hz). The spatial configuration of the signals was then manipulated. Specifically, each signal could either be spatially split: the two frequency bands of the signal had different ITDs; or consistent: the two frequency bands of the signal had the same ITD. Three ITDs were used (-500, 0 and 500 μs) and the SRT achieved in different combinations of split and overlapping signals was measured. The results showed that performance was best when the target and

[2] A good analogy for an *indispensable* attribute can be taken from vision. Consider two identical objects, they can either be separated in *time* or *space* in order for the viewer to see two. Hence, space and time can be considered indispensable attributes of vision (Kubovy 1981).

[3] Note that this example also illustrates an exception to the principle of exclusive allocation.

masker had different ITDs, regardless of whether the target or masker had an ITD that was consistent across frequency. This showed that although ITD was an important cue for segmentation, other cues were more important for grouping the target and masker.

Harmonicity

Frequency cues have already been discussed to some extent above. Indeed, separation in frequency is an important factor in determining whether frequency components will be grouped in to the same stream. However, other important observations have been made of grouping based on spectral cues. Grouping is found to be very likely if the frequency components form a harmonic series, and the auditory system is capable of identifying more than one harmonic series, provided they have different fundamental frequencies. In fact, the auditory system is able to infer any frequencies that may inadvertently be missing, including the fundamental. Density of the spectra is another important factor—the denser the spectra, the higher the likelihood of grouping. Relative intensities of partials is found to play a role, again, the higher the similarity, the higher the likelihood of grouping.

Amplitude and Frequency Modulation

Bregman refers to common AM and FM as the common fate principle, after the Gestalt psychologists. It is the idea that numerous frequency components can often be seen to be doing the same thing (albeit in different frequency ranges). In fact, Bregman argues that this is a very powerful grouping principle, since it is very unlikely that different sound sources will produce sounds that behave in the same way. In terms of frequency, modulations can be loosely divided into two subsets: gliding changes and micromodulation. Gliding changes refers to relatively slow and gradual shifts in partials, such as those exhibited in the shifting pitch of the voice during conversation. Micromodulation refers to smaller and faster changes in frequency that may also be exhibited by the human voice—during conversation and singing—and also by many musical instruments and sounds. These frequency changes may be of the order of one percent, although conscious vibrato may be as much as twenty percent. In terms of amplitude, again we may consider two subsets of modulation: onset/offset synchrony and changes in amplitude. Onset and offset synchrony has been discussed above and provides powerful evidence that partials are being produced by the same sound source. Changes in amplitude occur in many natural circumstances, from speech to different kinds of environmental noise. In the case of music, the modulation pattern may be periodic, such as the tremolo of a string instrument.

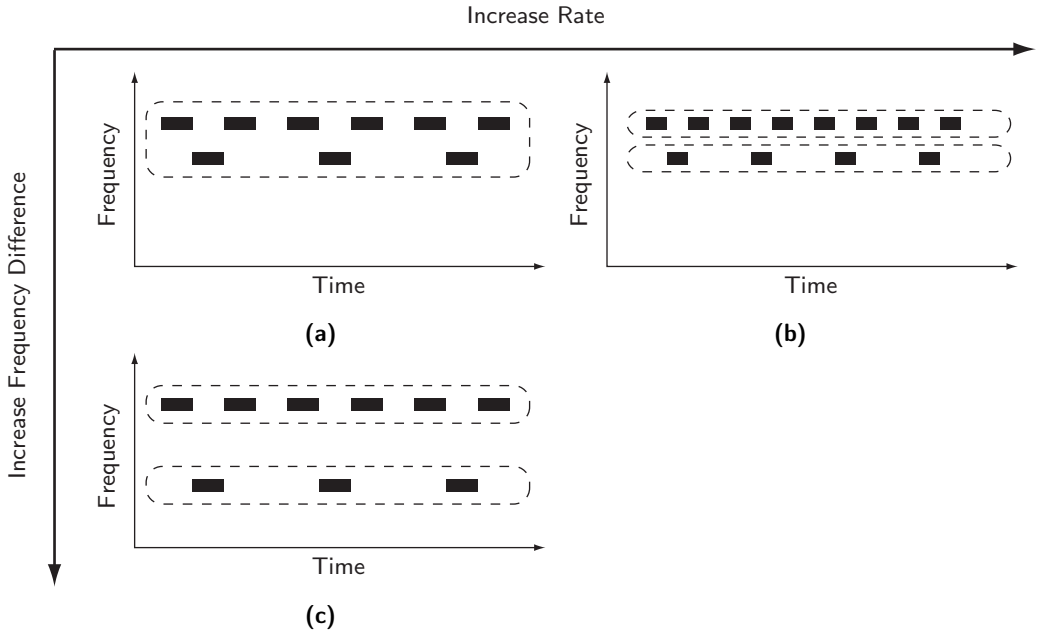


Figure 2.4: Sequential grouping of alternating pure tone patterns; adapted from Wang & Brown (2006) (see Section 2.4.2 for details).

2.4.2 Sequential Grouping

Sequential grouping aims to group segments occurring at different instances in time yet are likely to have originated from the same physical source. Sequential grouping can also be demonstrated using very simple stimuli and, like simultaneous grouping, there are numerous cues that the auditory system can utilise to inform grouping.

Temporal Relations

Van Noorden (1975) uses a pattern of two alternating pure tones and varies both the rate at which the tones sound and the frequency interval between the tones (see Figure 2.4). Firstly, the tones are presented at a slow rate, with the time between onsets being about 150 ms; the frequency difference is less than about four semitones. In this case the listener hears one stream of alternating tones (Panel (a)). As the tone rate is increased, the listener finds it increasingly difficult to hear one stream (see Panel (b)). Similarly, for the same slow tone rate, separate streaming of the tones becomes more likely as the frequency difference between the tones increases. With an interval of 12 semitones or more, two streams are heard (see Panel (c)); in the interim, listeners can choose to hear one or two streams (Bregman 1990; Van Noorden 1975).

Frequency

Unfortunately, the effect shown above can not be extrapolated on to similar experiments with complex tones. As Bregman points out, complex tones have three properties in

frequency: fundamental frequency, pitch, and spectral balance. It is interesting to note that pitch is perceived at the fundamental of a complex harmonic tone, independently of the presence of the fundamental (Licklider 1951). The spectral balance refers to the relative levels of the harmonics. Bregman finds that all of these characteristics have an additive influence on sequential grouping, with proximity and similarity being key factors for comparison.

Spatial Location

Contrary to spatial location in simultaneous grouping, Bregman argues that sequential grouping by spatial location is not as strong as one might expect. A good example of evidence that suggests this was provided by Deutsch (1975). In her experiment, an alternating ascending and descending scale pattern was presented binaurally such that the descending scale was presented to alternating left and right ears whilst the ascending scale was sent to the opposing ear (see Figure 2.5(a)). The expected outcome would be that notes were grouped based on the ear of presentation, i.e. by location (Figure 2.5(b)). However, most listeners reported a grouping by frequency, as shown in Figure 2.5(c). And although this experiment was carried out with relatively slow tone rates (each tone was 250 ms long), Bregman repeated the experiment at higher tone rates and obtained similar results. Bregman concludes that whist sequential grouping by spatial location may not be as strong as it is by utilisation of other cues, we should expect it to be a powerful multiplier when those other cues provide complimentary evidence.

2.5 Primitive versus Schema–based Grouping

The examples of grouping given in Section 2.4 are all examples of primitive grouping. Whilst each environment is different—different animals, languages and music to name but a few—and requires individual adaptation, there are some fundamental rules of environmental sound that apply to a broad range of sounds in the world. For example, when a complex sound changes over time, in most cases the harmonics of the sound will tend to change in a complimentary manner—in direction, frequency and amplitude. This is primitive grouping. Primitive grouping is innate: it is observable from birth and actively involves partitioning the sound. However, Bregman argues that this can not be the whole story; he states that separating sounds is not based entirely on uncontrolled mechanisms and that many instances of separation require prior knowledge and conscious effort.

Listeners gain knowledge about particular types of sound, such as speech, music, machine noises, etc. and store this data in units of mental control known as schemas. Each schema stores information about an individual regularity in our experience. To

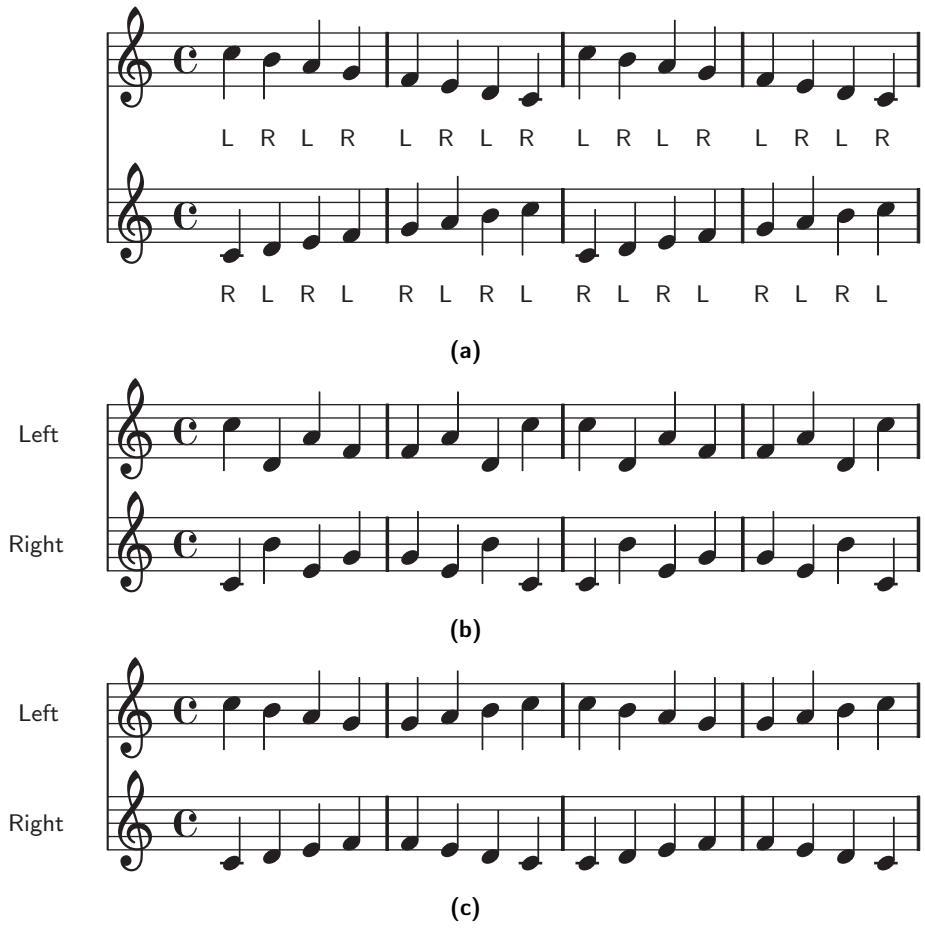


Figure 2.5: Evidence from Deutsch (1975) for the lack of sequential grouping by common spatial location. **(a)** The stimulus played to the listener. **(b)** The expected result with grouping by coincident ear. **(c)** The actual grouping, which appears to have been performed by frequency.

take language for example, schema may exist for the sound of "a", one for the word "apple", one for the grammatical structure of a particular sentence and one for a particular pattern of conversation exchange.

Schemas can make an important contribution to scene analysis. Two examples illustrate this contribution. One example can be observed when synthesising two different vowel sounds which have the same fundamental frequency, the same onset and offset time and are located at the same spatial location. Separating these vowels would be almost impossible using the primitive grouping principles discussed above, yet listeners are able to do so. Another example can be observed when trying to separate a phoneme from a sudden and abrupt loud noise (i.e. the noise is shorter than the phoneme). The auditory system is able to select the frequency components that it expects based on the schema from the noise and they are heard as part of the speech sound. Schemas hence do not actively partition sound, but instead select information from the evidence that is available. This process requires attention and as such is not innate like primitive grouping.

2.6 Summary

The aim of this chapter was to establish the mechanisms behind ASA. This was achieved in two steps: firstly the physiological mechanisms of the peripheral auditory system were established. Secondly, the mechanisms of ASA were presented. With regard to the first step, numerous observations were made in Section 2.1 with respect to auditory physiology. Firstly, it was established that the outer and middle ear provide directional filtering and match the impedance of the air to the impedance of the inner ear. Secondly, the inner ear filters the sound into numerous frequency channels by way of the cochlea and basilar membrane. Thirdly, the auditory nerve exhibits numerous interesting properties such as frequency selectivity, a kind of noise floor, non-linear compression and adaptation to steady stimuli. In response to the second step, the two stages of ASA were discussed. Firstly, segmentation was presented, which is the process that breaks the sounds arriving at the ear into local time–frequency regions. These segments are described by numerous properties, including AM, FM, $F0$, ITD and ILD. These segments are then recombined into streams that represent each sound source. This grouping takes place both simultaneously and sequentially in time by grouping segments that are similar in terms of AM, FM, frequency, spatial location, harmonicity or temporal relations. Furthermore, grouping can either take place using primitive mechanisms that are innate or using learned schemas, such as those that are used to group components of speech.

Chapter

3 Computational Auditory Scene Analysis

As discussed in Section 1.6, before the specifics of machine listening and source separation in reverberation can be discussed, it is first necessary to have an understanding of the techniques used in machine listening and source separation. Therefore the aim of this chapter is to explain some common CASA techniques and how they are implemented computationally. This will be achieved from the point of view of a typical CASA system architecture: each component of the architecture will be discussed in this chapter and the pertinent subsidiary techniques and implementations will be presented.

A typical CASA architecture is presented in Figure 3.1. As well as providing the structure for this chapter, it is useful for comparing how CASA is performed with how humans perform ASA, as described in Chapter 2. As Figure 3.1 shows, the first stage is to analyse the acoustic mixture to produce a representation of auditory nerve activity, this will be discussed in Section 3.1. Following this, acoustic features, or cues, such as periodicity, onset/offset time, AM and FM are extracted. This will be discussed in Section 3.2. Intermediate representations such as segments can then be formed. This will be discussed in Section 3.3. These segments are then grouped according to primitive grouping cues and trained (schema–based) models of individual sound sources to produce streams. This will be discussed in Section 3.4. Finally, the waveform can be re-synthesised such that the performance of the model can be assessed. This is discussed in Section 3.5.

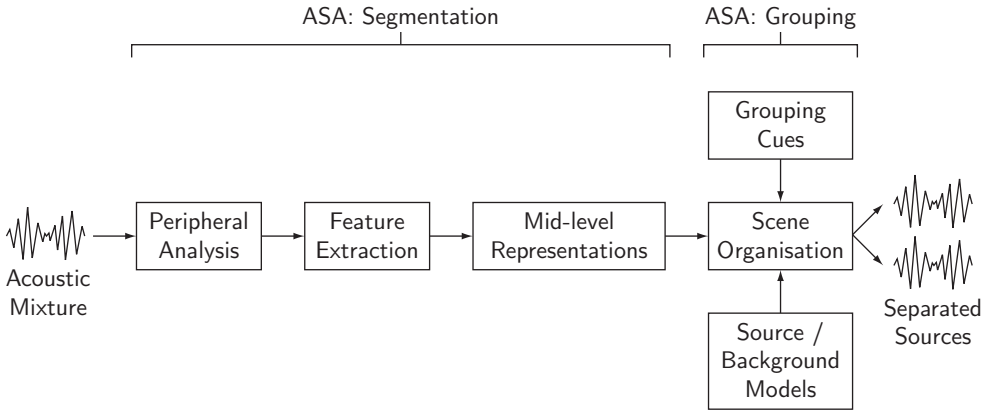


Figure 3.1: A typical CASA system architecture. Adapted from (Wang & Brown 2006).

3.1 Peripheral Analysis

Peripheral analysis forms the first part of Bregman's first stage of ASA and the first stage in a typical CASA system architecture. Usually this involves modelling the mechanical processing of the ear to produce an output that has physiological relevance to ASA. This section will introduce some of the techniques used in CASA systems to model human auditory processing. The techniques are presented in the same physiological order as the auditory system, i.e. the frequency analysis performed by the basilar membrane followed by the sensorineural transduction performed by the IHCs. Following this, higher order analyses such as those described in Section 2.1 (see "Beyond the auditory nerve") are described.

3.1.1 The Gammatone Filterbank

The gammatone filterbank was originally proposed by Patterson et al. (1987) as a model of the frequency analysis performed by the human cochlea. The model is popular in CASA systems for two reasons: firstly, it provides a good match with physiological data and secondly, it is computationally efficient. Patterson et al. (1987) propose a gammatone filter of the time-domain form:

$$gt(t) \propto t^{N-1} e^{-2\pi bt} \cos(2\pi f_0 t + \delta), \qquad (t \geq 0) \tag{3.1}$$

where N is the filter order, b is the bandwidth parameter, f_0 is the centre frequency of the filter and δ is the phase of the impulse response's fine structure. The name is derived from the two halves of the equation: the term before the cosine is the statistical gamma function and the cosine is simply a tone at the centre frequency of the filter (Patterson et al. 1987). As is shown in Figure 3.2(a), the impulse responses of the gammatone filterbank are not time–aligned. For the purpose of making across–frequency measurements and for graphical purposes it may be useful to phase compensate the peaks of the impulse responses. This is achieved by Holdsworth et al. (1988) in two steps: firstly, a lead $t_c = (N-1)/2\pi b$ is introduced to the filter output to align the peaks and secondly, the fine structure is aligned with a phase correction $\delta_c = -2\pi f_0 t_c$. This gives the following result:

$$\tilde{gt}(t) \propto (t + t_c)^{N-1} e^{-2\pi b(t+t_c)} \cos(2\pi f_0 t), \qquad (t \geq -t_c) \tag{3.2}$$

which aligns all impulse response peaks at $t = 0$. The (non-phase-aligned) impulse responses of eight filterbank channels are shown in Figure 3.2(a).

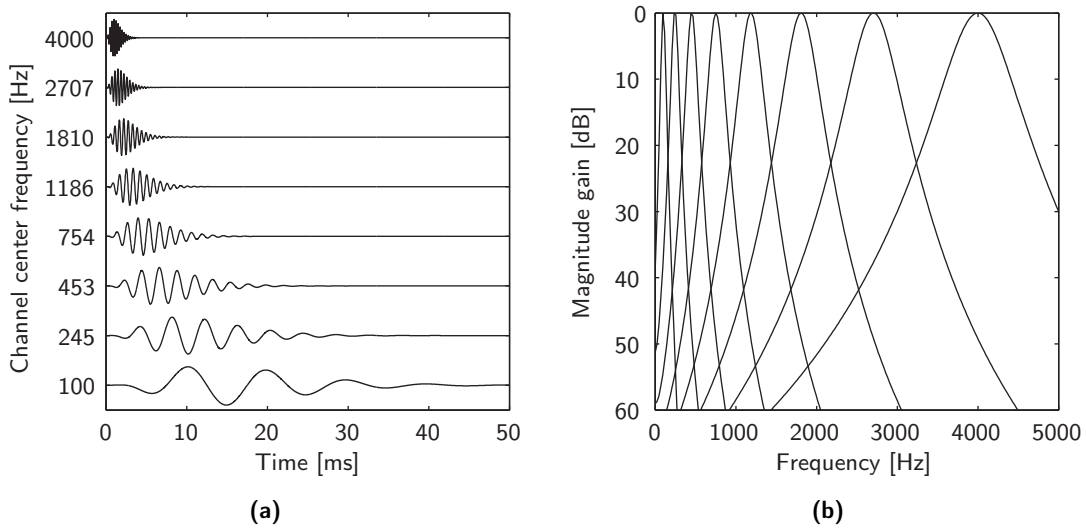


Figure 3.2: The gammatone filterbank with channels equally spaced on the ERB–rate scale. Adapted from (Wang & Brown 2006). **(a)** Channel impulse responses. **(b)** Channel bandwidths.

Cooke (1991) further defined the complex gammatone filter by substituting the cosine term with a complex exponential:

$$\begin{aligned} & gt(t) \propto t^{N-1} e^{-2\pi bt} e^{j2\pi f_0 t}, && (t \geq 0) \\ \implies \quad & gt(t) \propto t^{N-1} e^{-2\pi (b - jf_0) t}, && (t \geq 0) \end{aligned} \tag{3.3}$$

The fine structure output of the gammatone filter can be obtained from the real part of complex coefficients.

According to Holdsworth et al. (1988), in the frequency domain the response of the gammatone filter can be derived either by Fourier transform or by the fact that the time-domain product of the gamma and cosine transforms will correspond to frequency-domain convolution of the Fourier transform of the gamma function $(1 + jf/b)^{-N}$ with a two-point distribution at $\pm f_0$. For simplicity, phase δ is set to zero since it has no discernible effect on the frequency-domain characteristics of the filter. This gives the result:

$$GT(f) \propto \left[1 + j\frac{f - f_0}{b}\right]^{-N} + \left[1 + j\frac{f + f_0}{b}\right]^{-N}, \qquad (-\infty < f < \infty) \tag{3.4}$$

It can be seen from this equation that f_0 is the centre frequency of the filter and the shape is approximately symmetrical on a linear frequency scale. For a fixed order N—which controls the overall shape of the filter—b is proportional to the bandwidth of the filter. Furthermore, the second term of Equation 3.4 can be ignored since according to De Boer & Kruidenier (1990), f_0/b is sufficiently large when modelling the human

auditory periphery. This leads to the approximate frequency response function:

$$GT(f) \approx \left[1 + j\frac{f - f_0}{b}\right]^{-N}, \qquad (0 < f < \infty) \tag{3.5}$$

The bandwidth of the filters is chosen according to the Equivalent Rectangular Bandwidth (ERB) of human auditory filters. The ERB of a filter is the bandwidth of a rectangular filter that has the same peak gain and passes the same total power for a white noise input. This may be regarded as a measure of the critical bandwidth of human auditory filters (Glasberg & Moore 1990; Moore 2004). Moore (2004) states that a good match to human data is given by:

$$\mathrm{ERB}(f) = 24.7 + 0.108f \tag{3.6}$$

Typically, the filter order N is chosen to be 4 and the bandwidth parameter b is chosen thus:

$$b(f) = 1.019\,\mathrm{ERB}(f) \tag{3.7}$$

The filter centre frequencies are usually distributed according to the so-called ERB–rate scale. This is a warped frequency scale—similar to the human critical band scale—where centre frequencies are uniformly distributed according to their ERB. The ERB–rate scale is approximately logarithmic and relates to the number of ERBs, $E(f)$, such that:

$$E(f) = 21.4 \log_{10}(0.00437f + 1) \tag{3.8}$$

The frequency responses of eight gammatone filters are given in Figure 3.2(b); the channels are uniformly distributed on the ERB–rate scale. Note that the spacing results in bands being closer and narrower at low frequency. Furthermore, the bands are all shown to have the same peak gain, but in practice the peak gains can be altered to match the contours of the equal loudness curves (see for example BS EN ISO 226: 2003). The number of gammatone filters chosen for simulations is a trade-off between computational efficiency and physiological accuracy since one filter represents only a single point on the basilar membrane (Wang & Brown 2006).

3.1.2 Inner Hair Cell Modelling

As stated in Section 2.1, the movement of the basilar membrane is induced in IHCs that convert the movement into neural activity (Pickles 2008). A popular model of IHC processing was proposed by Meddis (1986, 1988) and Meddis et al. (1990). Whilst Meddis admits that the exact operation of his model may be controversial he states that it does provide a fast and useful simulation of many characteristics exhibited in auditory nerve activity, such as those discussed in Section 2.1.

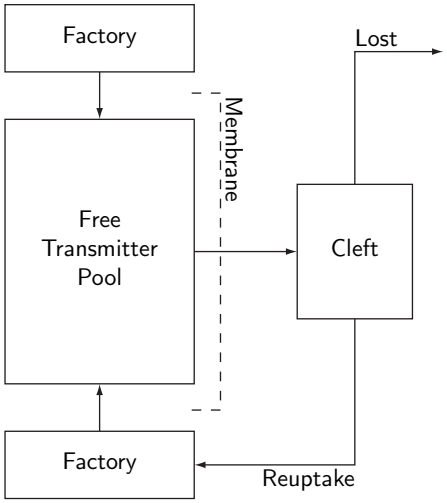


Figure 3.3: Meddis' inner hair cell model. Adapted from (Meddis et al. 1990)

The model takes the displacement of the basilar membrane, such as that given by the gammatone filter output, as its input and converts this into the "fluctuating instantaneous probability of a spike event in a post-synaptic auditory nerve fibre" (Meddis et al. 1990). Meddis' model works by assuming that each hair cell contains three reservoirs of transmitter substance: one is a source pool, one is a reprocessing store and one is a local reservoir between the factory and the source pool (not shown) (see Figure 3.3). Packets of transmitter substance are held in a free transmitter pool which lies near to the cell membrane. The rate at which this transmitter is released across the pre-synaptic cleft is related to the instantaneous displacement of the basilar membrane (or gammatone filter channel). The quantity of transmitter in the cleft determines the instantaneous probability of a post-synaptic spike occurrence.

The time-domain response of Meddis' model is shown in Figure 3.4. The figure was produced by simulating the neural activity in response to a 500 Hz sine wave which was subsequently passed through a gammatone filter with a centre frequency of 500 Hz.

3.1.3 Cochleagram

The cochleagram is simply a method of representing the output of some level of the auditory system (e.g. the cochlea or auditory nerve). Typically, this representation is similar to the familiar spectrogram (Figure 3.5(a)). A cochleagram for the utterance "or some other grease" spoken by a female voice is shown in Figure 3.5(b); notice the quasi-logarithmic ERB–rate distribution of the frequency scale in the cochleagram, allowing for a much more detailed representation of low frequencies (Wang & Brown 2006). The data in Figure 3.5(b) are estimates of the auditory nerve firing rate, calculated using the

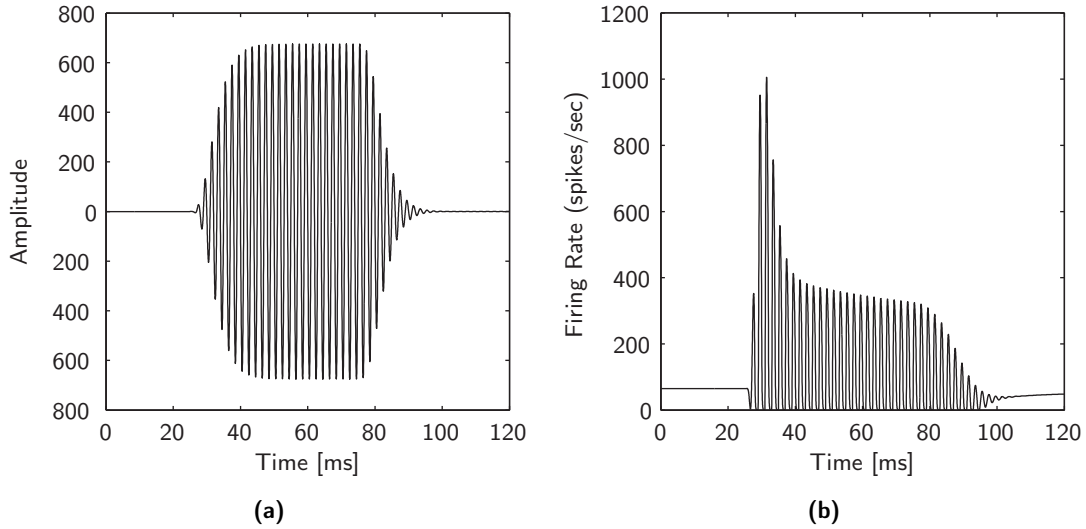


Figure 3.4: The response of Meddis' hair cell model to a 500 Hz pure tone. **(a)** Plot of the pure tone, which has been passed through a gammatone filter with a centre frequency of 500 Hz. **(b)** Simulated neural activity response to the tone. The onset it heavily exaggerated.

method first proposed by Roman et al. (2003) and described in detail in Section 6.1.1[1] (page 92).

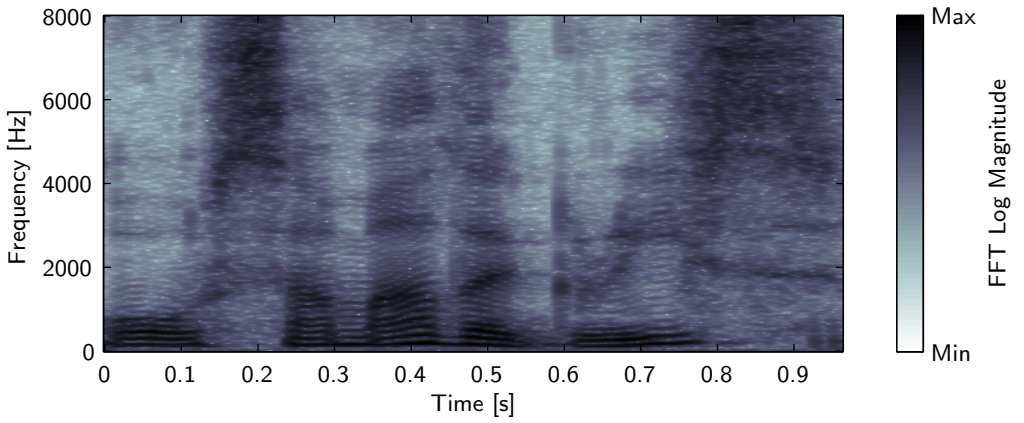
3.1.4 Correlogram

The correlogram (see Figure 3.6) is based on autocorrelation analysis of the signal arriving at each ear. Licklider (1951) first proposed this as a theory of pitch perception and his work now forms the basis for many models of $F0$ estimation (Wang & Brown 2006). Autocorrelation is a statistical method of measuring the correlation of a signal with itself at two different points in time. The correlogram is a time-domain autocorrelation of the simulated auditory nerve activity such as that output by the IHC model. Wang & Brown (2006) define the autocorrelation $\mathbf{a}$ as:

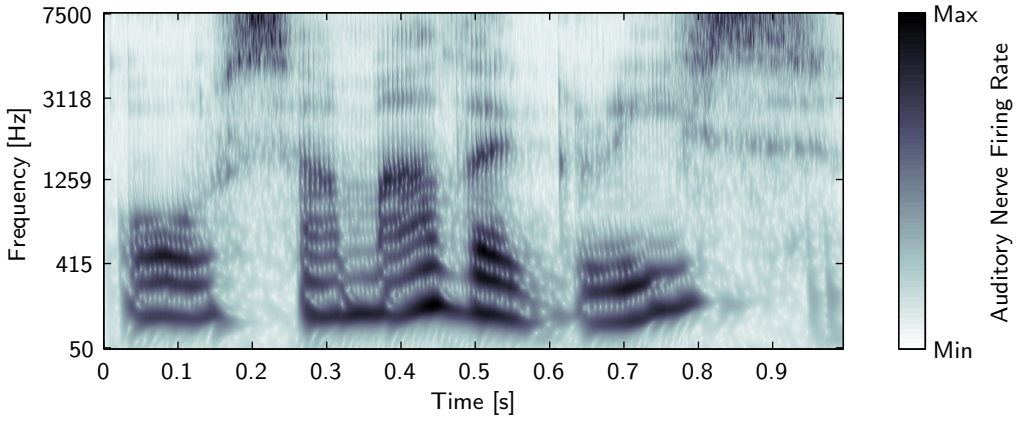
$$\mathbf{a}(i,n,\tau) = \sum_{d=0}^{M-1} \mathbf{h}(i,n-d)\mathbf{h}(i,n-d-\tau)w(d) \tag{3.9}$$

where $\mathbf{h}(i,n)$ is the simulated auditory nerve activity for frequency channel i at discrete time index n, and τ is correlation lag index (for autocorrelation, lags are usually chosen in the range [0,25] ms). The autocorrelation function is performed across M samples which are weighted with the window function w. The window function is typically chosen to be Hann, exponential or rectangular. The autocorrelation function can also be computed in the frequency domain using the Discrete Fourier Transform (DFT) and

[1]Note that there are 2 deviations in the plot compared to the method described in Section 6.1.1: the envelopes are *not* sampled at the frame rate and 128 frequency channels are employed, thus improving the time and frequency resolution.



(a)



(b)

Figure 3.5: Simulated neural activity representations. **(a)** Spectrogram for the female speech "or some other grease" taken from EBU SQAM (1988) calculated with a 512 point Hanning window (at a sampling frequency of 16kHz) and 90% overlap. **(b)** Cochleagram for the same female speech.

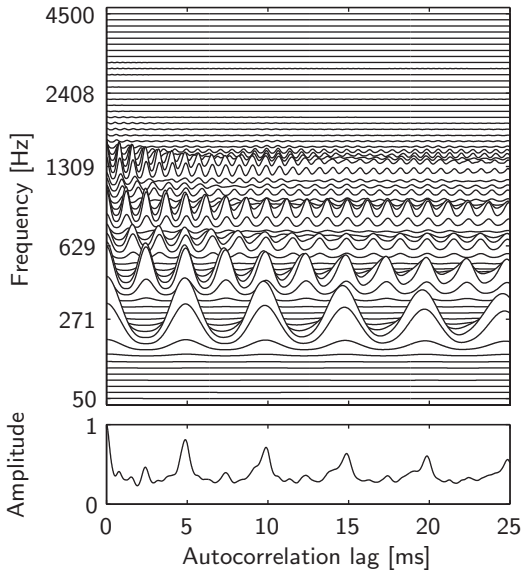


Figure 3.6: A correlogram and pooled correlogram for the vowel /ɑ/, spoken by a female, with a fundamental frequency of 208 Hz.

the Inverse Discrete Fourier Transform (IDFT) such that:

$$\mathbf{a}\big(\mathbf{h}(i,n)\big) = \mathrm{IDFT}\Big(\big|\mathrm{DFT}\big(\mathbf{h}(i,n)\big)\big|^{Q}\Big) \tag{3.10}$$

where the Q parameter adjusts the output of the function; setting $Q = 2$ will give a true autocorrelation output but smaller values can give sharper peaks (Wang & Brown 2006).

Finally, the data obtained from the correlogram can be summed into a pooled correlogram $\bar{\mathbf{a}}$ (shown in the lower panel of Figure 3.6) which is a sum of the correlogram outputs across each of the frequency channels thus:

$$\bar{\mathbf{a}}(n,\tau) = \sum_{i} \mathbf{a}(i,n,\tau) \tag{3.11}$$

Peaks in the pooled correlogram have been shown to correspond closely to perceived pitch. This technique will also show multiple peaks if more than one $F0$ is present which is useful for multi–pitch tracking and algorithms that use $F0$ for sound separation (Wang & Brown 2006).

3.1.5 Cross-correlogram

The cross-correlogram is based on the work of Jeffress (1948) as a model for binaural lateralisaton and especially ITD estimation, although subsequent studies (e.g. Brand et al. 2002) debate the exact physiological mechanisms behind the estimation process. The cross-correlogram is based on the cross-correlation function, which in turn is similar to the autocorrelation function except that the correlation is calculated between two

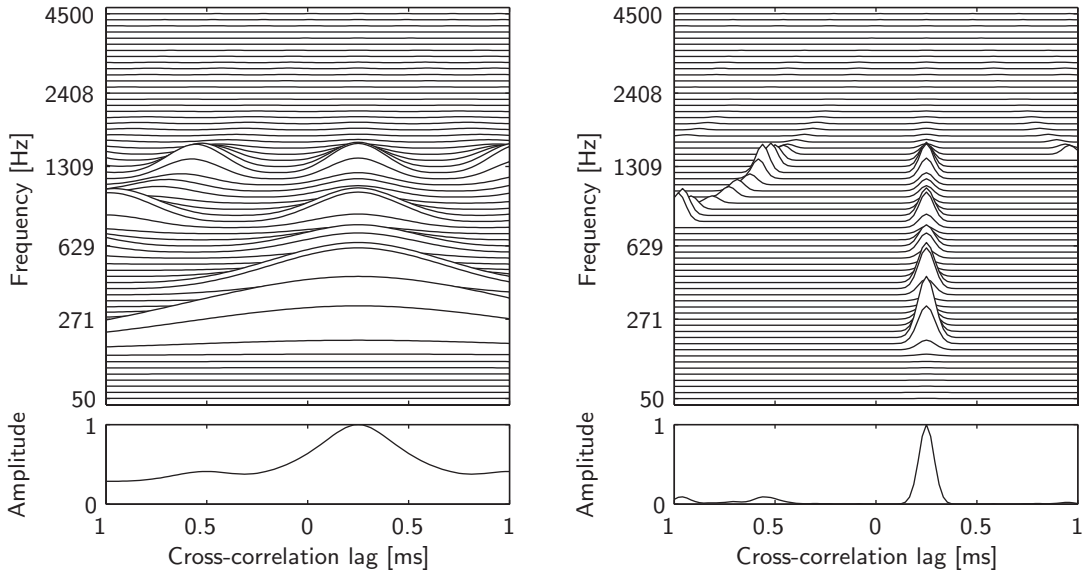


Figure 3.7: The cross-correlogram and skeleton cross-correlogram calculated for the utterance h/ɑ/d with a fundamental frequency of 208 Hz and an ITD of 0.25 ms. **(a)** A cross-correlogram and pooled cross-correlogram for the dichotic stimulus. **(b)** A skeleton cross-correlogram and pooled skeleton cross-correlogram for the same stimulus.

independent random processes (e.g. left and right ears). Wang & Brown (2006) define the cross-correlogram $\mathbf{c}$ as a time-domain cross-correlation of the simulated auditory nerve activity thus:

$$\mathbf{c}(i,n,\tau) = \sum_{d=0}^{M-1} \mathbf{h}_L(i,n-d)\mathbf{h}_R(i,n-d-\tau)w(d) \tag{3.12}$$

For cross-correlation, τ is chosen such that $\{\tau \in \mathbb{Z} : |\tau| \leq T\}$ and T is the maximum cross-correlation lag in samples (usually chosen to equate to 1 ms). An example of a cross-correlogram is shown in Figure 3.7(a). The ITD of the stimulus is indicated by a spine in the cross-correlogram; neighbouring peaks or *sidelobes* in each band are due to harmonic components and filter resonances. Note that this representation does not incorporate ILD (Wang & Brown 2006).

As with the correlogram, data from the cross-correlogram can be pooled across frequency into a pooled cross-correlogram $\bar{\mathbf{c}}$ that emphasises the spine at the stimulus ITD:

$$\bar{\mathbf{c}}(n,\tau) = \sum_{i} \mathbf{c}(i,n,\tau) \tag{3.13}$$

This minimises the contribution of the sidelobes because the position of each peak is frequency dependent. If the stimulus contains multiple sources originating from different azimuths they will show up as independent peaks in the pooled cross-

correlogram (Wang & Brown 2006). The pooled cross-correlogram is shown in the lower panel of Figure 3.7(a).

Finally, there is one last variant of the cross-correlogram: the skeleton cross-correlogram, as proposed by Roman et al. (2003) (see Figure 3.7(b)). This approach is introduced because the simulated cochlea filterbank introduces broad peaks in the cross-correlogram's output, especially at low frequencies. The skeleton cross-correlogram $\mathbf{s}$ is calculated thus:

$$\mathbf{s}(i,n,\tau) = \mathbf{q}(i,n,\tau) * \exp\left(\frac{-\tau^2}{2\sigma^2(i)}\right) \tag{3.14}$$

where, for $\{\tau \in \mathbb{Z} : |\tau| < T-1\}$,

$$\mathbf{q}(i,n,\tau) = \begin{cases} \mathbf{c}(i,n,\tau) & \text{if } \Big(\big(\mathbf{c}(i,n,\tau) - \mathbf{c}(i,n,\tau-1)\big)\big(\mathbf{c}(i,n,\tau) - \mathbf{c}(i,n,\tau+1)\big)\Big) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

where $*$ denotes convolution, σ are frequency–dependent standard deviations, and T is defined as the maximum cross-correlation lag in samples, again usually chosen to equate to 1 ms (see Equation 3.12). The resulting data can again be pooled across frequency to produce a pooled skeleton cross-correlogram with much more defined peaks. This effect is similar to applying lateral inhibition along the ITD/azimuthal axis (Lindemann 1986a,b; Albeck 2003). Note that in their paper, Roman et al. actually warp the cross-correlogram to azimuth before calculating the skeleton cross-correlogram whereas here, for simplicity, ITD has been used to calculate the skeleton cross-correlogram.

3.1.6 Cepstrum Analysis

According to Childers et al. (1977) (see also Bogart et al. 1963), cepstrum analysis comes in numerous flavours: the power cepstrum, the complex cepstrum and the phase cepstrum. CASA literature does not extensively discuss the phase cepstrum; the power cepstrum is the most common and usually referred to simply as the cepstrum. The word cepstrum is derived from the word spectrum—the first four letters having been placed in reverse order. The reasoning for this ties in with the definition of the spectrum. Essentially, the cepstrum can be considered as the power spectrum of the logarithm of the power spectrum of a function. The power cepstrum is often computed by using the DFT (Oppenheim & Schafer 1968, 1999):

$$x_{pc}(n) = \left|\mathrm{IDFT}\left(\log_e\left(\left|\mathrm{DFT}\big(x(n)\big)\right|^2\right)\right)\right|^2 \tag{3.16}$$

Interestingly, convolution in the time-domain is achieved by addition in the cepstral domain:

$$x(n) = x_1(n) * x_2(n) \tag{3.17}$$

or

$$|X(n)|^2 = |X_1(n)|^2 \cdot |X_2(n)|^2 \tag{3.18}$$

or

$$\log|X(n)|^2 = \log|X_1(n)|^2 + \log|X_2(n)|^2 \tag{3.19}$$

According to Childers et al. (1977), cepstral processing has useful applications in wavelet recovery and homomorphic deconvolution. More specifically, cepstral processing has been applied to numerous CASA–related topics such as $F0$ analysis (e.g. Unoki & Hosorogiya 2007), dereverberation (e.g. Van Eeghem et al. 1999) and speech recognition (e.g. Aikawa et al. 1996).

A variant of cepstral processing that is often observed in the literature is the mel–cepstrum. As the name implies, the cepstrum is calculated using the mel–frequency scale originally proposed by Stevens et al. (1937). Furthermore, there is some evidence that humans do perform some cepstrum–like processing in the central auditory system (see for example Wang & Shamma 1995).

3.2 Feature Extraction

Feature extraction forms the first part of Bregman's first stage of ASA and the second stage in a typical CASA system architecture (see Figure 3.1). The purpose of feature extraction is to extract auditory features that may later be useful for grouping signal components into streams, a process that will be discussed in Section 3.4. Most of the literature discusses feature extraction with regard to speech and hence speech feature extraction shall be the focus of this section. Wang (2006) lists 5 key features that are extracted in most CASA systems: pitch or periodicity, cross-channel correlation, onset and offset, AM and FM. The extraction of these features will be discussed in this section.

3.2.1 Pitch and Periodicity

$F0$ estimation has already been dealt with in part in Section 3.1.4 where the correlogram was introduced. Indeed, the correlogram is the most common representation of pitch

(Wang 2006). Notable implementations of this approach include Seneff (1984), Slaney & Lyon (1990), De Cheveigne (1991) and Rouat et al. (1997). However, there are other noteworthy methods of estimating $F0$. The above method can be considered a spectro–temporal approach since it relies on both frequency cues arising from the filterbank and time cues arising from the autocorrelation function. Other methods may tend to use either spectral cues or temporal cues almost exclusively.

In the frequency domain, an effective method of $F0$ estimation was first applied to speech by Schroeder (1968). Schroeder proposed a method whereby peaks in the spectrogram are divided by increasing positive integers. The results are distributed on a histogram, called a Schroeder histogram, with the right–most peak indicating the fundamental frequency. This method works for any periodic signal.

In the time domain, the autocorrelation function can be calculated on any periodic signal (it is not necessary to pass the signal through a filterbank if only the $F0$ is required). The $F0$ is indicated in the autocorrelation function by the first major peak with a non-zero lag (τ) (De Cheveigne 2006).

Matters are complicated somewhat if the signal contains multiple $F0$s. De Cheveigne (2006) states that cues from different voices can often be ambiguous, especially when their $F0$s are in simple ratios. As such, the pitch cues are weakened. Mathematically, the problem may be formulated in the following way:

$$z(t) = y_1(t) + y_2(t), \quad y_1(t) = y_1(t+U), \quad y_2(t) = y_2(t+V), \qquad \forall t \tag{3.20}$$

such that $z(t)$ is the observable signal and is the sum of the two signals $y_1(t), y_2(t)$ with different $F0$s. To extract the two $F0$s, the parameters U and V must be determined to best fit $z(t)$. Hence, De Cheveigne (2006) proposes three basic methods for determining the different $F0$s:

1. Use a single $F0$ algorithm in the hope that it will find multiple $F0$s
2. Use a single $F0$ algorithm iteratively to determine one $F0$; this information is then used to suppress that voice and the algorithm can be reapplied to find another $F0$
3. Estimate all the voices at the same time

It is beyond the scope of this thesis to investigate these algorithms in detail but the interested reader is referred to (De Cheveigne 2006) for a comprehensive overview of the topic.

3.2.2 Cross-channel Correlation

According to Wang (2006), the cross-channel correlation is quite simply a calculation of the correlation between neighbouring frequency channels. As can be seen from Figure 3.2(b), the responses of the filter channels overlap; the degree of this overlap is proportional to the centre frequency of the channel. Consequently, a number of frequency channels may respond to a given harmonic. This information is useful for segmentation since areas of the cochleagram can be compared and grouped based on their cross-channel correlation.

Cross-channel correlation is defined as the cross-correlation of neighbouring auto-correlation responses, which is possible because the phase of the autocorrelation is normalised at zero lag. Specifically, the cross-channel correlation $\mathbf{k}$ is calculated from the normalised autocorrelation $\hat{\mathbf{a}}$, which is calculated as in Equation 3.9 but is normalised to have zero mean and unity variance; the normalisation is necessary because neighbouring gammatone filters have different bandwidths. The operation is summarised thus:

$$\mathbf{k}(i,n) = \frac{1}{M}\sum_{\tau=0}^{M-1} \hat{\mathbf{a}}(i,n,\tau)\hat{\mathbf{a}}(i+1,n,\tau) \tag{3.21}$$

The normalisation of the autocorrelation (for example by ensuring the responses of each channel have zero mean and unity variance) removes any effects of Direct Current (DC) and of the separation of the frequency component and the channel centre frequency (Wang 2006).

3.2.3 Onset and Offset Detection

An onset is a sudden increase in sound level, usually corresponding with the beginning of a sound made by a voice or instrument. Similarly, an offset is usually a sudden drop in level caused when the instrument or voice ceases making the sound. Hence, since onsets and offsets are connected to the rate of change of the sound level, an effective technique for eliciting the onset/offset is to take the first order derivative of the envelope of the signal. However, intensity fluctuations within a voice or background noise can often cause spurious peaks not associated with the onset/offset. Therefore the derivative can be filtered in the hope that these spurious peaks will be removed. Typically the smoothing is achieved with a moving average filter that uses a Gaussian window (Hu & Wang 2004a, 2007). Mathematically, Wang (2006) states that the Gaussian function has the form:

$$G_o(t,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{3.22}$$

where σ is the standard deviation. Given that for any functions g_1 and g_2, $(g_1 * g_2)' = g_1 * g_2'$, the differentiated and smoothed output $o(t)$ is calculated by convolving the signal $x(t)$ with $G'_o(t, \sigma)$ thus:

$$o(t) = x(t) * G'_o(t, \sigma) \tag{3.23}$$

where

$$G'_o(t, \sigma) = \frac{-t}{\sqrt{2\pi}\sigma^3} \exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{3.24}$$

Hence, according to Hu & Wang (2004a), extracting onsets and offsets can be considered a three stage process:

1. Convolve $x(t)$ with $G'_o(t, \sigma)$ to obtain $o(t)$
2. Identify the peaks and valleys of $o(t)$
3. Mark peaks above a predefined threshold as onsets and valleys below a predefined threshold as offsets thus removing any spurious peaks or valleys

3.2.4 Amplitude Modulation

Extracting the AM of a signal basically amounts to extracting the envelope of the signal (Wang 2006). According to Wang (2006), one common method of extracting the envelope is by the Hilbert transform method (although no Hilbert transform is actually taken). The method is demonstrated by Hartmann (1998):

Any real signal $x(t)$ can be expressed as the inverse Fourier transform of the Fourier transform $X(\omega)$:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{j\omega t} X(\omega)\, \mathrm{d}\omega \tag{3.25}$$

The analytic signal $\tilde{x}(t)$ is obtained by removing negative frequencies and multiplying by 2:

$$\tilde{x}(t) = \frac{1}{\pi} \int_{0}^{\infty} e^{j\omega t} X(\omega)\, \mathrm{d}\omega \tag{3.26}$$

The Hilbert envelope $\varepsilon(t)$ is then derived directly from the analytic signal:

$$\varepsilon(t) = |\tilde{x}(t)| \tag{3.27}$$

Equation 3.26 can be re-written by introducing the unit step function $\theta(\omega)$:

$$\tilde{x}(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} e^{j\omega t} X(\omega)\theta(\omega)\, \mathrm{d}\omega \tag{3.28}$$

where

$$\theta(\omega) = \begin{cases} 0 & \omega < 0 \\ \frac{1}{2} & \omega = 0 \\ 1 & \omega > 0 \end{cases} \tag{3.29}$$

The theta function $\theta(\omega)$ creates a natural link between the analytic signal and the Hilbert transform of the real signal $\mathrm{H}[x(t)]$ and hence

$$\tilde{x}(t) = x(t) + j\,\mathrm{H}[x(t)] \tag{3.30}$$

where

$$\mathrm{H}[x(t)] = x(t) * \frac{1}{\pi t} \tag{3.31}$$

However, this final result is often unreliable and leads to slow convergence. A more efficient method is to use Equation 3.26, since fast Fast Fourier Transforms (FFTs) are computationally efficient. Wang (2006) summarises the procedure in the following way:

$$\varepsilon(t) = \Big|\mathrm{IDFT}\Big(2\theta(\omega) \cdot \mathrm{DFT}\big(x(t)\big)\Big)\Big| \tag{3.32}$$

Other methods are of course possible. One such method is to half–wave rectify the signal and then low-pass filter, a method which approximates the Hilbert envelope and is computationally very efficient (Wang 2006). The Hilbert envelope can also be obtained directly from the complex gammatone filter coefficients (see Equation 3.3) by taking the absolute magnitude (Cooke 1991).

3.2.5 Frequency Modulation

Wang (2006) proposes two methods for extracting FM information. In CASA, FM is usually considered to be a frequency transition of a sound component, such as the transitional harmonics of the voice. As such, the first technique involves extracting the spatial contours of a two–dimensional cochleagram. The second technique uses the responses of bandpass filters.

Spatial Contour Extraction

According to Wang (2006), the first technique is performed by convolving the cochleagram response with a set of two–dimensional time–frequency kernels, with the aim of producing a frequency transition map (see also Riley 1989; Mellinger 1991; Brown & Cooke 1994a). According to Brown & Cooke (1994a), this kernel function models a hypothetical set of neurones that are sensitive to different rates and directions of frequency transition.

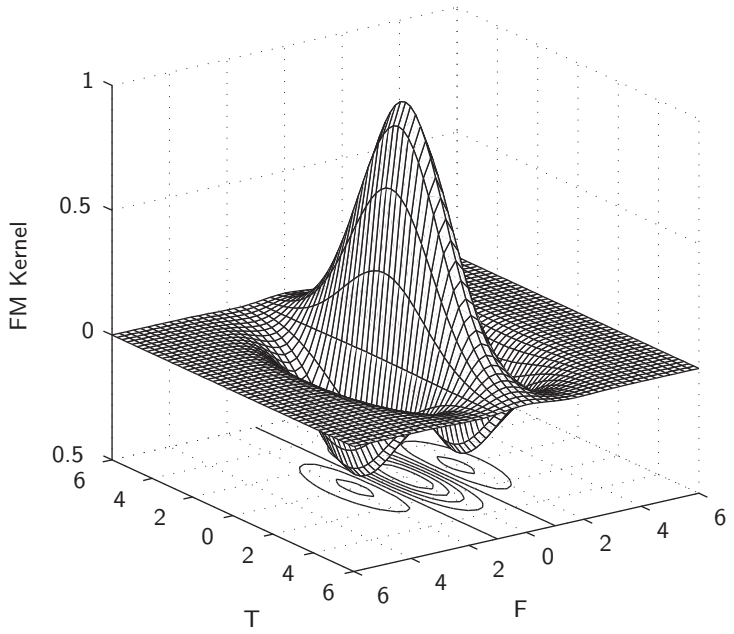


Figure 3.8: The Laplacian-of-Gaussian FM kernel function

According to Wang (2006), this kernel function is typically chosen to be Gaussian and of the form:

$$G_{FM}(t, f, \sigma_t, \sigma_f) = \frac{1}{2\pi\sigma_t\sigma_f} e^{[(t^2/2\sigma_t^2)+(f^2/2\sigma_f^2)]} \tag{3.33}$$

where σ_t and σ_f are widths (standard deviations) in the time and frequency dimensions respectively and are usually chosen such that $\sigma_t > \sigma_f$, causing the kernel to be elongated in the time dimension. Sensitivity to different frequency transition rates and directions is achieved by rotating the co-ordinate system of the function. The frequency change is detected using a Laplacian operation in the frequency dimension (Wang 2006). Hence, according to (Riley 1989), the FM operation becomes:

$$FM(t, f) = -\frac{\partial^2}{\partial f^2} G_{FM}(t, f, \sigma_t, \sigma_f) \tag{3.34}$$

This equation is plotted in Figure 3.8 and uses $\sigma_t = 2$ and $\sigma_f = 1$. The main Gaussian shape along the time dimension is clearly visible whilst the negative valleys above and below with respect to frequency have been created by the Laplacian operation. Hence, for a given frequency transition, the result of the convolution will be maximal when the kernel has the same orientation as the frequency transition. The results of these numerous convolutions can then be plotted to produce a frequency transition map.

Calculating Instantaneous Frequency

The second technique stated above, according to Wang (2006), is performed by calculating the Instantaneous Frequency (IF) of the response of each of the pass bands. The IF of a real signal is found by calculating the time derivative of the instantaneous phase of the analytic signal. However, this technique can be unstable, leading to a range of positive and negative values. One method that overcomes this was proposed by Kumaresan & Rao (1999) whereby the analytic signal is decomposed into two analytic signals, of which one has a constant envelope and positive IF. They then use a form of smoothing, analogous to linear prediction in the spectral domain, to produce a positive and smoothly varying IF. The variation in IF indicates the FM within the frequency channel.

3.3 Mid-level Representations

The purpose of the mid-level representation is to form a description of the post-peripheral analysis data based on the features extracted during feature extraction. This process forms the final part of Bregman's first stage of ASA and the third stage in a typical CASA system architecture (see Figure 3.1). The most frequently used representation is the segment—possibly owing to its perceptual relevance (Wang 2006)—and hence that will be the main focus of this section.

Further to the above, Wang (2006) suggests that the goal of segmentation is to group individual T–F units into segments such that each segment is a continuous region of the cochleagram. Note that this process takes place monaurally; binaural cues are used during grouping to collect segments arriving from the same spatial location (see Section 3.4). These segments can be considered as a mid-level representation since they bridge the gap between T–F units and streams, which can be considered the end product of ASA (Bregman 1990; Wang 2006). One of the most important properties of a segment is that its component units all originate from the same sound source. In this way, the segment can not be subdivided but can instead be grouped with other segments to form a stream. Furthermore, it must be true that adjacent segments belong to different sound sources, or at least exhibit differing properties. Hence, the sound energy belonging to the target source within the segment must be greater than the total sound energy for all other sound sources present within that segment. Note that the above definition of a segment is directly related to the statement made in Section 1.3 that the goal of CASA is to calculate the Ideal Binary Mask (IBM). The IBM is a collection of ideal segments as described above (Wang 2006).

The above definition of segmentation implies that the principle of exclusive allocation has been applied to each T–F unit (see Section 2.3), i.e. each T–F unit belongs to only one segment. In a similar manner to Bregman (1990) recognising that it is not

always true that segments are exclusively allocated, not all CASA models exclusively allocate T–F units. One such example is the model of Ellis (1996) in which sounds are represented either as noise elements, transient elements or wideband periodic elements, and may overlap. Another example is that of the aforementioned Mellinger (1991) model in which frequency partials are tracked from their onset through any adjacent frequency channels to which they may travel.

In terms of performance, Hu & Wang (2007) have proposed a method for assessing segmentation performance. They adapt a method from the field of computer vision and image analysis (Hoover et al. 1996) whereby performance is assessed on the extent to which the ideal segments and the experimental segments overlap. Only target segments are measured, all other segments are labelled as 'ideal background'. A T–F region can be labelled either as correct, under-segmented, over-segmented, missing or mismatch. Metrics can be formed based on this data to show the percentage of segments that fall into each of these categories, which can then be used to compare performance across different systems.

3.4 Scene Organisation

The purpose of scene organisation is to group segments into streams, such that a stream describes an auditory event. This process forms the second stage of Bregman's account of ASA and the fourth stage in a typical CASA system architecture (see Figure 3.1). As Bregman (1990) points out, grouping can be divided in two subsets: simultaneous and sequential. Within these subsets, numerous cues are utilised by the auditory system to achieve grouping. This section will therefore discuss grouping from this perspective and include cues commonly utilised in CASA systems.

3.4.1 Simultaneous Grouping

As stated in Section 2.4.1, simultaneous grouping applies to segments that overlap in time and have arisen from the same sound source. Numerous cues can be utilised to achieve grouping; they will be discussed in this section.

Spatial Location

Feng & Jones (2006) point out that many of the earlier binaural models based on the work of Jeffress (1948) are limited to localisation of one or two sources (e.g. Lyon 1983; Banks 1993) and the performance of these models is often poor for speech. However, more recent work such as that carried out by Liu et al. (2000; 2001) has improved this performance and increased the number of locatable sources from four to six.

However, grouping simultaneously with spatial cues does not occur frequently in the

literature, possibly due to the strength of other cues in this respect, and the fragility of ITD in reverberation. An example of a system that uses spatial cues simultaneously is the system of Nakatani & Okuno (1999), which attempts to segregate two simultaneous talkers. They combine $F0$ analysis with ITD in order to produce a more accurate estimate of pitch. Specifically, they state that:

> The fundamental frequency of each harmonic fragment is calculated more precisely by using only sinusoidal components coming from the same direction as the fragment.
>
> (Nakatani & Okuno 1999)

Harmonicity

For signals with strong harmonic components such as voiced speech, segregation can simply be achieved using a comb filter, provided that the fundamental frequency has already been extracted (see Section 3.2.1). However, it is crucial that the pitch estimate is accurate, otherwise the comb filter could instead destroy the harmonics that it was intended to extract (Wang 2006). Parsons (1976) suggests numerous techniques for resolving overlapping speech. His method has the following stages:

1. *Peak Separation.* Firstly, local maxima in the spectrum are identified as harmonics and added to a peak table which includes estimates of frequency, amplitude and phase. In the case of overlapping peaks, this information will be incomplete. To overcome this, firstly the overlap must be detected via tests of symmetry, distance from adjacent peaks and 'well–behaved' phase. Secondly, the components must be separated. Prior knowledge of the peak shapes is combined with the simple additive nature of the overlap to calculate each harmonic's shape and thus extract it.

2. *Pitch Extraction.* This stage uses an adaptation of Schroeder's (1968) method (described in Section 3.2.1, page 30). Following this, the peaks of the peak table are assigned by comparing the values to predicted values based on the fundamental frequency that has just been estimated.

3. *Tracking.* Once the harmonics of each talker have been established, it is necessary to track them so that the same talker is followed throughout. This is done by assuming that the pitch will not change much from each 51 ms segment to the next. Specifically, each harmonic set (voice) is assigned to a 'track' and values for subsequent segments are predicted so as to increase the likelihood of following the same voice.

Numerous models for the separation of simultaneous sounds based on harmonicity have been suggested (e.g. Weintraub 1985; Cooke 1991; Brown & Cooke 1994a; Hu & Wang

2004b) and all work in slightly different ways. For example, Cooke's (1991) approach works by first calculating "synchrony strands". These are his proposed time–frequency representation which computes frequency, amplitude and AM rate along the strand's length by applying local constraints of similarity and continuity to the output of a cochlear model. These strands are then grouped based upon harmonic relations (and also AM similarities).

The model of Hu & Wang (2004b) performs grouping in speech differently depending on whether the harmonics are resolved, i.e. harmonics are often unresolved at higher frequencies (Wang 2006). Furthermore, Hu & Wang (2004b) define a harmonic as resolved if an auditory filter channel responds primarily to it, otherwise it is unresolved. For resolved harmonics, an initial grouping is carried out based on the dominant pitch in each time frame using the oscillatory correlation model of Wang & Brown (1999). Specifically, a comparison is made between correlogram response of the T–F unit and the dominant pitch per frame and grouping is done on this basis. This grouping is used to estimate a pitch track for the target sound source. Higher frequencies are grouped using an AM criterion, which will be discussed later in this section.

Common Onset and Offset

The model of Brown & Cooke (1994a) includes grouping by common onset and offset. In their model, the tracking algorithm has a preference for breaking an auditory element rather than making a tracking error. In this case, the start and end of an auditory element do not necessarily correspond to an onset and offset. However, during the feature extraction process, onsets and offsets are plotted on a map. They use the following logic to decide on grouping:

> Auditory elements which start or end synchronously are more likely to form a group, providing that there is sufficient activity in the onset and offset map at the appropriate time.
>
> (Brown & Cooke 1994a)

However, they find that onsets and offsets are rarely exactly synchronous and allow a tolerance of 20 ms difference. Acoustic elements are subsequently checked against the onset/offset map to check that the element has actually started and/or stopped. Again, Brown & Cooke allow a tolerance due to the impulse response of the filters; as before this tolerance is set to 20 ms. Furthermore, Brown & Cooke state that the grouping of acoustic elements based on their onset or offset time is not guaranteed, which is in keeping with data provided by Darwin & Sutherland (1984). If both are similar then the likelihood is increased, but also similarity in $F0$ contour has a multiplying effect and dramatically increases the chance of grouping if it demonstrates significant similarities between elements.

AM and FM

As stated above, the model of Hu & Wang (2004b) treats resolved and unresolved harmonics differently. To group unresolved harmonics—which are typically found above about 1 kHz—the appropriate regions of the cochleagram are first segmented by common AM rates and temporal continuity. Subsequent grouping is then performed based on these AM rates.

The model of Brown & Cooke (1994a) incorporates FM analysis. The model computes a number of auditory maps for cues such as FM, pitch and onsets/offsets from the cochleagram. Segments are formed based on smoothly varying spectral peaks in the FM map and frequencies where there is high cross-channel correlation. Grouping is performed by first summing the correlogram responses of the segment within each time frame. Following this, a dynamic algorithm is used to identify a pitch contour over the segment based on the summed correlogram response. This pitch contour is then compared to neighbouring contours and grouped accordingly.

3.4.2 Sequential Grouping

As stated in Section 2.4.2, sequential grouping applies to segments that do not overlap in time but are likely to have arisen from the same sound source. Sequential grouping follows on from simultaneous grouping by aiming to join these groups into continuous streams (Wang 2006). Numerous cues can be utilised to achieve grouping; they will be discussed in this section.

Pitch

Grouping by pitch contour was first demonstrated by Atal (1972) in the context of ASR. This can only apply to continuous voiced pitch tracks since un-voiced sounds will break the pitch track. But pitch is useful for segregating voices with very different pitch ranges such as male and female speakers (e.g. Weintraub 1985). However, in most scenarios, pitch range will vary considerably and pitch tracks are likely to overlap (Wang 2006).

Consequently, Shao & Wang (2006) perform sequential grouping of two competing talkers using the pitch track generated by the algorithm of Wu et al. (2003). Firstly, overlapping sections are removed since they are not useable. Subsequently, the algorithm must decide on the grouping of consecutive pitch tracks. This is performed using two criteria: the frequency difference between the final pitch of the first track and the initial pitch of the following track and on the time gap between the tracks. The bigger either of these metrics are, the less likely it is that the two tracks belong to the same sound source or voice. The thresholds are set by training the model to estimate the distribution parameters. Thereafter it is simply a binary decision as to whether a

pitch track is grouped with an adjacent track.

Spectral Content

Spectrum-based sequential grouping assumes that the spectral properties of a speaker are more similar across time for the given speaker than they are similar to a different speaker. Hence, the key to grouping sequentially is in obtaining a robust measure of the properties of the spectrum that can be used for comparison across frames or segments (Wang 2006). Such a system was proposed by Morgan et al. (1997), which is designed to separate two simultaneous talkers. The problem is formed as one of speaker assignment, whereby in each frame, each voice must be assigned to one of the output channels. They hence borrow a method of spectral comparison originally devised by Carlson & Clements (1991) which makes comparisons of the current frame against the last 50 frames that contained voiced sounds. The comparison is a measure of divergence based on autocorrelation, Linear Prediction (LP) coefficients and residual energy.

For musical signals, an approach was proposed by Brown & Cooke (1994b) based on timbre. Despite the difficulties in defining 'timbre', Brown & Cooke use only two dimensions to measure it: brightness, which is a measure of the spectral centroid taken from the correlogram, and onset asynchrony, which is a measure of the relative differences in onset time of frequency partials belonging to a continuous stream. These two dimensions are then clustered for each group and comparisons of groups across time can be performed on these clusters. Godsmark & Brown (1999) took this approach one step further by proposing a "timbre track". This timbre track plots changes in spectral centroid against changes in amplitude. Sequential grouping is then simply performed by comparing these timbre tracks.

Spatial Location

Sequential grouping by spatial location usually occurs through spatial subtraction in which interfering noise sources are subtracted from the total sound, thus enhancing the target. One such model is that of Lockwood et al. (2004), which utilises ITD. Their system is a variation of the adaptive beamforming technique and utilises only two microphones. Specifically, their technique is a minimum variance distortionless response beamformer which works by constraining the combining weights such that there is no change in gain or phase (hence distortionless) and minimising the average energy of the output (hence minimum variance). Minimisation is achieved by computing a 2×2 correlation matrix for each frequency band which is updated every N samples, thus allowing signals to be tracked quickly within each band. Feng & Jones (2006) show that this system is more efficient than the system of Liu et al. (2001).

The ILD cue is utilised in another model by Lockwood et al. (2003) which employs two or more very directional microphones that thus elicit a high ILD. Since this

level difference is a result of the directivity of the microphone(s), this change in level consequently indicates direction. The algorithm is essentially the same as that of Lockwood et al. (2004) except that the steering vectors are calculated using the level difference rather than the phase difference.

One last notable model is that of Roman et al. (2003), which incorporates both ITD and ILD cues generated through convolution of signals with a Head–Related Impulse Response (HRIR). ITD is extracted using the cross-correlation model described above; ILD is extracted by calculating the ratio of signal power at the two ears for each frequency channel. They find that there is a strong correlation between the relative strength of signals in a mixture and the estimated ITD/ILD and that the ITD and ILD can be seen to cluster for each frequency channel when compared in a 'binaural space'. Consequently, they use a nonparametric classification method to estimate a binary mask that is subsequently used to separate the target and interfering sound sources.

3.5 Re-synthesis

The final stage of a typical CASA system, according to Figure 3.1, is to re-synthesise the audio waveform from a group of segments that hopefully originated from the same physical source. Re-synthesising the waveform allows the performance of the system to be assessed through subjective means or by measuring physical parameters such as changes in signal-to-noise ratio (Wang & Brown 2006).

Re-synthesis is typically achieved by inverting a time–frequency representation such as synchrony strands (e.g. Cooke 1991) or some other representation such as the correlogram (e.g. Slaney et al. 1994). For systems that use T–F masking, re-synthesis of the target waveform from an auditory filterbank output is relatively straightforward. The process is described by Weintraub (1985) and Brown & Cooke (1994a):

1. Phase discrepancies arising from the filter must be removed if the phase–corrected filter has not been used (see Equation 3.2, page 21). To achieve this, the filter response of each channel must be time–reversed, passed through the filter and then time–reversed again.

2. The outputs of each filterbank channel are windowed into lengths equivalent to the length of the T–F units. Windowing is achieved with a raised cosine.

3. The level of each windowed T–F unit is then weighted by the corresponding value of the T–F mask (either real or binary).

4. The weighted channels are then summed to produce the reconstructed waveform.

3.6 Summary

The aim of this section was to explain some common CASA techniques and their respective computational implementation. The mechanisms have been presented in terms of a typical CASA system architecture. In this typical architecture, the first stage is to model the peripheral processing of the auditory system. The gammatone filterbank was presented as a commonly employed basilar membrane model due to its correlation with physiological data and computational efficiency. Following this, Meddis et al.'s (1990) model of the IHCs was shown to be a popular computational model of the conversion from basilar membrane displacement to neural activity. The neural activity data can then be used to extract other cues such as periodicity via the correlogram and ITD via the cross-correlogram. Visual representations of these data can also be formed such as the cochleagram, which is a spectrogram–type plot of neural activity.

From these data, other acoustical cues or *features* can be extracted such as AM, FM, onsets, offsets and cross-channel correlation. AM is calculated by simply extracting the amplitude envelope of each channel. FM is calculated either by convolving the cochleagram with a two-dimensional Laplacian-of-Gaussian function or by calculating variations in IF. Onsets and offsets are extracted by simply differentiating the amplitude envelope and smoothing the result to remove spurious peaks caused by other noises or voice fluctuations. The onset or offset is then identified by the differential crossing a pre-defined threshold value.

Thereafter, mid-level representations are created as an intermediate step between T–F units and groups; this process can be considered as the final computational version of Bregman's first stage of ASA (segmentation). The segment is commonly used due to its perceptual relevance. A segment is a continuous region of the cochleagram that collects adjacent T–F units belonging to a single sound source; they are created monaurally. Subsequently, scene organisation—equivalent to the second stage of Bregman's framework—attempts to collect segments together to form streams such that a stream represents the sound originating from a single sound source. As in ASA, acoustical cues are used to inform grouping and grouping takes place both simultaneously and sequentially. Specifically, segments can be grouped according to spatial location, spectral content, common onset and offset, AM, FM and pitch. Finally, once the T–F mask has been calculated the waveform needs to be re-synthesised from the filterbank output by applying the mask.

Chapter

4 CASA in Reverberant Environments

The previous chapters have provided some important background information on human auditory perception, ASA and machine source separation (in terms of CASA). This chapter will address the following questions that were given in Section 1.6:

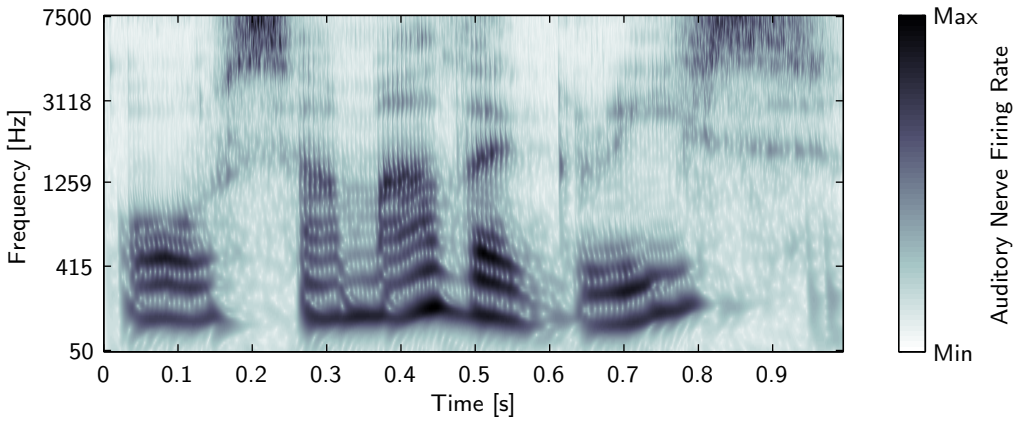
1. What are the problems posed by reverberation to human auditory perception in general?
2. What are the problems posed by reverberation to machine listening in general?
3. What are the human solutions to reverberation?
4. What are the machine listening solutions to reverberation, in particular in terms of source separation? How do machine listening solutions relate to human solutions?
5. Which reverberant source separation solution has most scope for improvement?

In response to these research questions, Questions 1 and 2 are addressed in Section 4.1 and Section 4.2 respectively, which present issues posed by reverberation to human auditory perception and machine listening. Section 4.3 addresses Question 3, where human solutions to reverberation are presented. Question 4 is addressed in Section 4.4, where machine solutions to reverberation are presented and their relation to human solutions is discussed. Lastly, the findings of the chapter are summarised in Section 4.5 where the machine solution that is shown to have the most scope for improvement is selected for further work.

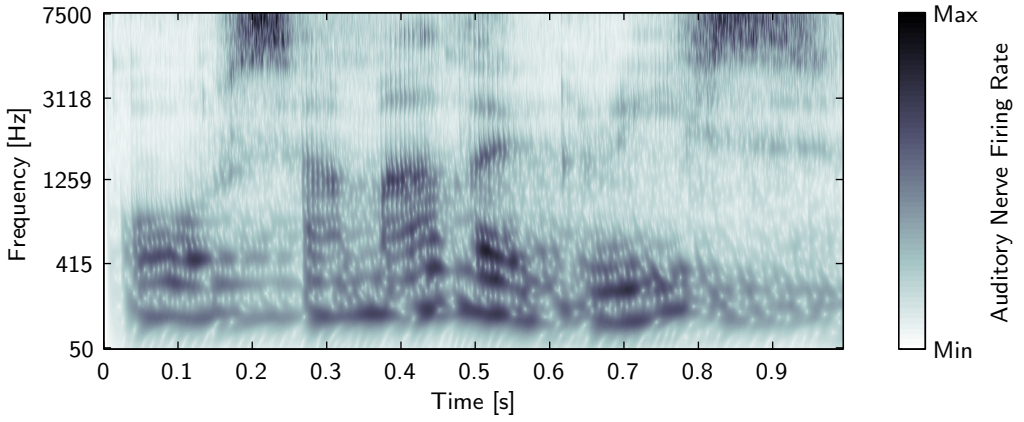
4.1 Reverberation Issues: Human

1. What are the problems posed by reverberation to human auditory perception in general?

Reverberation presents numerous problems to human auditory perception including degradations in speech perception, source separation and sound localisation (Brown & Palomäki 2006). These effects are discussed in this section.



(a)



(b)

Figure 4.1: Cochleagram of reverberated speech. **(a)** Cochleagram for the clean female speech "or some other grease" (as seen in Figure 3.5(b), page 26). **(b)** Cochleagram for the same speech convolved with a room impulse response with an RT_{60} = 0.68 s and C_{50} = 17.4 dB (Room C, Table 5.1, page 77).

4.1.1 Speech Perception

The reverberation of a typical room causes speech, or indeed any other signal, to be smeared across time, as can be seen in Figure 4.1. Several observations can be made from this figure. Firstly, areas that were once devoid of speech energy, such as the gaps between words, now contain reverberant energy that has been smeared from the preceding sound. Consequently, the onsets that follow are masked by the reverberant energy and are not as prominent. In the study by Gelfand & Silman (1979), it was found that small room reverberation (Reverberation Time (RT_{60}) (to -60 dB relative to the direct sound) ≈ 0.8 s) had a significant impact upon listeners' ability to recognise articulation and stop and frication consonants, i.e. sounds where onsets and offsets are important. On the other hand, sibilance and semivowel sounds, i.e. sound where onsets and offsets are less important, were barely affected. This effect is known as overlap–masking, which is the masking of the onset of an utterance by the reverberant energy of the previous utterance (Libbey & Rogers 2004).

A second observation that can be made of Figure 4.1 is that formant transitions are blurred and almost indistinguishable. This effect is referred to as self–masking. Bolt & MacDonald (1949) explain self–masking as the blurring of a given utterance by the reverberant energy arising from the onset of the utterance, thus causing such transitions and onsets and offsets to be smeared. However, both Bolt & MacDonald (1949) and Libbey & Rogers (2004) assume that overlap–masking is the most detrimental of the two. Bolt & MacDonald (1949) argue this for two reasons: firstly, because the initial part of a speech sound contains most of the information that is crucial for intelligibility and secondly, because of the observation that intelligibility is improved at slower speaking rates.

Three more observations are made by Brown & Palomäki (2006) that are demonstrated in Figure 4.1. The first is that the voice onset has become blurred; specifically the gap between a stop release and a voiced sound, such as the "g" of "grease" at about 0.62 s. The second observation is that noise–like sounds of affricates and fricatives are somewhat extended, such as the /s/ of "some" at about 0.15 s. Lastly, reverberation can be seen to mask amplitude modulations of harmonics which appear as vertical bands in the cochleagram.

Brown & Palomäki (2006) state that the commonly acknowledged structure of reverberation, consisting of early reflections and dense, late reverberations, has a two–sided effect on our perception of speech. The early reflections help to reinforce speech by effectively amplifying it (although they do introduce a level of comb filtering which can be detrimental), whilst the dense late reflections, being poorly correlated to the speech, act as additive noise that is detrimental. Lochner & Burger (1964) introduced the concept of the ratio of useful-to-detrimental sound energy. This effectively equates

to a measure of the ratio of early-to-late-arriving energy. Such a ratio has been defined in BS EN ISO 3382: 2000:

$$C_{te} = 10 \log_{10} \left(\frac{\int_0^{te} P^2(t)\,\mathrm{d}t}{\int_{te}^{\infty} P^2(t)\,\mathrm{d}t} \right) \mathrm{dB} \tag{4.1}$$

where $P(t)$ is the instantaneous sound pressure of the room impulse response at time t, C_{te} is the early-to-late or clarity index (in dB) and te is the early time limit and is set according to the signal that is to be assessed, being 50 ms for speech and 80 ms for music (BS EN ISO 3382: 2000).

4.1.2 Source Segregation

Reverberation affects most of the cues utilised by humans to group perceptual segments (Brown & Palomäki 2006). This subsection will discuss how reverberation affects these cues.

Fundamental Frequency

Culling et al. (1994) state that grouping by harmonicity depends upon the extent to which the $F0$ fluctuates. As stated in the previous section, reverberation causes sounds to be blurred across time. For harmonicity cues, this will have little effect if the harmonics are stationary. However, if the harmonics fluctuate, as they do in natural speech, then the strength of the harmonicity will be reduced as the harmonic structure is smeared. In one of their experiments, Culling et al. tested this hypothesis by measuring their listener's ability to separate a target vowel stimulus from a vowel–like masker—with and without reverberation—where the target had either a static or fluctuating $F0$. They found that when reverberation was added, there was no change in the listener's ability to separate the target vowel for a static $F0$. However, in the second part, the $F0$ of the target was sinusoidally modulated, which resulted in the listeners being unable to separate the second vowel due to self– and overlap–masking (Culling et al. 1994).

In a later study Culling et al. (2003) extended their paradigm to running speech, this time testing the intelligibility of monotonous speech and naturally intonated speech of a target male voice in the presence of interfering female speech. From the above study, Culling et al. predicted that the monotonous speech would be more intelligible in reverberation than the intonated speech, due to the same issue of $F0$ fluctuations. However, this was not found to be the case: the monotonous and naturally intonated speech were equally intelligible in reverberation. Interestingly, the monotonous speech was less intelligible in the anechoic condition and hence the natural speech demonstrated a greater drop in intelligibility when reverberation was introduced.

Binaural Cues

In reverberant conditions, the distribution of ILD and ITD cues are broadened, due to the late sound from reflections. Many authors (e.g. Plomp 1976; Darwin & Hukin 2000; Culling et al. 2003) agree that reverberation disrupts our ability to separate concurrent sounds using spatial or binaural cues. Plomp (1976) investigated the binaural advantage (i.e. the advantage of binaural listening over monaural listening; see Section 4.3.3) by measuring the intelligibility of speech against a spatially separated speech masker. For anechoic conditions, an intelligibility gain of 4–5 dB was observed whilst for reverberant conditions with an RT_{60} of just 0.4 s, the intelligibility gain was lowered to 2–3 dB. The intelligibility gain was further lowered as the RT_{60} was increased. Furthermore, ITD is found by Culling & Summerfield (1995) and Drennan et al. (2003) to be only a weak cue for simultaneous grouping in reverberation, due to the altered interaural coherence. However, Culling et al. (1994) state that ITD may maintain its usefulness in reverberation as a sequential grouping cue by directing auditory attention to specific spatial locations.

Common Onset and Offset

As discussed in the previous section, reverberation serves to smear what is perceived by the ear. This results in troughs in the temporal envelope being filled with reverberant energy. Consequently, stronger onsets are maintained whilst weaker onsets are likely to be masked by the reverberation. Offsets tend to be almost completely masked due to the temporal decay of the reverberation. This means that whilst grouping by common onset maybe preserved (depending on the magnitude of the onset and the preceding energy), grouping by common offset is likely to be unreliable (Brown & Palomäki 2006).

4.1.3 Sound Localisation

Localisation performance in a reverberant environment depends upon numerous factors such as the position of the listener and the nature of the stimulus. Both Hartmann (1983) and Giguère & Abel (1993) found that broadband noise was much more difficult to locate in reverberation than in an anechoic environment. Hartmann also found that spectral density was an important factor for localisation, with broadband noise being easier to locate than spectrally sparse complex tones. Also, localisation performance decreased for broadband noise as reverberation time increased. However, Hartmann also notes that stimuli with strong attacks were easily located independently of reverberation time.

4.2 Reverberation Issues: Machine

2. What are the problems posed by reverberation to machine listening in general?

Reverberation presents numerous problems for machine listening. These problems relate to the extraction of features such as pitch tracking, binaural cues, onsets and offsets, and hence affect applications such as ASR. These problems will be discussed in this section.

4.2.1 Feature Extraction

Pitch Tracking

The YIN algorithm of De Cheveigne & Kawahara (2002) was tested under reverberant conditions by Brown & Palomäki (2006). The addition of moderate reverberation (RT_{60} = 0.5 s, Direct-to-Reverberant Ratio (DRR) = −3 dB) was seen to have two effects. In cases where the pitch is strong, the pitch track was seen to be extended slightly due to the pitch tracking algorithm following the reflected sound. Where the pitch is weaker it tends to be corrupted and/or masked by the reverberation and is lost completely. Roman & Wang (2005) have shown that the deterioration of periodicity in reverberation can have a large detrimental effect on separation performance, even for relatively small RT_{60}s, with the Signal-to-Noise Ratio (SNR) dropping by as much as 8 dB between 0 and 0.35 s (see Figure 4.2). Furthermore, they show that periodicity can only be a robust cue if steps, such as inverse filtering, are taken to undo the effects of the reverberation.

Binaural Cues

Some work has been done on developing binaural CASA systems that aim to separate the target from a spatially separated noise or interferer (e.g. Lyon 1983; Bodden 1993; Liu et al. 2001; Palomäki et al. 2004b). These systems have been tested in reverberant conditions. Whilst they offer some robustness to reverberation, their performance is seen to degrade as either the RT_{60} or DRR is increased. For example, the model of Palomäki et al. (2004b) shows a drop in ASR accuracy of 38% between RT_{60}s of 0 and 0.3 s. This is because these systems need to localise each sound source in order to separate them. As previously noted, reverberation causes the distribution of spatial cues to be blurred. Consequently, localising the sound sources is more difficult and separation is less successful. This was demonstrated in a study by Woodruff & Wang (2010) in which they found that the average azimuth estimation error increased from 1° in an anechoic environment to as much as 10° for an RT_{60} of 0.8 s. This resulted in a drop in T–F unit labelling accuracy of 25% over this range of RT_{60}s. However, many studies (e.g. Faller & Merimaa 2004; Palomäki et al. 2004b; Woodruff & Wang 2010)

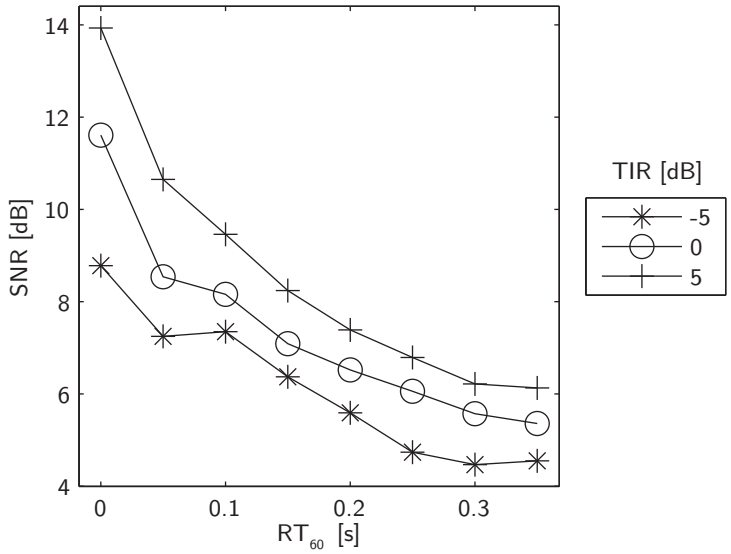


Figure 4.2: Separation based on pitch tracking in reverberation, from Roman & Wang (2005). Results show the Signal-to-Noise Ratio (SNR) for three Target-to-Interferer Ratios (TIRs) (i.e. the ratio in dB of the target and interferer sound sources).

have shown that appropriate cue selection can significantly improve the robustness of binaural cues to reverberation.

Onset and Offset Detection

As may be inferred from discussions in the previous section with regard to perceptual onset and offset detection, reverberation can be very detrimental to automated onset and offset detection. As shown in Section 3.2.3, onsets and offsets are detected by differentiating the signal envelope; they are high frequency components of the envelope. However, reverberation acts as a low-pass filter on the temporal envelope (discussed further in Section 4.3.1), thus reducing their magnitude. Brown & Palomäki (2006) show that whilst some strong onsets are retained, most offsets are masked by the reverberation. This is demonstrated in Figure 4.3[1], which shows onsets (in white) and offsets (in black) for the signals plotted in Figure 4.1, i.e. anechoic female speech (Figure 4.3(a)) and the same female speech convolved with a room impulse response with an $\mathrm{RT}_{60} = 0.68$ s and DRR ≈ 9 dB (Figure 4.3(b)). The addition of reverberation has resulted in 60% of units labelled as onsets being retained, compared with only 32% of units labelled as offsets. In the model of Hu & Wang (2007), segmentation is performed by onset and offset analysis. Whilst their system is untested in reverberation (but works well anechoically), they point out that the information provided by onsets and offsets is likely to break down in reverberant conditions. For these reasons, few

[1]The figure was calculated from the mean of the Hilbert envelope for each 10 ms frame and then calculating the difference in dB between adjacent frame. Onsets are indicated where the difference exceeds 8 dB, offsets are indicated where the difference is less than −8 dB.

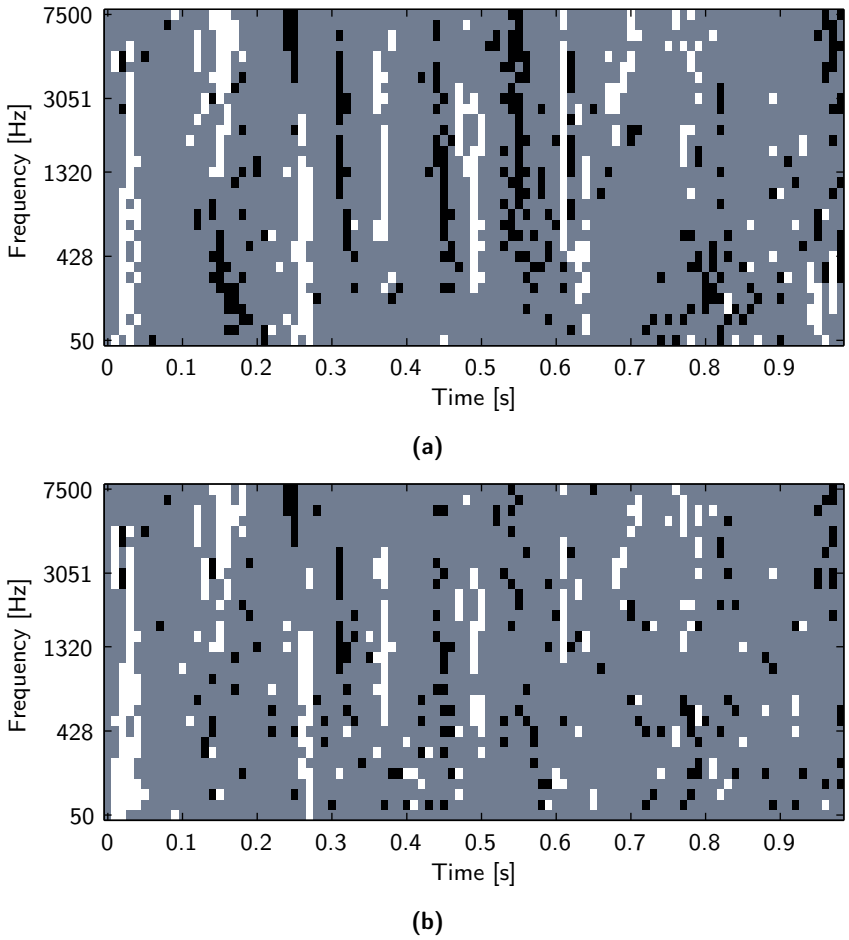


Figure 4.3: Onset and offset detection in reverberation. **(a)** Onsets (white) and offsets (black) for anechoic female speech (as in Figure 4.1(a)). **(b)** Onsets and offsets for the same speech convolved with a room impulse response with an $RT_{60} = 0.68$ s and DRR ≈ 9 dB (Room C, Table 5.1, page 77) (as in Figure 4.1(b)). The plot shows that whilst some strong onsets have been retained, many offsets are masked by the reverberation.

systems utilise onsets and offsets in reverberation, although the system proposed by Palomäki et al. (2002) attempts to extract onsets in reverberation in order to produce a reverberation mask (see Section 4.4.4).

4.2.2 Automatic Speech Recognition

The impact of reverberation on acoustic features has a resultant effect on ASR, which relies to some extent on the extraction of some of these features. Specifically, most modern ASR systems are based on Hidden Markov Models (HMMs). In these systems, the HMM creates a series of vectors for each word or phoneme. These vectors are based on cepstral coefficients calculated over short time frames of 10–20 ms (Jurafsky & Martin 2009). Most systems work by matching a speech input to an acoustic model of each type of speech sound. The acoustic model is often trained on anechoic speech.

Hence it is clear that reverberation will have a significant impact on these systems because it affects the cepstrum and temporal envelope of the speech, reducing its similarity to the acoustic model and thus reducing the likelihood that the word or phoneme will be recognised (Brown & Palomäki 2006).

Generally, most studies agree that reverberation, and in particular late reverberation, is very problematic to ASR because the distortions, and in particular temporal smoothing, extend over several analysis frames. Early reflections can be beneficial, by boosting the level of the direct speech, and spectral changes can be counteracted by single–frame processing techniques (Brown & Palomäki 2006). The literature discusses the effects of three reverberation parameters on ASR performance: RT_{60}, early and late reflections and DRR. These three parameters and their corresponding effects are discussed below.

Reverberation Time

The effects of reverberation time on speech recognition are perhaps the most documented and most studies suggest that ASR performance decreases as RT_{60} increases (e.g. Giuliani et al. 1996; Kingsbury 1998; Eneman et al. 2003; Couvreur & Couvreur 2004; Palomäki et al. 2004a). The findings of these numerous researchers are summarised in Figure 4.4(a). There is a reasonable degree of variability in these results for two reasons: firstly each experiment uses a different speech recogniser and/or speech corpus and secondly, there is a wide range of dates and hence technological advancement in the algorithms employed. However, they all have a consistent trend of reducing accuracy with increasing RT_{60}. It is likely that, as discussed above, the temporal smoothing caused by reverberation leads to overlap- and self-masking, creating a significant difference between the acoustic model and the separated speech.

Early and Late Reflections

Gölzer & Kleinschmidt (2003) investigated the relative effects of early and late reflections on ASR. Specifically, they attempted to modify Equation 4.1 to determine a value for *te* such that room reflections arriving after this time would be detrimental to ASR accuracy. They carried out tests on a range of impulse responses that were modified in one four ways: 1. *Growing gap*, reflections starting 5 ms after the direct sound up to a variable time are muted; 2. *moving gap*, a 5 ms muted region is introduced at the variable time starting from 5 ms after the onset; *cutting tail*, the impulse response is muted after the variable time; 4. *filling gap* the region from 5–100 ms is initially muted, and un-muted up to the variable time. They found that *te* was in the range 25–50 ms, depending on the room impulse response and the specific recogniser. However, it is also noteworthy that this effect was generally relatively small, often not affecting ASR accuracy by more than 10%.

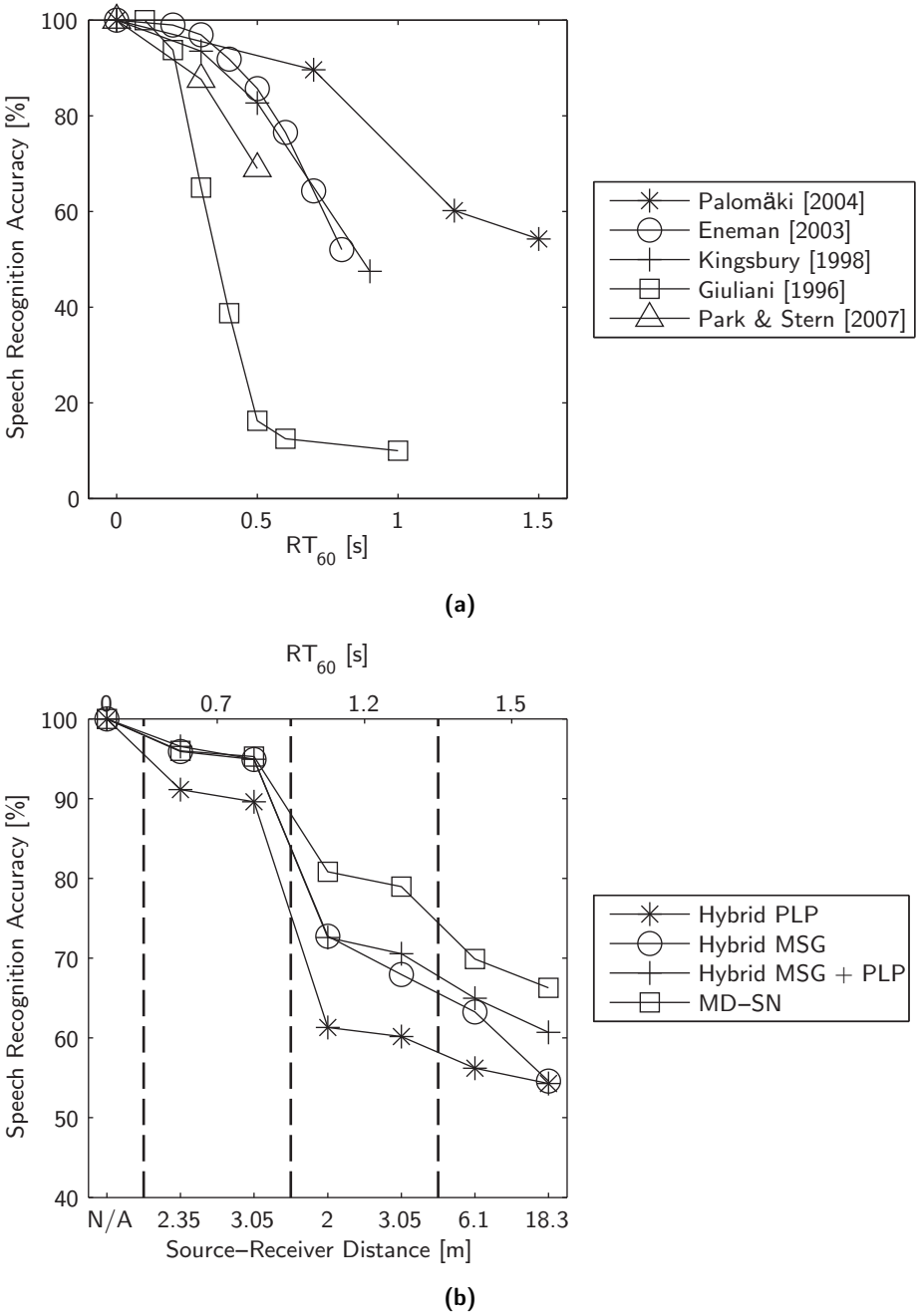


Figure 4.4: Speech recognition performance in reverberant conditions. **(a)** Speech recognition accuracy for varying RT_{60}. Due to the differences in implementations between the studies, the data have been normalised to 100 % for $RT_{60} = 0$ s. **(b)** Speech recognition accuracy for varying DRR. Adapted from (Palomäki et al. 2004a).

Direct-to-Reverberant Ratio

Numerous studies (e.g. Gillespie & Atlas 2002; Palomäki et al. 2004a) also state that recognition accuracy falls with DRR for a constant RT_{60}, although this effect is less significant than altering the RT_{60}. The results of Palomäki et al.'s (2004a) experiments, in which they tested numerous ASR systems for a variety of RT_{60} and distances from the microphone (i.e. changing the DRR), are shown in Figure 4.4(b). Notice that recognition accuracy falls more dramatically for changes in RT_{60} than for changes in DRR. The result is expected given previous observations. Specifically, reducing the DRR of the reverberation will not increase the duration of the self- and overlap-masking—whilst it increases the masking effect, the overall impact of this will be partially balanced by the increase in level of the early reflections that, as previously noted, are conducive to speech recognition. In contrast, increasing the RT_{60} increases the duration of the self- and overlap-masking.

4.3 Reverberation Solutions: Human

3. What are the human solutions to reverberation?

The problems posed to humans by reverberation have been established in the previous sections of this chapter and include issues of speech perception, source separation and sound localisation. Humans have numerous mechanisms that can be utilised in order to be robust to reverberation and remain effective in areas such as speech perception. Such mechanisms include utilising the slow temporal modulation of speech, the binaural advantage, spectral envelope distortion compensation and precedence. These mechanisms will be discussed in this section.

4.3.1 Utilising Slow Temporal Speech Modulation

The perception of speech by humans is remarkably tolerant to reverberation. Numerous researchers have commented on this including Dudley (1939), Houtgast & Steeneken (1973; 1985) and Drullman et al. (1994a; 1994b). These studies have found that the slow temporal modulations of speech are robust to the effects of reverberation and that humans are able to take advantage of this. Specifically, Houtgast & Steeneken (1985) found that the modulation rates in speech are in the range 1–16 Hz, with a strong component at 4 Hz that is found to reflect the syllable rate. As mentioned above, reverberation is seen to smooth the temporal envelope of a signal that it affects and can be considered as a low-pass filter (Houtgast & Steeneken 1973). Hence, reverberation often has little effect on the temporal envelope at these frequencies since they fall into the pass-band. Houtgast & Steeneken (1985) take this principle one step further by predicting speech intelligibility in a room based on the room's Modulation Transfer

Function (MTF). According to Houtgast & Steeneken (1973), the MTF of a room is established using a signal with a sinusoidal envelope. Reverberation hence only affects the modulation depth; the amount of change is dependent on the modulation frequency. The MTF is simply the change in modulation depth as a function of frequency. The prediction of speech intelligibility—the Speech Transmission Index (STI)—is derived from the MTF (Steeneken & Houtgast 1980).

4.3.2 Spectral Envelope Distortion Compensation

It was noted in Section 4.1.1 that reverberation, particularly due to early reflections, introduces comb filtering which distorts the spectrum of the reverberated signal. According to Watkins (1991), the spectral envelope is a major factor for identifying many types of sound, including speech. Watkins presents a series of experiments aimed at establishing the nature of the perceptual mechanism that has been shown to counteract the spectrally distorting effects of reverberation (see also Haggard 1974; Repp 1987; Summerfield et al. 1987). These experiments distorted the spectral envelope of a carrier sound (a short sentence) followed by a test word taken from a continuum between /itch/ and /ɛtch/. The magnitude of the spectral distortion compensation was hence quantified in terms of the phoneme boundary shift. Several key observations arise out of these experiments:

- Perceptual compensation was maintained when a change of talker was suggested by the stimulus
- Perceptual compensation was reduced when there was a change in the characteristics of the distorting channel
- Speech carriers gave larger boundary shifts than noise carriers

Watkins concludes that these distortions are likely to be counteracted in the central auditory system, rather than in the periphery, through a mechanism that assesses the rate of spectral change. He argues that the rate of spectral change in the transmission channel (e.g. the room) is likely to be much slower than the rate of change for the sound source (e.g. speech), whose spectral content is likely to change frequently (Watkins 1991).

4.3.3 The Binaural Advantage

The literature gives several advantages of binaural over monaural listening. Koenig (1950) provides some of the earliest examples, obtained by providing a listener with stereo headphones connected to either one or two microphones such that the listener may choose between the two configurations. He makes several interesting observations:

- Using the binaural system the listener is able to subjectively suppress the reverberation present during monaural listening
- The listener is able to separate conversations using the binaural system whereas they are heard as a "hopeless jumble" with one microphone
- Listeners are able to recognise speech in extremely high noise conditions with the binaural system whereas the speech was lost during monaural listening
- In the experiment discussed in the previous section conducted by Watkins (1991), the phoneme boundary shift is smaller when the carrier and test sound are presented to different ears

It will be shown in Section 4.3.4 that precedence may account for some of these effects by weighting the first few wavefronts and hence suppressing a majority of the reverberation. Additional mechanisms have been discussed in Section 4.1.2 and involve utilising multiple cues including harmonicity, ITD, ILD and common onsets and offsets. The auditory system may also utilise instant-by-instant differences in SNR for each ear such that the "better ear" can be weighted over the ear with the worse SNR (Devore & Shinn-Cunningham 2003).

Another well–documented binaural advantage is the improvement in our ability to detect masked signals compared to monaural presentation. Moore (2004) summarises the effect in the following way. A noise signal and a sine tone are replayed through headphones and identical in both ears. The level of the sine tone can be lowered until it becomes inaudible because it is masked by the noise; the masking threshold is denoted L_0. If the phase of the tone is inverted in one ear, it suddenly becomes audible. The level of the tone can be lowered again until it is masked by the noise; the new masking threshold is denoted L_1. This difference in masking threshold, $L_0 - L_1$, is referred to as Binaural Masking Level Difference (BMLD). At low tone frequencies (c. 300 Hz), BMLD can be as high as 15 dB, reducing to 2–3 dB above 1500 Hz (Durlach & Colburn 1978). Similar effects have been observed with other types of masked signal, by introducing other phase shifts, and by presenting the stimulus to one ear only (Moore 2004). Finally, Moore points out that, in reality, such differences in interaural phase only occur when the signal and masking noise are located in different spatial locations. Hence, at least some of these phenomena are closely related to localisation and to the ASA problem.

4.3.4 The Precedence Effect

The term "precedence effect" was originally coined by Wallach et al. (1949). They use this term to refer to our apparent ability to locate sounds in reverberant environments that present very difficult circumstances. Wallach et al. are referring to the fact that

in an enclosed environment, a listener not only receives the direct sound from the sound source, but also numerous reflections propagating from the surfaces of the room. These reflections—the strength of which are determined by the size, shape and surface materials of the room—conspire to provide conflicting cues that should make locating the sound quite difficult. However, as Wallach et al. point out, this is not the case and more often than not, localisation of sounds within a room is perfectly possible.

Although it is well documented that precedence is an essential mechanism that assists localisation of sounds in reverberation, it remains unclear whether it plays a role in other areas of auditory perception. For example, precedence may account for at least some of the binaural advantages due to the suppression of reflections. Continuing his work on spectral envelope distortion compensation, Watkins (1999) investigates whether a precedence–type effect is present in the perception of vowels: whether the spectral shape is derived from the first wave front. The experiment left the first portion of the sound (equivalent to the direct sound) un-filtered whilst filtering the remainder of the sound (equivalent to reflected sound); ITD was also changed separately for the direct and filtered sounds. Whilst they confirmed the precedence effect in terms of localising the direct sound, no evidence was found to suggest a mechanism that determined the spectral envelope from the direct sound alone. Furthermore, Watkins points out that the filtered part of the sound had a significant influence on the perceived identity of the vowel. Brown & Palomäki (2006) point out that such an outcome makes sense since reflections do contain useful information about the speech (see also Libbey & Rogers 2004).

Litovsky et al. (1999a) have given a comprehensive overview of the precedence effect and an insight into some of the physiological and perceptual mechanisms behind it. They list several important observations arising from experiments involving just two clicks (a 'lead' and a 'lag') investigating the precedence effect: fusion, localisation dominance and lag discrimination suppression. A point of reference that will be used throughout this discussion is the *echo threshold*. This is the point at which the lag sound no longer perceptually fuses with the lead sound and becomes a separate auditory event; for transient signals this usually occurs for a lead–lag interval range of about 3–5 ms, although some estimates vary in the range 2–50 ms, depending on a variety of factors (see Table 4.1). These factors include relative amplitude, the nature of the stimulus, spatial separation and listener instruction. It is important to note that the echo threshold is not equivalent to the threshold of detectability, since a lag can be detected based on any aspect of overall sound quality. Note that the term 'lead' is used here instead of 'sound source' and 'lag' instead of 'reflection' since the latter terms are reserved by Litovsky et al. for 'real' acoustic environments or 'real world' experimental applications; this document will adopt these conventions. An overview of some pertinent precedence figures is given in Table 4.1 (Litovsky et al. 1999a).

Table 4.1: Experimental thresholds for precedence effects using different stimuli types. Adapted from (Litovsky et al. 1999a).

Study	**Stimulus**	**Threshold [ms]**	**Criterion**
Fusion echo thresholds			
Haas (1951)	Speech	30–40	"Echo annoying"
Lochner & Burger (1958)	Speech	50	Lead and lag "equally loud"
Schubert & Wernick (1969)	Noise		
	a) 20 ms duration	5–6	Lead and lag "equally loud"
	b) 50 ms duration	12	
	c) 100 ms duration	22	
Ebata et al. (1968)	Clicks	10	Fused image at centre of the head
Freyman et al. (1991)	Clicks	5–9	Lag heard on 50% of trials
Yang & Grantham (1997b)	Clicks	5–10	Lag clearly audible on 75% of trials
Litovsky et al. (1999b)	Clicks	5–10	Lag clearly audible on 75% of trials
Localisation critical thresholds			
Litovsky et al. (1997a)	Clicks	8	Lead location chosen on 75% of trials
Litovsky et al. (1997b)	Clicks	11.4	Lead location chosen on 75% of trials
Discrimination critical thresholds			
Freyman et al. (1991)	Clicks	5–9	$d' = 1$
Yang & Grantham (1997a)	Clicks	5–10	Discrimination 75% correct
Litovsky et al. (1999b)	Clicks	5–10	Discrimination 75% correct

Fusion

The fusion effect refers to listeners' ability, at relatively short delays (< 5 ms for click sources equating to a small room), to fuse together a sound and its reflection, while two or more spatially separated sounds are present, into a single auditory event. Fusion can be considered as taking place below the echo threshold discussed above. Data about fusion is usually elicited by asking listeners, under controlled experimental conditions, to state how many sounds they hear for a range of lead–lag delays. The delay is then plotted against the percentage of trials in which two sounds were identified. For lead–lag delays of $>$ 8–10 ms, two sounds are usually heard on every trial (for click stimuli). Other observations can also be made during such experiments. Listeners frequently report other perceptual changes in the fused image, such as loudness, spatial extent and pitch, although the nature of the changes will depend on other factors including the type of stimulus used (Litovsky et al. 1999a).

Litovsky et al. (1999a) also note that there is considerable inter-subject variation with regard to the echo threshold and the strength of fusion (e.g. Freyman et al. 1991). For example, some listeners have shown no experience of fusion for delays as low as 2–4 ms whilst others report fusion beyond 10 ms for click stimuli. It is possible that listener instruction plays a role in this. Spatially separating lead and lag has also been found to lower the echo threshold.

One last observation that may be of particular relevance to CASA is that Litovsky et al. (1999a) argue that several precedence effects, including fusion, occur at comparable delays both monaurally and binaurally. They point out that whilst many authors (e.g. Blauert 1997) consider precedence to be a binaural phenomenon, eliciting similar data in monaural studies appears to discount this (Litovsky et al. 1999a).

Localisation Dominance

Localisation dominance refers to a listener's apparent ability to locate a sound source, despite the presence of reflections that may otherwise contradict the directional cues provided by the source. This is because the listener weights the direct information more highly than the reflections, although the directional information from the reflection is not completely ignored. Like fusion, localisation dominance is thought to be most prevalent when the delays are below the echo threshold. Data about localisation is normally elicited using one of three experimental paradigms: headphones, free field and room studies in the azimuthal plane, and free field studies in the median sagittal plane (i.e. the plane that travels from top-to-bottom of the body and divides it into equal left and right portions) (Litovsky et al. 1999a). These three paradigms are discussed below.

The first paradigm is studies using headphones. Most headphone studies are conducted such that listeners are required to match the stimulus position to a reference or to the

midline by varying a binaural parameter (e.g. ITD or ILD) of either the lead or the lag. The study of Wallach et al. (1949) implies that the lead is weighted at about four times more than the lag since for a lead–lag delay of 2 ms, a lead ITD of 100 μs required a lag ITD of 400 μs to move the image to the centre of the head.

According to Litovsky et al. (1999a), a more precise approach to estimating localisation dominance is to use an acoustic pointer, whereby the ITD of a noise burst is adjusted by the listener until it matches the stimulus (see for example Zurek 1980). This technique allows a direct estimation of the perceptual weight of the lead and lag using very few parameters. Subsequent studies (e.g. Shinn-Cunningham et al. 1993) have confirmed Wallach et al.'s (1949) findings in terms of lead and lag weighting and indicate a weighting factor of as much as 80–90% for the lead.

Another paradigm is free field and room studies in the azimuthal plane. Whilst the headphone measure can successfully elicit perceptual lead and lag weights, it has questionable ecological validity and can certainly be considered 'un-realistic'. Hence, free field studies are conducted to provide a more realistic scenario. Typical test scenarios involve two speakers, with variations in the lead–lag delay and relative level. Such tests indicate that localisation dominance is a trade-off between these two variables. Some tests use three or more speakers (e.g. Litovsky et al. 1997b) and provide strong evidence for localisation dominance. The data shows that for short delays (1–2 ms) the leading source was chosen 95% of the time but for delays above 5 ms lead and lag were chosen equally often. Such data are consistent with headphone studies but reveal little new information on the weighting of lead and lag for localisation (Litovsky et al. 1999a).

The final paradigm is free field studies in the median sagittal plane. Studying precedence in the median sagittal plane allows the investigator to assess monaural cues such as relative spectra and level without presenting binaural cues such as ITD and ILD. Such studies (e.g. Litovsky et al. 1997b; Dizon et al. 1997; Dizon & Litovsky 2004) show similar results to those of azimuthal investigations for lead–lag delays of 1–2 ms although the effect is slightly weaker possibly owing to a poorer localisation ability in this plane (Litovsky et al. 1999a).

Lag Discrimination Suppression

Lag discrimination suppression refers to a listener's ability to process spatial information about the lag stimulus. Experiments using headphones (e.g. Zurek 1980; Gaskell 1983; Shinn-Cunningham et al. 1993) attempt to elicit the just noticeable difference in ITD and ILD of the lead and lag stimuli. Free field experiments (e.g. Freyman et al. 1991; Yang & Grantham 1997a,b) attempt to elicit the discrimination of positional changes. These experiments generally agree that for lead–lag delays of < 5 ms (for clicks) it is difficult to discriminate changes in the lag, whilst changes in the lead are

easier to discriminate. The headphone experiments tend to agree that detecting changes in the ITD or ILD of the lag is made difficult when the lead–lag delay is of the order of 2–3 ms (Litovsky et al. 1999a).

Dynamic Processes in the Precedence Effect

In addition to the above effects, numerous researchers have observed dynamic processes in the precedence effect (e.g. Thurlow & Parks 1961; Clifton & Freyman 1989; Freyman et al. 1991; Clifton 1987; Freyman et al. 1991; Blauert & Col 1992; Blauert 1997). To summarise, these dynamic processes occur when "implausible" reflection patterns are heard. This may include, for example, an ITD that exceeds the maximum possible ITD in free-field conditions, implying that the time of arrival difference is due to a reflection. This implausible reflection pattern causes the precedence effect to breakdown whilst the listener rescans the room. This breakdown affects echo suppression, localisation, externalisation and fusion, and the listener is able to localise both lead and lag stimuli. The precedence effect then builds-up again, raising the echo threshold, in response to the new reflection pattern. Further discussion on these dynamic processes is given in Section 7.1 (page 111).

4.4 Reverberation Solutions: Machine

4. What are the machine listening solutions to reverberation, in particular in terms of source separation? How do machine listening solutions relate to human solutions?

The problems posed to machines by reverberation have been established in the previous sections of this chapter and include issues of ASR, pitch tracking, binaural cues and onset and offset detection. This section presents the machine solutions to reverberation. There are six main approaches to CASA (and ASR) in reverberant environments: pre-processing dereverberation, utilising robust acoustic features, reverberation masking, precedence modelling, spatial filtering and utilising multiple cues. These approaches will be discussed in this section. The relationship between these approaches and human solution to reverberation will be discussed at the end of the section.

4.4.1 Dereverberation

Dereverberation involves removing the echoes caused by room reflections such that any subsequent processing can be carried out on a 'clean' signal. Dereverberation has traditionally been used for speech enhancement with one of the earliest examples being provided by Allen (1973). The motivation behind this technique is clear: removing the reverberation will allow the use of established algorithms that already work in anechoic conditions.

Three techniques for dereverberating a signal have been suggested: feature–based techniques, inverse filtering and spatial filtering (see Section 4.4.2). These techniques are discussed below.

Feature–based Techniques

Several properties of speech can be exploited in order to dereverberate it. Yegnanarayana & Murthy (2000) find that the DRR of speech changes over its duration, both in the short term—within each glottal cycle—and over a longer term due to overlap masking. Furthermore, the speech signal is categorised according to whether it has a high DRR, a low DRR, or is purely reverberant. Using this information they are able to modify the LP residual signal (for a review see Makhoul 1975, 1976) in both short 2 ms regions and longer 20 ms regions. Yegnanarayana & Murthy report significantly less reverberation without any significant loss in speech quality.

In Section 4.3.1, the role of slow temporal speech modulation as a factor that aids in human robustness to reverberation was highlighted. Consequently, Avendano & Hermansky (1996) proposed the Inverse Modulation Transfer Function (IMTF), which attempts to reverse the effects of reverberation on the modulation spectrum of speech. Specifically, they split the signal into frequency bands and then learn filter coefficients for each band that will reduce the distance between the clean speech's temporal envelope and the reverberated speech's filtered temporal envelope. They find that the obtained filters are a good approximation of the ideal IMTFs. Unfortunately, this approach failed to dereverberate the speech and actually added audible artefacts.

One last acoustic feature that can be utilised is the harmonic structure of speech and exemplary models include those of Brandstein (1999) and Wu & Wang (2003). Both models make the assumption that time–frequency areas of the speech that show a clean harmonic structure remain uncorrupted by reverberation. In the case of Brandstein (1999), this knowledge is used to estimate the time delay of speech received by a pair of spatially separated microphones by highly weighting those areas that remain unaffected. Wu & Wang (2003) attempt to estimate the RT_{60} of the signal by using the correlogram under the assumption that pitch strength is inversely proportional to reverberation time; the correlogram model is borrowed from (Wu et al. 2003). Wu & Wang calculate a histogram in each time frame of the relative time lags between the pitch period and the closest peak in the corresponding frequency channel of the correlogram. As RT_{60} increases, the distribution of this histogram is shown to broaden, thus RT_{60} can be measured directly from the spread of the relative time lag distribution. With this estimate, the speech can be enhanced by identifying and subtracting echoes. Wu & Wang report that the output has appreciably reduced reverberation effects.

Inverse Filtering

Inverse filtering attempts to calculate a filter that reverses the effects of the room impulse response (Brown & Palomäki 2006). There are numerous approaches to the task of estimating this inverse filter. Hatziantoniou & Mourjopoulos (2004) show that robust inverse filters can be obtained from smoothing the room impulse response. In contrast, some authors attempt so-called 'blind deconvolution' or 'blind dereverberation' whereby the inverse filter is estimated. Approaches to this estimation include utilisation of the complex cepstrum (e.g. Tohyama et al. 1993) and ICA (e.g. Bell & Sejnowski 1995).

Another method was suggested by Nakatani et al. (2004) dubbed 'HERB' (Harmonicity based dEReverBeration). This is a hybrid method that exploits robust features but also blindly estimates an inverse filter. Specifically, periodic or quasi–periodic time segments are used to estimate the inverse filter. This filter aims to make each local time segment periodic; Nakatani et al. argue that if reverberation destroys periodicity then a periodic signal must be free of reverberation. With some enhancements to the algorithm made in (Nakatani et al. 2005), which improved frequency resolution and removed a constraint requiring a static $F0$ in each analysis window, the system achieves very effective dereverberation.

A CASA model that incorporates inverse filtering was suggested by Park & Stern (2007). The aim of the model was to separate a target from an interferer separated by 30° in the azimuthal plane. In their model, dereverberation forms the first stage of their model, i.e. before any peripheral processing or segmentation. To dereverberate the speech they estimate the transfer function from the autocorrelation of the LP residual signal. They choose to concentrate on early reflections at this stage in order to reduce the size of the required filter. They also note that early reflections are particularly problematic for source separation because their temporal proximity to the direct sound has a deleterious effect on the cross-correlation for a given frame. Following this, the dereverberated signals are passed through a cochlear model and each channel is cross-correlated. Late reflections are subsequently suppressed by an inhibitory signal, the nature of which is motivated by the precedence effect. Their work follows that of Palomäki et al. (2004b) (see Section 4.4.5); the inhibitory signal in each channel is the result of low–pass filtering the instantaneous envelope with a time delay and then re–applying the envelope. Finally, the T–F mask is calculated by combining information about the relative strength of the target with corresponding information about the echo suppression.

4.4.2 Spatial Filtering

Spatial filtering aims simply to handle reverberation by enhancing the target location and suppressing sounds from other directions such as room reflections. Whilst

multiple microphone approaches (beamforming) are common, some CASA systems have employed this approach (see for example Bodden 1993; Wittkop et al. 1997; Liu et al. 2001; Aoki & Furuya 2002; Roman & Wang 2004).

An early model that attempted to incorporate spatial filtering was that of Lyon (1983). His binaural model computed cochleagrams for each ear using a cochlear model and calculated ITD by cross-correlation. To estimate the cochleagram for each sound, a time–variable gain was utilised in each frequency channel and the gain was determined by the ITD. Lyon's paradigm involved the separation and dereverberation of two sound sources and employed 8 time–variable gains in order to produce cochleagrams for left and right direct sound and for left and right reverberant sound. In his informal evaluation, Lyon notes that whilst separation is achieved with some success, performance is poorest in the most reverberant parts of the signal due to the inability of the model to get an accurate estimate of source location. This could be overcome by integrating source–location estimates from less reverberated regions in a manner analogous to the precedence effect.

The model of Liu et al. (2001) (see also Liu et al. 2000) consists of two stages: firstly, multiple sound sources are located and secondly, noise sources are cancelled to leave only the target. In the first stage, a dual delay line is combined with coincidence detection. Cancellation is achieved by subtracting the two input signals such that nulls are created at the locations of each of the noise sources. These nulls are steered independently in each frequency channel to achieve maximum noise cancellation. Although the system achieves cancellation of three interfering talkers by 3–11 dB, the addition of moderate reverberation reduces this by about 2 dB. Liu et al. note that this drop in performance in reverberation is likely to limit the real–world usefulness of the algorithm.

A different approach was adopted by Roman & Wang (2004). Contrary to the above approach, their system aims to cancel the target sound source. Specifically, they aim to estimate the IBM by suppressing T–F units that are dominated by the target, as opposed to the traditional approach that keeps T–F units where the target is stronger than the noise. The motivation for this approach is their observation of the correlation between the amount of cancellation in a T–F region and relative level of the target and interfering sound sources in that region. They calculate this output-to-input energy ratio $OIR(i,l)$ in the following way:

$$\mathbf{OIR}(i,l) = \frac{|\mathbf{z}(i,l)|^2}{|\mathbf{y}(i,l)|^2} \tag{4.2}$$

where $\mathbf{z}(i,l)$ is the post–cancellation signal residue in the T–F region (i,l) and $\mathbf{y}(i,l)$ is the corresponding signal input. From this, simple logic can be used to test whether the target in a region has been suppressed and from this the IBM can be estimated. Specifically, if perfect cancellation is achieved then $\mathbf{OIR}(i,l) \to 0$. If T–F units remain dominated by noise then $\mathbf{OIR}(i,l) \gg 0$. By setting a threshold value Θ_i in each

frequency channel, the binary mask $\mathbf{m}(i,l)$ is calculated thus:

$$\mathbf{m}(i,l) = \begin{cases} 1 & \text{if } \mathbf{OIR}(i,l) > \Theta_i \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

4.4.3 Utilising Robust Acoustic Features

Utilising robust acoustic features involves representing speech, or some other signal, with features that are identified as being relatively robust to reverberation. This approach is common in ASR to reduce the dissimilarity between the training data gathered in anechoic conditions and the captured data that may have additional reverberation (Brown & Palomäki 2006).

Whilst representations such as Cepstral Mean Normalisation (CMN) (see for example Liu et al. 1993) can handle convolutional distortions that have a short impulse response, it is ineffective at handling strongly reverberated speech (Palomäki et al. 2004a). Hence, Brown & Palomäki (2006) suggest that exploiting the slow temporal modulations of speech may provide a more robust representation. Such an approach has been implemented by Langhans & Strube (1982) and Schlang (1989) who use a form of modulation filtering. Other approaches include that of Hermansky & Morgan (1994) (see also Hermansky 1990) who propose RelAtive SpecTrAl (RASTA)–Perceptual Linear Prediction (PLP) (RASTA–PLP). Again, this approach is concerned with exploiting the temporal differences in speech. In the case of convolutional distortions, the resulting artefacts are likely to be slowly changing in time. As such, a long term average can be subtracted in the cepstral domain. In the case of other noises, RASTA takes advantage of the rate of change of these sounds—assuming that they change at a different rate to speech and as such can be suppressed in the cepstrum. The steps involved in RASTA–PLP processing are summarised below:

1. Compute the power spectrum of the critical bands
2. Compress the spectral amplitudes through a non-linear transformation
3. Filter the time trajectory of each of these spectral components
4. Expand the filtered speech
5. Apply the equal loudness contour and then take the cube root to simulate the power law of hearing
6. Compute an all-pole model of the resulting spectrum

Unfortunately, Kingsbury (1998) points out that this approach, like CMN, is ineffective for convolutional distortions with a long impulse response, because the the impulse response (perhaps 0.5–2 s) is longer than the typical analysis window (16–32 ms). Furthermore, Kingsbury argues that convolutional distortions are only approximately additive in the cepstral domain if the analysis window in the spectral domain is two to four times longer than the distortional impulse response. Consequently, Kingsbury et al. (1998) enhance RASTA–PLP with the addition of the so-called modulation spectrogram. The modulation spectrogram is obtained by firstly using a filterbank similar to that of Greenwood (1961) which approximates cochlear filtering. Following this, the amplitude envelope of each channel is extracted by half–wave rectifying and low–pass filtering the signals with a cut-off frequency of 28 Hz. The envelopes are then down-sampled by a factor of 100 and normalised to the average level for each utterance. The slow modulations are extracted by filtering the envelopes and passing information below 8 Hz; the modulation spectrogram is essentially a spectrographic plot of the logarithm of this data. Kingsbury et al. show that this representation is quite robust to reverberation and additive noise.

Brown & Palomäki (2006) point out that there is a general problem with utilising robust acoustic features: they work well for the specific paradigm for which they were designed but are often not transferrable to other paradigms. Indeed, Kingsbury et al. (1998) show their system performed worse on anechoic speech than the standard PLP system of Hermansky (1990). Consequently, Brown & Palomäki suggest that a hybrid approach may be more suitable whereby conventional and reverberation–robust features are combined. Such a method was demonstrated by Kingsbury et al. (1998) in which PLP and modulation spectrogram features were combined resulting in better recognition performance.

4.4.4 Reverberation Masking

Reverberation masking involves utilising areas of the signal that appear relatively uncorrupted by reverberation and treating them differently to those that are more corrupted. Specifically, Palomäki et al. (2002) (see also Palomäki et al. 2004a) propose an approach built on the 'missing data' ASR framework. This technique provides the ASR system with acoustic features and a binary T–F mask indicating regions that are either reliable (1) or corrupted by noise or reverberation (0). This binary mask is calculated using a modulation filter to identify regions with strong speech energy, i.e. least contaminated by reverberation or noise.

Firstly, the signal is passed through a gammatone filterbank and the envelope of each channel is extracted. This envelope is then filtered with a low–pass filter to smooth speech modulations and differentiated to emphasise onsets which are likely to correspond to direct sound and early (non-detrimental) reflections. The reverberation

mask is set to 1 if the envelope lies above a threshold value and 0 otherwise. The threshold is calculated using a so-called "blurredness" metric. This metric is calculated as the ratio of average and maximum energy obtained from the envelope in each frequency channel. The final stage of the system is to spectrally normalise the signal, the normalisation factor is calculated from the regions marked as reliable. The results are similar to those obtained by Kingsbury (1998). However, this system has the advantage of being more flexible in terms of implementation because it can be used on reverberant speech using a recogniser that is trained on clean speech (Palomäki et al. 2002).

4.4.5 Precedence Modelling

Modelling the precedence effect explicitly aims to copy its perceptual counterpart. Several authors have suggested computational models of precedence (see for example Lindemann 1986a,b; Macpherson 1991; Martin 1997; Faller & Merimaa 2004). Whilst these models vary in their methods, they all at least agree on the psychoacoustical principles of the law of the first wave front. This aids in our ability to localise in reverberation, because the first wave front, which is the direct sound from the source and hence most informative in terms of direction, is weighted over subsequent echoes. As such, all of these models have the distinct function of localising sound sources.

To date, only two models aimed at separating sources have been suggested that incorporate a precedence model: that of Palomäki et al. (2004b) and Park & Stern (2007) (which is based on the former). The model of Palomäki et al. (2004b) is an ASR system that aims to separate speech from a spatially separated noise intrusion in small room reverberation. A schematic of their model is shown in Figure 6.1. The model is discussed in detail in Section 6.1 (page 91). To summarise, the model calculates an inhibitory signal from the Hilbert envelope. This signal retains onsets and suppresses information immediately following each onset that is likely to be corrupted by reverberation. This inhibitory signal is subsequently subtracted from the fine structure, which is then used to localise the sound sources and grouping decisions are made by estimating the relative strength of the two sources.

Interestingly, Palomäki et al.'s (2004b) model also includes a routine to normalise the spectral energy in order to reduce the aforementioned spectral effects of reverberation (see Section 4.3.2). The technique is similar to previous methods such as that adopted by (Kingsbury 1998) whereby acoustic feature vectors in each frequency band are normalised by the mean and variance of the spectrum. Palomäki et al. (2004b) point out that this method is ineffective for CASA since the normalisation factor will also be calculated from any additional interference that may be present. Hence, the difference in their system is that the normalisation factor is calculated only from *reliable* components indicated by the T–F mask.

In evaluating their model, Palomäki et al. make several observations. Firstly, in terms of ASR, they note that their system is more robust for its set of circumstances than a Mel–Frequency Cepstral Coefficient (MFCC) based system. Secondly they note that performance of the system depends on the angular separation of the noise intrusion, with performance falling as the angle is decreased. Thirdly, they note that the nature of the intrusion has consequences for the performance: it is improved when the spectra of the speech and noise are similar, since this aids the extraction of location cues. Lastly, the performance drops in increasing reverberation. However, the precedence model, although basic, appears to have increased localisation accuracy.

It is interesting to note that a large body of work on computational precedence exists, but only two separation algorithms exist that include precedence processing (and they are almost identical). This gap suggests that there is much more work to be done on incorporating precedence into source separation.

4.4.6 Utilisation of Multiple Cues

It has been noted in Section 4.2 that cues such as pitch and binaural cues become unreliable in reverberation. Hence, one solution to this problem is to utilise multiple cues in order to achieve a higher degree of robustness in reverberation. Such a system was proposed by Shamsoddini & Denbigh (2001), which combined both binaural and harmonicity cues. The first stage of the algorithm is to process binaural cues by calculating the cross-correlation over 33 ms frames. The angle of the dominant source is calculated from this function; it is presumed to be the target if it lies within $\pm 10^{\circ}$ on the median plane and presumed to be an interferer otherwise. The next stage of the algorithm estimates the target and interference using different strategies depending on which is dominant. In the case where the target is dominant, initial enhancement is achieved by adding frequency coefficients from the two microphones. Simultaneously, an estimate is made of the interference by subtracting the outputs of the microphones to place a null at the target location. For the case when the interference is dominant, the target is first enhanced by steering a null to the interference direction. In either case, the algorithm has now established estimates for both the target and interference. The magnitudes of the coefficients are subsequently adjusted to coincide with initial estimates. For cases where the interference coefficient is stronger than the target, the coefficient is simply discarded. Alternatively, when the target coefficient is stronger than the interference, the magnitude of the target coefficient magnitude is reduced by the interference coefficient magnitude. Shamsoddini & Denbigh show that this results in a substantial improvement in intelligibility for both one and two spatially separated interferers. Also, ASR accuracy increased by 65% (from 30% to 95%) for a signal-to-interference ratio of 12 dB and one interferer. Unfortunately, they do not state how performance changes with reverberation time.

A similar approach was proposed recently by Woodruff & Wang (2010). The model performs simultaneous and sequential grouping separately. In the simultaneous grouping stage, a T–F unit is labelled as voiced speech from its cross-channel correlation. For each of these units, two pitch points are chosen from peaks in the pooled autocorrelation. These points are then linked together to form pitch contours and grouping into segments is performed by measuring pitch deviation and spectral continuity. These segments are then linked over time using estimates of the source azimuth. Specifically, ITD is extracted using normalised cross-correlation and ILD is extracted as the ratio of signal energies. This information is combined with a trained likelihood function to derive the azimuth. Following a cue–weighting procedure inspired by the precedence effect that weights signal onsets, the segments are grouped based on the derived azimuth. The results show that the binaural system reduces azimuth errors when compared to previous approaches (e.g. Liu et al. 2000) and the system is robust to RT_{60}s up to 0.8 s, only dropping in labelling error by 5% over the range.

A different approach was proposed by Kollmeier & Koch (1994) (see also Wittkop et al. 1997). They utilise the modulation spectrogram (as described in Section 4.4.3) and also the observed interaction between modulation detection and binaural space which, according to Kollmeier & Koch, helps to separate different acoustic "objects" which are characterised both by a location in binaural space and by a particular range of modulation frequencies. The envelope cues and modulation characteristics are argued to be more robust to the effects of reverberation and additive noise than the fine structure. The model therefore first calculates the complex modulation spectrum by firstly splitting the left and right ear signals into frequency bands using the DFT. A second DFT is then taken of the envelope in each band to obtain the complex modulation spectrum. From this, binaural cues are extracted by dividing the complex modulation spectra by each other and taking the logarithm of the result; the ITD is the real part of the result whilst the imaginary part corresponds to the Interaural Phase Difference (IPD). Subsequently, a weighting function is derived that passes sounds from the source direction relatively unchanged but suppresses sounds arriving from other directions. The signal can then be simply reconstructed by applying this weighting function and then twice taking the inverse DFT and overlap adding. The system was evaluated subjectively by assessing speech intelligibility with one, two or four interfering sources present. The results indicated an effective increase in signal-to-noise ratio of 2 dB for both anechoic and reverberant conditions (RT_{60} = 1.33 s).

Recent work by Barker et al. (2010) uses a fragment–based approach to grouping. Fragments are similar to segments, as discussed in Section 3.3. The system performs primitive grouping by firstly identifying multiple pitch tracks using an algorithm based on autocorrelation (proposed by Ma et al. 2007). Simultaneous grouping is performed by comparing the autocorrelation responses of neighbouring units. Sequential grouping

is performed using a multi-pitch tracking algorithm proposed by Coy & Barker (2007). Additional schema–based grouping is performed with a hypothesis–driven process that attempts to identify foreground and background source fragments.

Additional fragment–based work has been conducted by Christensen et al. (2007) (see also Christensen et al. 2009) for robust extraction of binaural cues. The system uses a multi-pitch tracking algorithm to identify speech fragments. Localisation is then performed within these fragments. This system improved relative frame localisation accuracy by approximately 35%.

4.4.7 Perceptual Relevance of Machine Solutions

The solutions to reverberation described in this section have varying levels of relevance to human solutions to reverberation. For *Dereverberation*, the extent to which humans perceptually or physiologically dereverberate the sounds they perceive remains unclear. Calculating the IMTF had explicit perceptual relevance, but was found to be ineffective. As previously discussed in Section 4.3.3, some studies have hinted that the binaural auditory system is able to "squelch" reverberation when compared to monaural listening. Some work has been carried out by Libbey & Rogers (2000) where speech intelligibility tests were used to assess the effects of binaural listening, reverberation level and deconvolution processing. However, the results seem to be inconclusive. Furthermore, no perceptual mechanisms have been suggested that might accomplish such a feat, resulting in the utilisation of this technique having questionable perceptual relevance. For *Spatial Filtering*, there is no evidence to date that the auditory system performs this kind of processing and hence spatial filtering has questionable perceptual relevance. However, it is possible that spatial filtering may be achieved, albeit crudely, via an Equalisation–Cancellation (EC) mechanism (Durlach 1963, 1972). EC theory attempts to predict BMLD by first transforming the stimulus at the two ears (through the introduction of interaural time delays and level differences) and then cancelling the signals in order to at least partially reveal the desired masking component. Revealing the signal amongst noise in this way may be analogous to spatial filtering. EC models have been consistently shown to provide a good match to psychoacoustic data (e.g. Zurek et al. 2004; Culling & Lewis 2010). For *Utilisation of Robust Acoustic Features*, many of the techniques discussed in this section have little or no perceptual relevance and the implementations stated such as RASTA–PLP and CMN appear more like engineering approaches. *Reverberation Masking* seems like an engineering solution to the issue of reverberation and it remains unclear whether there is an equivalent perceptual mechanism. For example, Libbey & Rogers (2004) suggest that binaural overlap masking release may make some contribution to the binaural advantage. Conversely, Watkins (2005) suggests that late reverberant tails, which may be suppressed by a reverberation mask, are crucial to the perceptual ability to

compensate for reverberation. However, since this technique has mainly been applied to ASR, it is unclear whether it will be effective for CASA. *Precedence Modelling* explicitly aims to mimic one perceptual technique that is known to account for human robustness to reverberation for some auditory tasks such as localisation. *Utilisation of Multiple Cues* is a perceptually–relevant methodology, since it is clear from Chapter 2 that humans utilise many cues in order to accomplish auditory tasks, including ASA.

4.5 Summary and Conclusions

This chapter aimed to answer the following questions:

1. What are the problems posed by reverberation to human auditory perception in general?
2. What are the problems posed by reverberation to machine listening in general?
3. What are the human solutions to reverberation?
4. What are the machine listening solutions to reverberation, in particular in terms of source separation? How do machine listening solutions relate to human solutions?
5. Which reverberant source separation solution has most scope for improvement?

1. What are the problems posed by reverberation to human auditory perception in general?

It was established in this chapter that reverberation poses several problems for human auditory perception. These problems include degradations in speech perception, source separation and sound localisation. This is because reverberation blurs or destroys many cues—such as periodicity, the temporal and spectral envelopes and binaural cues—that humans rely on for these tasks.

2. What are the problems posed by reverberation to machine listening in general?

Similarly to the effects on human auditory perception, and for the same reasons, reverberation has deleterious effects on numerous aspects of machine listening including the extraction of features such as pitch, binaural cues, onsets and offsets, and on applications such as ASR.

3. What are the human solutions to reverberation?

In response to the problems posed by reverberation, humans have numerous mechanisms that are used in order to attempt to overcome its effects. These mechanisms include: utilising the slow temporal modulation of speech, which occurs at rates below the envelope–filtering effect of reverberation; the binaural advantage, whereby listeners

gain a significant advantage in many areas of perception by having two ears rather than one; spectral envelope distortion compensation, which counteracts the spectral distortion introduced by reverberation; and precedence, which weights the first few wavefronts of the direct sound over later wavefronts arriving as reflections from other surfaces.

4. What are the machine listening solutions to reverberation, in particular in terms of source separation? How do machine listening solutions relate to human solutions?

Several machine listening techniques were demonstrated that were designed to reduce the deleterious effects of reverberation. *Dereverberation* removes reverberation before any further processing, the motivation being that its use permits the use of existing algorithms that are untested in reverberation. This is an effective technique but its perceptual relevance remains unclear. *Spatial filtering* aims to enhance the target location and suppress sounds, including reverberation, arriving from other directions. However, this approach depends on the ability of the algorithm to locate the sound source, an ability that may be severely impeded by reverberation. Although it is possible that spatial filtering may be achieved via an EC mechanism, this link requires further research. *Utilising robust acoustic features* represents the signal using features that are robust to reverberation. Unfortunately, many of the approaches described in the literature are not usable in paradigms other than the one for which they were developed. Furthermore, this technique has little or no perceptual relevance. *Reverberation masking* attempts to identify T–F regions that show minimal corruption by reverberation. This technique has questionable perceptual relevance and remains untested for CASA. *Precedence modelling* attempts to enhance source localisation estimates by modelling the perceptual precedence effect. The localisation data can then be used to inform grouping. This technique has perceptual relevance. However, whilst much work has been carried out on computational precedence, relatively little work has been carried out on incorporating this into CASA. *Utilisation of multiple cues* is motivated by the idea that if individual cues break down in reverberation, gathering data from many cues may achieve greater robustness to its effects. This approach has perceptual relevance since it is clear that humans use many acoustical cues in order to accomplish auditory tasks.

5. Which reverberant source separation solution has most scope for improvement?

From the answers to the previous questions it can be concluded that, within the scope of the current investigation, modelling the precedence effect offers the most scope for improvement. There are four reasons for this: firstly, it is perceptually–relevant. Secondly, it remains relatively untested for source separation. Thirdly, there is a comprehensive existing body of work on computational precedence. Lastly, previous work has shown that with suitable processing, the reverberation–robustness of spatial cues can be improved. Furthermore, for other cues, it was shown that

onsets and offsets are likely to be unreliable in reverberation, and pitch is only robust to reverberation if dereverberation processing is introduced, which has questionable perceptual relevance. It is for these reasons that the study documented in Chapter 6 will investigate precedence modelling for source separation.

Chapter

5 Evaluating Source Separation in Reverberant Environments

Before investigating modelling the precedence effect for source separation, it is important to determine an assessment procedure that is suitable for reverberant environments. Investigating models of the precedence effect also places specific requirements upon the experimental procedure. Specifically, since the precedence effect is predominantly known to assist in localisation, the separation system needs to be based on source localisation. Research Question 6 is therefore adapted (and re-numbered 6′) to fit this criterion: *How should the performance of separation algorithms incorporating different precedence models be evaluated? What signals? What metrics?* In order to answer these questions, this chapter is broadly split into two parts: firstly, the experimental procedure and mixture parameters are described in Section 5.1. Secondly, the choice of metric is discussed in Sections 5.2–5.4. The chapter is summarised and concluded in Section 5.5.

As previously stated, investigating the precedence effect for source separation effectively requires the separation system to be based on source localisation. Typically, research centred on this paradigm aims to separate signals arising from two spatially–separate sound sources located in a range of reverberant rooms (e.g. Roman et al. 2003; Palomäki et al. 2004b; Woodruff & Wang 2010). Hence, it is this paradigm upon which subsequent investigations into precedence modelling for source separation are centred.

5.1 Experimental Procedure

6a. How should the performance of separation algorithms incorporating different precedence models be evaluated? What signals?

In the aforementioned paradigm of separating signals arising from two spatially–separate sound sources, there are four parameters that can and have often been varied in previous research. These parameters are: the spatial separation of the target and interfering sound sources, the signals produced by each source, the relative loudness of each sound source and the room. These parameters are herein referred to as the mixture parameters, since they determine all of the characteristics of the binaural recording that is subsequently used for separation. The room is typically simulated using a set of Binaural Room Impulse Responses (BRIRs). These responses fully describe the

transmission of sound from a sound source to each ear of a Head And Torso Simulator (HATS). The BRIRs permit the simulated reproduction of any signal from the sound source position within a room for which the BRIRs were captured. Furthermore, it is also common to test a given separation algorithm in every combination of mixture parameters. Note that this is a trade-off between maximising the number of variables (in order to ensure the results are representative of realistic situations) and minimising the processing time.

Each of the mixture parameters are discussed in this section. The subsequent work described in this thesis is based on that proposed by Palomäki et al. (2004b). Consequently, the mixture parameters employed in this investigation are similar to those employed by Palomäki et al.

5.1.1 Spatial Separation of Target and Interferer

When spatially–separating the target and interfering sounds, in order to fairly compare the effect of different separations it is useful to retain a constant distance from the HATS in order to eliminate the effect of distance on the separation. Distance could be employed as an additional variable. However, distance has been neglected in this study because it primarily affects the DRR of the mixture. Changing rooms will affect numerous acoustical parameters, including RT_{60}. As noted in Section 4.2 (page 49), changing RT_{60} is more likely to have a deleterious effect on separation perfromance because it increases the duration of overlap- and self-masking. Furthermore, most studies of source separation in reverberation have considered RT_{60} as the primary acoustic parameter. Therefore, separations are defined in terms of their azimuth relative to the HATS, with 0° lying on the intersection of the median sagittal and horizontal planes at ear height and the sources at a distance of 1.5 m from the HATS. The azimuthal separations were chosen to be the same as those used by Palomäki et al. (2004b): 10°, 20° and 40° (i.e. ±5°, ±10° and ±20°); the target was consistently chosen to be the source on the left. These azimuths present a range of challenges: at wider azimuths localisation should be straightforward and separation relatively successful. As the separation reduces, the algorithm may no longer be able to distinguish two separate sources and hence separation may be less successful.

5.1.2 Relative Loudness

The relative level of the target and interferer is described by the Target-to-Interferer Ratio (TIR), i.e. the ratio in dB of the RMS levels of the target and interfering sound sources. At higher TIRs, it may be more difficult to localise the interfering sound source and hence separation may be less successful, and vice versa. Similar TIRs to those used by Palomäki et al. (2004b) were chosen: 0, 10 and 20 dB.

5.1.3 Signals

Many source separation algorithms are developed as front-end processing for ASR systems and hence ASR is often chosen as the evaluation metric. Consequently, the target signal is often taken from a corpus of utterances that have been specifically developed for speech recognition tasks. This is the approach chosen by Palomäki et al. (2004b). In other studies that use SNR–based metrics, which have no specific requirements in terms of signals, the corpus developed by Cooke (1991)[1] is often used. The interferers in Palomäki et al.'s (2004b) are drawn from this set. As will be discussed later, the following investigations do not use ASR as the metric and hence there are no specific requirements on the signals used for separation. Despite this, to retain a similarity with Palomäki et al.'s (2004b) study, the stimulus set is based on their set. Specifically, the target signal was a 4 second excerpt of female speech taken from EBU SQAM (1988). The interfering signals were chosen to be: a rock music track ("Action!" by Razorlight; an up-to-date version of Cooke's rock music excerpt), white noise and an excerpt of male speech also taken from EBU SQAM. The speech segments were chosen to incorporate a wide range of phonemes. These interferers present a range of challenges: speech is a sparse signal that should not always overlap the target; the rock music is more noise–like, demonstrating significant spectral overlap, but with onsets that should allow relatively successful localisation in reverberation; the noise also has significant spectral overlap but no onsets and hence may be more difficult to localise in reverberation.

5.1.4 Binaural Room Impulse Responses

In their paper, Palomäki et al. (2004b) generate artificial BRIRs by combining the Gardner & Martin (1994) Head–Related Transfer Function (HRTF) database with a model of small room acoustics. However, it was decided to use BRIRs captured in real rooms rather than simulating them due to the generally poor subjective quality of responses calculated using acoustic models.

The responses from four rooms were captured, with an additional set captured for the anechoic condition. The rooms were chosen to demonstrate a range of RT_{60}s in the interval [0,1] s, since these times are typical of rooms used everyday in the real world and in studies of this type. The following paragraphs describe how the responses were captured.

Capturing the Responses

The impulse responses were captured using a Cortex Instruments Mk.2 HATS. They were obtained from sinesweeps replayed through a Genelec 8020A active loudspeaker

[1] Available from `http://www.dcs.shef.ac.uk/~martin/`

Table 5.1: Room acoustical properties, including RT_{60}, Initial Time Delay Gap (ITDG), Direct-to-Reverberant Ratio (DRR) and clarity index C_{50} (for speech).

Room	RT_{60} [s]	ITDG [ms]	DRR [dB]	C_{50} [dB]
A	0.32	8.72	6.09	16.5
B	0.47	9.66	5.31	11.4
C	0.68	11.9	8.82	17.4
D	0.89	21.6	6.12	9.43

and the responses were deconvolved to produce the impulse responses. The recordings were made at a sampling frequency of 48 kHz (and subsequently resampled to 16 kHz). The loudspeaker was placed around the HATS on an arc in the horizontal plane with a 1500 mm radius between $\pm 90^\circ$ and measured at 5° intervals. The acoustic centre of the loudspeaker was placed at the same height as the ears. Diagrams of each of the rooms are provided in Appendix A. A summary of the acoustical propoerties of each of the rooms is provided in Table 5.1. Measurements of RT_{60} were obtained according to BS EN ISO 3382: 2000 using an interrupted pink noise method with six microphone and two loudspeaker positions (12 measurements in total). In accordance with the standard, the overall room RT_{60} is calculated by averaging the 500 Hz and 1 kHz bands. The octave-band RT_{60}s are given in Appendix A. Other parameters were measured post-hoc directly from the impulse responses.

Anechoic Condition

To test the anechoic condition (later referred to as 'X'), two options were available:

Utilise an available HRTF database This method is advantageous because many of the HRTF databases have been comprehensively tried and tested. However, no database exists that uses the the aforementioned experimental combination of loudspeaker, HATS and distance.

Generate pseudo-anechoic responses Since no anechoic chamber was available, this method would facilitate use of the loudspeaker, HATS and distance used in capturing the other BRIRs and maximises the commonality across the set of BRIRs.

Consequently, the second approach was chosen and a pseudo-anechoic method based on that suggested by Fincham (1985) was utilised whereby the responses were captured in a large room and simply truncated to before the first reflection so that subsequent reflections did not colour the frequency reponse. The room measured $17.04 \times 14.53 \times 6.5$ m ($l \times w \times h$), the HATS and loudspeaker were placed in the centre of the room at a height of 2.8 m and separated by 1.5 m. An example of this approach is shown in Figure 5.1 which was captured in the same space using a Genelec 8020A loudpseaker and a B&K

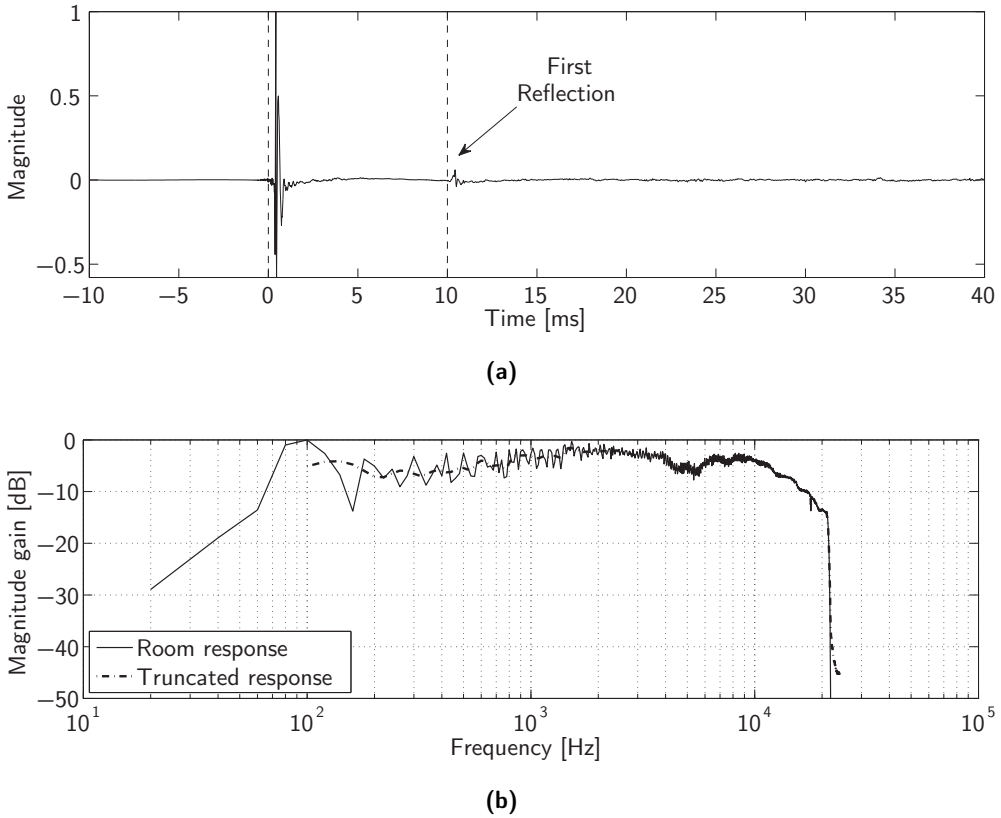


Figure 5.1: Example of a pseudo-anechoic impulse response measurement. **(a)** Impulse response captured in a large room. **(b)** Frequency response of the truncated and un-truncated impulse responses.

4003 omnidirectional microphone at the same positions. It shows the captured impulse response (panel (a)) and frequency responses calculated using both the full impulse response and the impulse response truncated to before the first reflection indicated at about 10 ms (panel (b)). It is clear that much of the frequency colouration has been removed.

5.2 Ideal Binary Masks and Metrics

6b. How should the performance of separation algorithms incorporating different precedence models be evaluated? What metrics?

A popular metric for assessing the performance of source separation algorithms is the estimation of a form of SNR (Li & Wang 2009), which is typically calculated thus:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_n s_t^2(n)}{\sum_n \big(s(n) - s_t(n)\big)^2}\right) \tag{5.1}$$

where s_t is the target signal, s is the estimated target signal and n is the sample index. Note that the denominator is a summation of a difference signal and thus incorporates

any and all differences between the target and estimated target.

As can be seen in Equation 5.1, an important point to note about SNR–based metrics is their incorporation of convolutional distortions such as room reverberation. A reverberated signal s_r can be considered in the following way:

$$s_r(n) = s_t(n) + \sum_{d=1}^{D} a_c(d) s_t(n-d) \tag{5.2}$$

where $|a_c| < 1$ are reflection coefficients, $d \in \mathbb{N}$ and $D \leq \infty$ (for signal processing D will be considerably smaller: of the order of a few seconds (in samples)). Therefore, because reverberation can be considered as an additive component that contributes only to the estimated target, substituting s_t with s_r in Equation 5.1 increases the magnitude of the denominator and lowers the SNR. Furthermore, the calculated SNR is likely to vary dramatically according to the nature of the reverberation. Hence, for the same signals and binary mask, SNR is likely to demonstrate large inconsistencies between different acoustic environments. This prevents meaningful comparison of separation algorithms across different acoustic conditions. Source separation in reverberation is an important research goal and testing and comparing separation algorithms in a range of reverberant conditions is a common task in this field. The comparison of algorithms across different acoustic conditions is also an important requisite for this thesis.

The importance of reverberation to the output is dependent upon the application of the algorithm. For applications such as ASR, the resulting distortions may be undesirable because many speech databases are not trained on reverberant speech. However, Zurek (1987) notes that reverberation makes a significant contribution to the timbral and spatial characteristics of a perceived sound. Thus reverberation may be essential for applications such as auditory scene reconstruction (i.e. the separation and subsequent manipulation or reconfiguration of spatial auditory objects). With so many potential applications for source separation, each with slightly different requirements, it is important that the assessment procedure remains independent of application and retains a common ground on which algorithms may be compared. Furthermore, when considering reverberant conditions, it is desirable for a metric to assess the separation performance of the algorithm in the reverberant conditions, without assessing the effect of the reverberation on the output.

A recent study by Mandel et al. (2010) has suggested a metric for assessing the separation of reverberated speech. The metric, termed Direct-path, Early echoes, and Reverberation of Target and Masker (DERTM), measures the suppression of the direct sound, early reflections and late reverberation of both the target and interfering sounds. This is because Mandel et al. find that, for speech intelligibility, suppressing late reverberation is an important goal for a binary mask. The metric is shown to be very effective for reverberated speech, but this limits its application, since speech is

not necessarily the only signal that might need to be extracted (musical instrument separation is also a common task). Furthermore, it assumes that intelligibility is the ultimate goal for source separation, which, as discussed above, may or may not be the case.

A common goal for source separation algorithms—and the goal proposed for CASA by Wang (2005)—is to estimate the IBM. The IBM $\mathbf{m}_{\text{ibm}}$ is set to one at frequency bin i and time frame l when the ratio of the target sound source energy $\acute{\mathbf{u}}_t$ and total interference energy $\acute{\mathbf{u}}_i$ exceeds a threshold value, and zero otherwise, thus:

$$\mathbf{m}_{\text{ibm}}(i,l) = \begin{cases} 1 & \text{if } 10\log_{10}\left(\dfrac{\acute{\mathbf{u}}_t(i,l)}{\acute{\mathbf{u}}_i(i,l)}\right) > \theta \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

where θ is a threshold value in dB and usually chosen to be 0. This criterion is based upon the principle of psychoacoustical auditory masking whereby stronger energy within a critical band masks weaker energy (Roman et al. 2003; Moore 2004). This point of view was supported in a recent paper in which Li & Wang (2009) suggest that estimating the IBM remains a good objective for sound source separation and provides a good indication of performance. Furthermore, other studies have also shown that, at least for speech recognition, estimating the IBM remains a reasonable objective for source separation in reverberant environments (Roman & Wang 2004; Palomäki et al. 2004b; Jin & Wang 2009)

Hu & Wang (2004b) point out that SNR does not take perceptual phenomena such as auditory masking and phase spectrum insensitivity (Helmholtz 1885; Moore 2004) into account. Consequently they utilise the target resynthesised from the IBM, s_{ibm}, as the ground truth when calculating SNR. This modified version of SNR is referred to as Signal-to-Ideal-Noise Ratio (SINR), such that:

$$\text{SINR} = 10\log_{10}\left(\frac{\sum_n s_{\text{ibm}}^2(n)}{\sum_n \big(s(n) - s_{\text{ibm}}(n)\big)^2}\right) \tag{5.4}$$

One further option is to use the reverberated target as the ground truth in calculating SNR. This is referred to as Reveberant-Signal-to-Noise Ratio (RSNR), such that:

$$\text{RSNR} = 10\log_{10}\left(\frac{\sum_n s_{\text{r}}^2(n)}{\sum_n \big(s(n) - s_{\text{r}}(n)\big)^2}\right) \tag{5.5}$$

Whilst these approaches address some of the issues of SNR discussed above—by incorporating the reverberation into the numerator and denominator of Equation 5.1—for SINR, unless the estimated mask is identical to the IBM, s will, in most practical situations at least, differ from s_{ibm} and that difference will include reverberant energy (as well as target energy and interferer energy). For RSNR, s will include some interferer reverberation and exclude some target reverberation (again, in most practical situations

at least). These contributions of reverberant energy to the denominator of Equation 5.1 may differ dramatically from one environment to the next and so (as discussed above for SNR) the calculated SINR and RSNR are both likely to be inconsistent across reverberant environments. As stated above, this inconsistency is undesirable for a separation metric, which should not consider the effect of reverberation on the output of the system, and it prevents easy comparison between studies.

In addition to the above considerations, Li & Wang (2009) show that for:

- an acoustic mixture that is a sum of two signals with no additional convolutional distortion,
- rectangularly windowed non-overlapping masks,

the IBM is optimal in terms of SNR. This is an important result, because it means that any deviation from the IBM will produce a sub-optimal separated output. As previously discussed, the addition of convolution distortion is likely to have a significant impact on the calculated SNR. However, the DRR of a discontinuous signal such as speech is time–dependent due to the time–varying nature of the signal energy (Yegnanarayana & Murthy 2000). Therefore, a mask may exist that minimises the presence of reverberation whilst maximising the contribution of the target. This may undermine the optimality of the IBM in terms of SNR.

Therefore, to test the consistency of SNR, SINR and RSNR in reverberation, and the effect of reverberation on the optimality of the IBM in terms of SNR and RSNR, an experiment was conducted that compared the separation of an un-convolved mixture with that of mixtures created with additional reverberation obtained from a range of real rooms. In all cases the separation performance of the IBM is compared with a notional binary mask. The study is detailed in the following section.

5.3 The Ideal Binary Mask in Reverberant Conditions

This section details a study that investigated the optimality of the IBM in terms of SNR and RSNR in reverberant conditions and the effects of reverberation on SNR, RSNR and SINR. The study investigated the separation performance of the IBM and a range of notional masks. The notional masks were *likely* experimental masks, calculated using techniques representative of those used in existing algorithms, as detailed in Section 5.3.1. The inputs were monaural mixtures of a target speech signal and interferer with varying TIRs. The mixtures were created anechoically (with no convolutional distortion) and by convolving the sources with impulse responses captured from the rooms described above. The separation procedure is described in the following section. The masks were tested with a range of mixture conditions; the experimental procedure is described in Section 5.3.2.

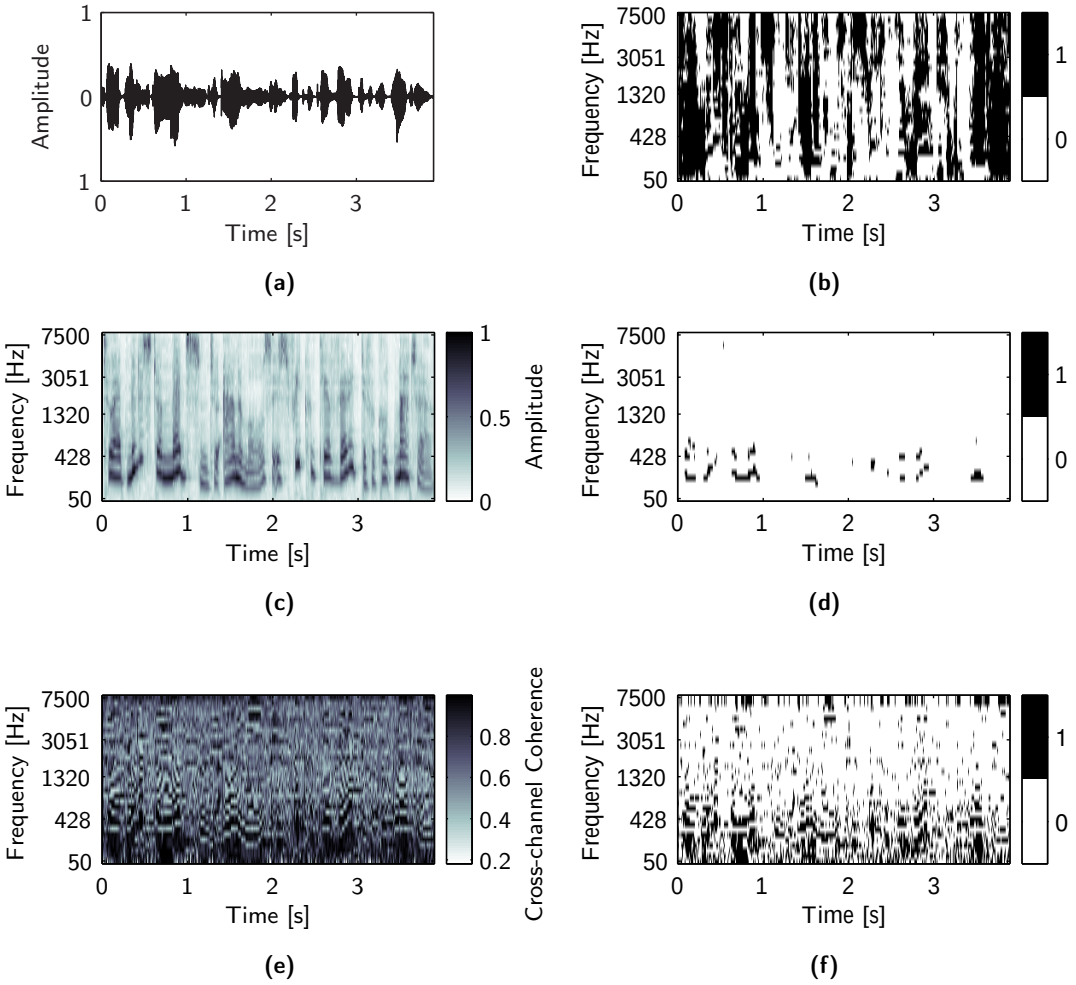


Figure 5.2: Examples of the processing employed in the metric study (no reverberation). **(a)** The target waveform (female speech). **(b)** The IBM with a male speech interferer and a TIR of 0 dB. **(c)** Cochleagram of the target. **(d)** Mask A with $\Theta_m = 0.7$. **(e)** The cross-channel coherence. **(f)** Mask B with $\Theta_m = 0.9$.

5.3.1 Mask Calculation

This section describes the procedure used to calculate both notional and ideal masks. Two processing techniques were utilised to create two sets of masks: A and B. For each processing technique, a range of masks was created by varying a threshold value Θ_m in the interval $[0, 0.99]$. Examples of the T–F representations and binary masks calculated using these processing techniques are given in Figure 5.2.

Notional Mask A

Notional mask A used a procedure based on target signal energy. A range of masks was created with each T–F unit set to one when the target signal energy exceeded a variable threshold.

The peripheral analysis procedure is loosely based on that described by Palomäki et al.

(2004b) and is almost identical to the procedure described in Chapter 6. Firstly, the clean target is passed through a gammatone filterbank (see Section 3.1.1); 32 channels were utilised with centre frequencies equally spaced on the ERB–rate scale in the range 50–7500 Hz. The Hilbert envelopes $\boldsymbol{\varepsilon}(i,n)$ (for sample index n) of each of these signals—which were obtained directly from the complex gammatone coefficients (see Equation 3.3, page 22)—were used to estimate $\hat{\mathbf{u}}(i,l)$ the normalised auditory nerve firing rate:

$$\hat{\mathbf{u}}(i,l) = \frac{\mathbf{u}(i,l)}{\mathring{\mathbf{u}}} \tag{5.6}$$

where

$$\mathring{\mathbf{u}} = \max_{i,l} \mathbf{u}(i,l), \tag{5.7}$$

$$\mathbf{u}(i,l) = \acute{\boldsymbol{\varepsilon}}\big(i,(l-1)M+1\big)^{0.3}, \tag{5.8}$$

$$\acute{\boldsymbol{\varepsilon}}(i,n) = \boldsymbol{\varepsilon}(i,n) - e^{-\alpha_s}\acute{\boldsymbol{\varepsilon}}(i,n-1), \tag{5.9}$$

M is the frame length in samples (10 ms), $\mathbf{u}$ denotes the auditory nerve firing rate and α_s is a time constant set in samples to 8 ms. This representation was used to calculate the notional mask $\mathbf{m}_A$:

$$\mathbf{m}_A(i,l) = \begin{cases} 1 & \text{if } \hat{\mathbf{u}}(i,l) > \Theta_m \\ 0 & \text{otherwise} \end{cases} \tag{5.10}$$

where Θ_m is the threshold value that is varied to create a set of masks. Note that since the mask was calculated using the clean target signals, notional mask A was independent of the acoustic conditions. Specifically, for a given mixture and threshold value, the mask will be identical in all of the rooms. Hence, any differences across the rooms seen in metric performances later can only be attributed to the differences in convolutional distortion.

Notional Mask B

Notional mask B used a procedure based on normalised cross-channel correlation (loosely based on that described by Wang (2006)). Specifically, following the gammatone filterbank used to calculate mask A, the cross-channel coherence $\hat{\mathbf{k}}$ was calculated in the following way:

$$\hat{\mathbf{k}}(i,l) = \max_{\tau} \frac{\hat{\mathbf{a}}(i,l,\tau)\hat{\mathbf{a}}(i+1,l,\tau)}{\sqrt{\sum_\tau \hat{\mathbf{a}}^2(i,l,\tau) \cdot \sum_\tau \hat{\mathbf{a}}^2(i+1,l,\tau)}}, \tag{5.11}$$

where

$$\hat{\mathbf{a}}(i,l,\tau) = \frac{\mathbf{a}(i,l,\tau) - \frac{1}{M}\sum_\tau \mathbf{a}(i,l,\tau)}{\frac{1}{M}\sum_\tau \big(\mathbf{a}(i,l,\tau) - \frac{1}{M}\sum_\tau \mathbf{a}(i,l,\tau)\big)^2}, \tag{5.12}$$

$$\mathbf{a}(i,l,\tau) = \sum_{n=0}^{M-\tau-1} \mathbf{h}\big(i,(l-1)M+n+\tau\big)\mathbf{h}\big(i,(l-1)M+n\big), \tag{5.13}$$

$\{\tau \in \mathbb{Z} : 0 \leq \tau \leq M-1\}$ is the discrete correlation lag, $\mathbf{a}$ is the autocorrelation, $\hat{\mathbf{a}}$ is the normalised autocorrelation and $\mathbf{h}$ is the half-wave rectified fine structure output of the gammatone filterbank. Finally, the binary mask $\mathbf{m}_B$ was set using the following logic:

$$\big\{\mathbf{m}_B(i,l),\mathbf{m}_B(i+1,l)\big\} = \begin{cases} 1 & \text{if } \hat{\mathbf{k}}(i,l) > \Theta_m \\ 0 & \text{otherwise} \end{cases} \tag{5.14}$$

As with mask A, the mask was calculated using the clean target signals.

The Ideal Binary Mask

In order to calculate the IBM, the (un-normalised) auditory nerve firing rate was calculated as for notional mask A, except that the inputs were the clean target and interfering signals. The estimate of the auditory nerve firing rate was used to estimate the auditory energy according to Palomäki et al. (2004b):

$$\acute{\mathbf{u}}(i,l) = \big(\mathbf{u}(i,l)^{3.333}\big)^2 \tag{5.15}$$

These data were used to calculate the IBM as in Equation 5.3. Note that since some mixture parameters were varied, the IBM also varied (see next section).

5.3.2 Experimental Procedure

A range of conditions was employed to ensure that the performances (reported later) were representative of a range of realistic conditions offering a varying degree of difficulty. However, only the rooms will be compared in the results, with model performances reported as means calculated across the other variables.

The masks were tested using the procedure described in Section 5.1, with the exception of the azimuthal separations, since the experiment is monaural. To summarise, the masks were tested with the following conditions:

- three TIRs of 0, 10 and 20 dB Root Mean Square (RMS) (i.e. the ratio of the clean target and interferer in terms of their RMS level)
- three interfering signals: white noise, male speech and a modern piece of rock music
- a range of reverberant conditions from real rooms (A–D) and an anechoic mixture (X) where no convolutional distortion was introduced

5.3.3 Results and Discussion

The results from the experiment are given in Figure 5.3. The figures are the mean values calculated across the target stimuli, interferer stimuli and TIR experimental variables. The main plots demonstrate the performance of the two notional masks. The performance of the IBM in terms of SNR is shown in the right hand plot of Figure 5.3(a) and 5.3(b) (the data are the same and repeated only for comparison). The performance of the IBM in terms of RSNR is shown in the right hand plot of Figure 5.3(c) and 5.3(d). The IBM data are calculated for each mixture condition and are hence independent of the variable threshold Θ_m.

A number of important observations can be made about the results:

- For the anechoic condition (Room X), the IBM is optimal in terms of SNR, which agrees with Li & Wang's (2009) findings.
- With the addition of reverberation, SNR demonstrates large inconsistencies across the different acoustic conditions, both in terms of absolute values and data trends.
- In some conditions, the notional masks are seen to out-perform the IBM in terms of SNR, which has undermined the optimality of the IBM.
- In some rooms, the SNR is seen to increase with the threshold value, contrary to SINR and anechoic conditions. This implies that these masks, calculated with very high thresholds, are optimal. However, in reality they retain very little of the target sound.
- For RSNR, where the target is reverberated, the IBM remains optimal in all conditions
- RSNR and SINR demonstrate a more consistent pattern of results across the anechoic and reverberant conditions.
- There are still significant variations in the values of RSNR and SINR that can only be attributed to the acoustic conditions.
- In the anechoic condition all of the plots show a general agreement in data trends.

The inconsistencies across the tested acoustic conditions shown in the SNR results can only be due to the contribution of the reverberation, a finding that is in agreement with the discussion in Section 5.2. The reverberation increases the difference between the target and estimated target signals and hence increases the magnitude of the denominator when calculating SNR. In cases where the notional masks are seen to out-perform the IBM, the notional masks may choose areas of high target energy that are likely to have a high DRR. Conversely, the IBM may incorporate areas with low

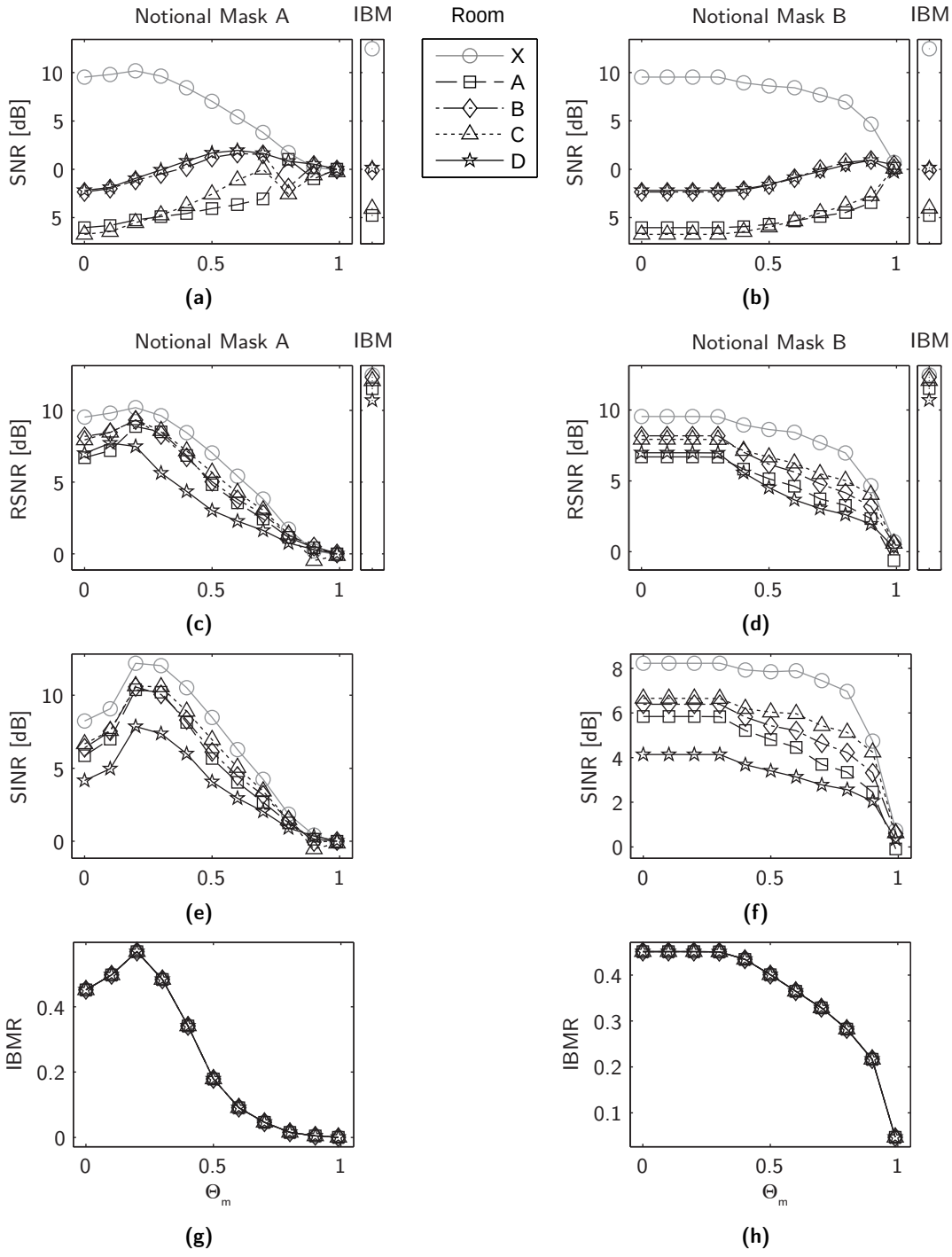


Figure 5.3: Results for the two notional masks showing the variation in results with the threshold values and room, averaged over other variables. **(a)** SNR results for notional mask A and IBM. **(b)** SNR results for notional mask B and IBM. **(c)** RSNR results for notional mask A and IBM. **(d)** RSNR results for notional mask B and IBM. **(e)** SINR results for notional mask A. **(f)** SINR results for notional mask B. **(g)** IBMR results for notional mask A. **(h)** IBMR results for notional mask B.

target energy (it only needs to be greater than the interferer) that are likely to have a low DRR. For the notional mask, the reverberation contributes less to the denominator and hence it appears to out-perform the IBM. The RSNR and SINR data are quite different to the SNR data. In almost all acoustic conditions the RSNR and SINR are positive and demonstrate a higher degree of consistency across the tested acoustic conditions in terms of data trends. The positive results are due to the reduction in the contribution of reverberant energy to Equation 5.1.

However, these results demonstrate that, with all mixture parameters remaining constant apart from the room reverberation, SNR, RSNR and SINR are unable to provide a consistent score for the same binary mask. As discussed in Section 5.2, comparison of algorithms across different acoustic conditions is a common and important task. However, the reverberation has directly affected the calculated SNR, RSNR and SINR and this is problematic for a performance metric.

5.4 The Ideal Binary Mask Ratio

The experiment conducted in the previous section demonstrated that metrics based on Signal-to-Noise Ratio are unable to provide a consistent score for a given binary mask when convolutional distortions are introduced. It is therefore desirable to find a metric that can provide a consistent score for a given binary mask independently of convolutional distortions. Hence, if estimating the IBM is the goal of source separation algorithms that utilise binary masks, then a metric that quantifies the extent to which a calculated mask is ideal should be a suitable choice. Furthermore, observations made by Li & Loizou (2008) point out that the pattern of the binary mask is more important for speech intelligibility than the local SNR of each T–F unit because the pattern of the mask may help to direct auditory attention. This suggests that the metric should consider the pattern of the mask without weighting the contributions of each T–F unit according to its local SNR.

Such a metric was proposed by Hu & Wang (2007). Their metric assesses segmentation performance and is based on a metric proposed by Hoover et al. (1996) for assessing image segmentation. Hu & Wang's (2007) metric compares ideal segments with calculated segments. Consequently, in their approach there are several outcomes of the comparison; segments can be identified as:

- *Correct*: The calculated and ideal segments significantly overlap
- *Under-segmented*: A calculated segment covers two or more ideal segments
- *Over-segmented*: An ideal segment covers two or more calculated segments

- *Mismatch*: The calculated segment significantly covers a T–F region belonging to the ideal background.
- *Missing*: The calculated segment completely covers a T–F region belonging to the ideal background.

However, not all algorithms utilise segmentation in this way and hence this metric may not be employable by all algorithms.

The aforementioned study performed by Li & Loizou (2008) demonstrated the effects on speech intelligibility of binary mask error, i.e. the percentage of T–F units that are incorrectly labelled when compared to the IBM. Their study demonstrated a strong negative correlation between binary mask error and speech intelligibility. This implies that, at least for anechoic speech, estimating the binary mask error can predict the speech intelligibility of a binary mask.

When comparing the ideal and calculated masks, each T–F unit from the calculated mask can be either correct (if it matches the corresponding unit in the ideal mask) or incorrect in one of two ways. Cases where the ideal target is incorrectly identified (the calculated mask is 0 when it should be 1, or "miss" error (Li & Loizou 2008)) may, in a worst case scenario, result in an important target source unit not contributing to the output. Cases where the ideal background is incorrectly identified (the calculated mask is 1 when it should be 0, or "false alarm" (Li & Loizou 2008)) may result, in a worst case scenario, in masking of the source by the interferer or other noise. Li & Loizou (2008) find that for speech intelligibility false alarm errors are more detrimental than miss errors. Empirical evidence for the effects of these two error types in other applications has not been found but the relative significance of each error type may well be application–specific, with miss errors being more important in some applications where speech intelligibility is not the primary concern. Therefore to calculate the metric, and to retain its independence of application, both errors are here weighted equally. Note that this could be adapted to suit a particular application by adjusting the error weighting to be more sensitive to either error type.

Consequently, the Ideal Binary Mask Ratio (IBMR) is proposed as a metric for assessing source separation algorithms that utilise binary masks. IBMR is an adapted and generalised form of binary mask error (Li & Loizou 2008) or labelling accuracy (Woodruff & Wang 2010). IBMR provides an intuitive score in the interval [0,1] for a mask, based on its correspondence to the IBM, rather than assessing the resynthesised output. IBMR is obtained by comparing the calculated and ideal masks:

$$\text{IBMR} = \frac{\lambda}{\lambda + \rho} \tag{5.16}$$

where

$$\lambda = \sum_{i,l} \mathbf{m}(i,l) \wedge \mathbf{m}_{\text{ibm}}(i,l), \tag{5.17}$$

$$\rho = \sum_{i,l} \mathbf{m}(i,l) \oplus \mathbf{m}_{\text{ibm}}(i,l), \tag{5.18}$$

$\wedge$ denotes binary logical AND and $\oplus$ denotes binary logical XOR. It can be seen from the above equation that good performance is achieved by minimising the difference between the calculated and ideal masks, ρ.

The IBMR is demonstrated in Figure 5.3(g) and Figure 5.3(h). The data are in general agreement, in terms of trends, with the anechoic SNR data and the RSNR and SINR data. IBMR is consistent across all of the acoustic conditions, thus eliminating the inconsistencies demonstrated by SNR, RSNR and SINR because the calculation of the metric does not consider the re-synthesised output. Furthermore, the similarity in trends provides further justification for the employed error weighting procedure.

The experiment described in this chapter was also conducted with a wider range of stimuli as suggested by Cooke (1991). The results were reported in (Hummersone et al. 2011) and found to be very similar to those presented in this chapter.

5.5 Summary and Conclusions

This chapter addressed the research question *6′. How should the performance of separation algorithms incorporating different precedence models be evaluated? What signals? What metrics?* The algorithms will be tested in a range of mixture conditions that incorporate a range of source–target azimuthal separations, TIRs, interferer signals and RT_{60}s, using a metric that facilitates meaningful comparison between different algorithms and across different acoustic conditions. The target signals will be female speech; the interfering signals will be male speech, music and noise. The Binaural Room Impulse Responses (BRIRs) will be captured in real rooms. The chapter proposed a novel metric that meets the above criterion and is suitable for assessing source separation algorithms that aim to calculate the IBM. Specifically, it was shown that whilst the IBM may, in certain conditions, be optimal in terms of SNR (a widely used metric), this was shown not to always be the case when convolutional distortions are introduced. Furthermore, with all other factors being equal (including the calculated mask), SNR–based metrics show inconsistency across different acoustic conditions. To address this problem, the proposed metric—Ideal Binary Mask Ratio (IBMR)—compares the calculated binary mask with the IBM. The metric is robust to the contribution of convolutional distortion to the output because it compares the pattern of the calculated and ideal masks without weighting the contribution of

each unit according to its local SNR. The proposed metric facilitates meaningful and direct comparison of separation algorithms, in particular in situations where acoustic conditions can not be held constant, or where it is important that the results should not be skewed by a particular set of acoustic conditions.

Chapter

6 Modelling Precedence for Source Separation

In Chapter 4 it was determined that modelling the precedence effect for source separation offered the most scope for improvement in reverberant environments. This chapter therefore addresses the research question: *7. Which approaches work best and are there any lessons to be learned for future development?*

Work done so far on incorporating precedence into source separation is based on the work of Palomäki et al. (2004b) (see also Park & Stern 2007). Their separation algorithm includes a simple precedence model that is based on the work of Zurek (1987) and Martin (1997). However, many computational models of precedence have been suggested that are all markedly different in their implementation. Therefore this chapter describes a study that has two objectives: firstly, to implement the model of Palomäki et al. (2004b) and test it using the procedure described in Chapter 5 and secondly, to replace the precedence model with those computational precedence models already proposed in the literature. The study is intended to answer the above research question by indicating perceptually–relevant processing techniques that give the greatest performance of the separation algorithm. Therefore, the chapter will firstly describe the baseline algorithm (Section 6.1). Following this it will describe each of the implemented precedence models and their incorporation into the baseline algorithm (Section 6.2). The experimental procedure is summarised in Section 6.3 and the results are presented and discussed in Section 6.4. The chapter is summarised and concluded in Section 6.5.

It should be noted at this point that this study is designed to test the performance of the combination of the numerous computational precedence models and the source separation algorithm. No judgements are or will be made about the technical quality, biological plausibility or even the localisation accuracy of the models, although clearly the latter will have a significant influence on the separation performance.

6.1 The Baseline Algorithm

This section will first present the baseline separation algorithm (section 6.1.1), which is heavily based upon the aforementioned work of Palomäki et al. (2004b) (note: although every attempt has been made to follow the principles of this algorithm, due

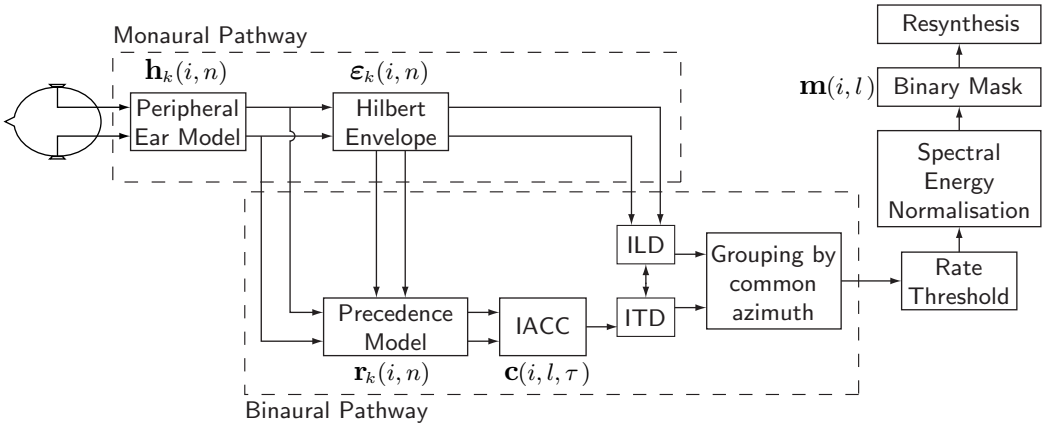


Figure 6.1: Schematic of the baseline separation algorithm and precedence model based on Palomäki et al.'s (2004b) model.

to implementation issues and modifications required to enable the evaluation method described in Chapter 5, the processing utilised is not identical). The work includes a simple precedence model, detailed in section 6.1.2. The architecture of the baseline algorithm is summarised in Figure 6.1.

6.1.1 The Baseline Separation Algorithm

The algorithm attempts to estimate the relative strength of two competing signals arising from spatially–separate sound sources. As shown in Figure 6.1, the binaural left and right signals are first passed through a gammatone filterbank (see Section 3.1.1, page 21) to simulate cochlear frequency selectivity (32 channels are employed, in the range 50–7500 Hz, equally spaced on the ERB-rate scale). The outputs of the gammatone filterbank are then half-wave rectified as a crude model of the IHCs; the results are denoted $\mathbf{h}_L$ and $\mathbf{h}_R$. The Hilbert envelopes ε_k (for ear k) of each of these signals are used to estimate the auditory nerve firing rate $\mathbf{u}_k$ as in Equation 5.8 (page 83). The precedence model, discussed below, is then introduced to inhibit the fine structure of the gammatone filterbank outputs. The cross-correlograms $\mathbf{c}$ for each frame are obtained by cross-correlating this precedence–modelled fine structure $\mathbf{r}_k$ (described in Section 6.1.2) over a three-frame rectangular window thus:

$$\mathbf{c}(i,l,\tau) = \sum_{d=0}^{3L-\tau-1} \mathbf{r}_L\big(i,(l-1)L+d+\tau\big)\mathbf{r}_R\big(i,(l-1)L+d\big) \tag{6.1}$$

where τ denotes the discrete lag (representing ITD) of the cross-correlation such that $\{\tau \in \mathbb{Z} : -T \leq \tau \leq T\}$, $T = 1$ ms (in samples), L denotes the frame length (10 ms, in samples) and $\mathbb{Z}$ is the set of integers.

The data from the cross-correlograms are subsequently warped from ITD to azimuth to yield $\mathbf{c}(i,l,\phi)$, where ϕ denotes azimuthal angle such that $\{\phi \in \mathbb{Z} : -90^\circ \leq \phi \leq 90^\circ\}$,

since the relationship between ITD and azimuth is frequency-dependent (Kuhn 1977). The warping function is derived from Kuhn's work. Specifically,

$$\mathrm{ITD} = \frac{\Pi r \sin\phi}{c_0} \tag{6.2}$$

where Π varies with frequency f (in Hz) such that

$$\Pi = \begin{cases} 3 & f \leq 500 \\ 2.5 + 0.5\cos\left(\pi \dfrac{\log_2 \frac{\sqrt{6}f}{1250}}{\log_2 6}\right) & 500 < f < 3000 \\ 2 & f \geq 3000 \end{cases} \tag{6.3}$$

where c_0 is the speed of sound (344 ms^{-1}) and r is the effective radius of the head, which Kuhn derives as 0.093 m, somewhat larger than typical skull perimeter measurements, perhaps due to protruding features such as the nose and pinnae. Since Kuhn is not specific about the change in Π between 500 and 3000 Hz, a raised cosine function is chosen to vary Π "smoothly".

The azimuthal-domain cross-correlograms are then transformed to skeleton cross-correlograms (see Section 3.1.5, page 27), except that it is performed in the azimuthal domain. The standard deviations utilised in the procedure are chosen thus:

$$\sigma(i) = 4.5 - (i-1)\frac{3.75}{I-1} \tag{6.4}$$

where $\{i \in \mathbb{N} : 1 \leq i \leq I\}$, $\mathbb{N}$ is the set of natural numbers and I is the number of channels (32). The skeleton cross-correlograms are subsequently pooled across frequency and time thus:

$$\bar{\mathbf{s}}(\phi) = \sum_{i,l} \mathbf{s}(i,l,\phi) \tag{6.5}$$

This pooled skeleton cross-correlogram is used to obtain 'global' estimates of the target signal and interferer azimuths (ϕ_t and ϕ_n respectively), which are identified using the following procedure:

$$\phi_t = \min(\phi_1, \phi_2), \tag{6.6}$$

$$\phi_n = \max(\phi_1, \phi_2) \tag{6.7}$$

where

$$\phi_1 = \arg\max_{\psi_\phi} \bar{\mathbf{s}}(\psi_\phi), \tag{6.8}$$

$$\phi_2 = \arg\max_{\psi_\phi} \{\bar{\mathbf{s}}(\psi_\phi) : \phi_1 \notin \psi_\phi\} \tag{6.9}$$

and $\big\{\psi_\phi \in \phi : \big(\bar{\mathbf{s}}(\phi) - \bar{\mathbf{s}}(\phi-1)\big)\big(\bar{\mathbf{s}}(\phi) - \bar{\mathbf{s}}(\phi+1)\big) > 0\big\}$. Note that the target is

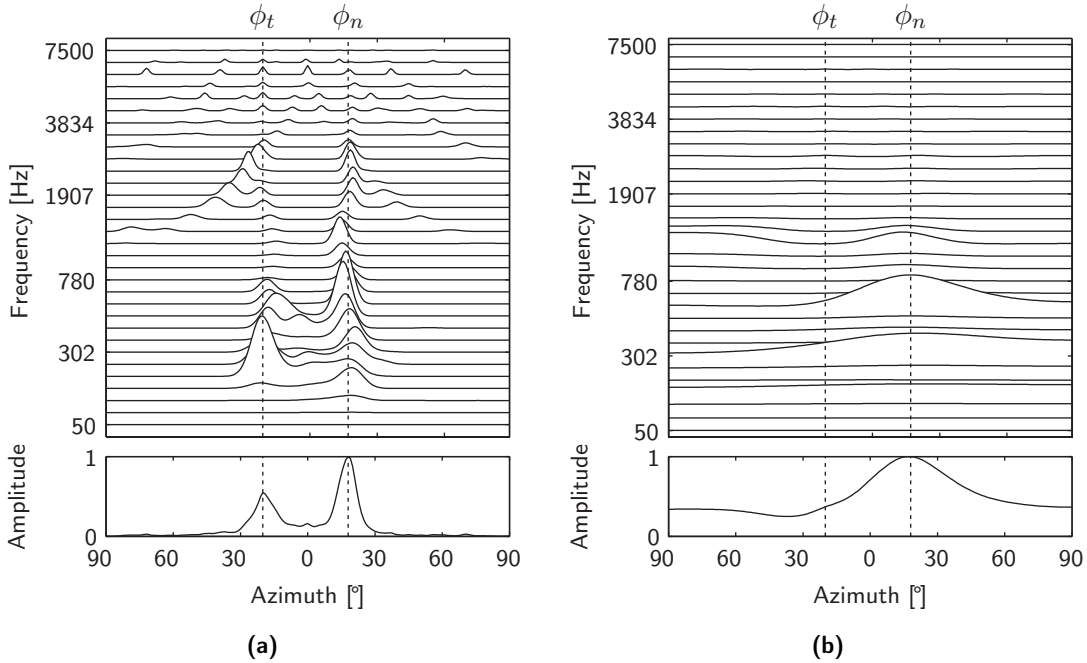


Figure 6.2: The grouping procedure for a mixture of female and male speech speech at $\mp 20^\circ$. **(a)** The skeleton cross-correlogram and pooled skeleton cross-correlogram for the entire stimulus, indicating the target and interferer azimuths (ϕ_t and ϕ_n respectively). **(b)** The cross-correlogram and pooled cross-correlogram for a single frame of the mixture in which the interfering male speech is more prominent.

consistently placed on the left and thus the azimuths are assigned accordingly. The azimuthal cross-correlograms are used to calculate the binary T–F mask $\mathbf{m}$ by making 'local' estimates of the relative strength of the target and interfering signals at the obtained global azimuths thus:

$$\mathbf{m}(i,l) = \begin{cases} 1 & \text{if } \mathbf{c}(i,l,\phi_t) > \mathbf{c}(i,l,\phi_n) \\ & \text{and } 10\log_{10}\left(\dfrac{\mathbf{c}(i,l,\phi_t)}{\mathring{\mathbf{c}}}\right) > \Theta_c \\ 0 & \text{otherwise} \end{cases} \tag{6.10}$$

where

$$\mathring{\mathbf{c}} = \max_{i,l,\phi} \mathbf{c}(i,l,\phi) \tag{6.11}$$

Generally Θ_c was set to -160 dB. An example of the grouping procedure is shown in Figure 6.2.

Two further checks are then performed on the mask. Firstly, the ILD value for each T–F unit in frequency channels above 2.8 kHz (denoted v) that has a corresponding mask value of one is checked against an ILD template $\boldsymbol{\zeta}$ to ensure azimuthal estimate consistency. The ILD template is the ideal value of ILD in each frequency channel v at the target azimuth and was calculated using white noise. A zero is written to the

mask if the ILD value deviates from the template by more than 1 dB:

$$\mathbf{m}(v,l) = \begin{cases} 0 & \text{if } |\mathrm{ILD}(v,l) - \boldsymbol{\zeta}(v,\phi_t)| > 1 \text{ dB} \\ \mathbf{m}(v,l) & \text{otherwise} \end{cases} \tag{6.12}$$

where

$$\mathrm{ILD}(i,l) = 10\log_{10}\left(\frac{\acute{\mathbf{u}}_L(i,l)}{\acute{\mathbf{u}}_R(i,l)}\right) \tag{6.13}$$

and auditory energy $\acute{\mathbf{u}}$ was calculated as in Equation 5.15. Secondly, energy values where the corresponding mask value is one are compared to a running energy average $\boldsymbol{\Xi}$, calculated in each frequency channel over a 200 ms (20 frame) window with 100 ms (10 frame) overlap. If the ratio of these values exceeds a rate threshold then a zero is written to the mask thus:

$$\mathbf{m}(i,l) = \begin{cases} 0 & \text{if } 10\log_{10}\left(\dfrac{\acute{\mathbf{u}}_{LR}(i,l)}{\boldsymbol{\Xi}(i,l)}\right) > \Theta_r \\ \mathbf{m}(i,l) & \text{otherwise} \end{cases} \tag{6.14}$$

where

$$\mathbf{u}_{LR} = \left(\frac{1}{2}\left(\mathbf{u}_L^{3.333}(i,l) + \mathbf{u}_R^{3.333}(i,l)\right)\right)^{0.3}, \tag{6.15}$$

$\acute{\mathbf{u}}_{LR}$ was calculated as in Equation 5.15 (page 84) and Θ_r is the rate threshold set to -11 dB.

Lastly, in order to undo the spectral envelope distortion introduced by reverberation, a normalisation factor is calculated that is applied at resynthesis. The resynthesis procedure is described by Brown & Cooke (1994a); the normalisation factor is divided by its corresponding frequency channel before they are summed. The factor is calculated as the mean of the largest values of the auditory nerve firing rate $\mathbf{u}_{LR}$ for which the corresponding mask value is one. The number of units over which this calculation is performed depends upon the input signal channel: the maximum is K/B, where K is the number of time frames in the input signal and $B = 15$; if the number of reliable units is less than this number then it is set to the number of reliable units. Note that although this procedure is performed in order for the output to be auditioned, it is not taken into account in the evaluation since the metric only considers the binary mask and not the resynthesised output.

In order to calculate the Ideal Binary Mask Ratio (IBMR) and assess the performance of the model, the IBM is calculated as in Section 5.2 (Equation 5.3, page 80).

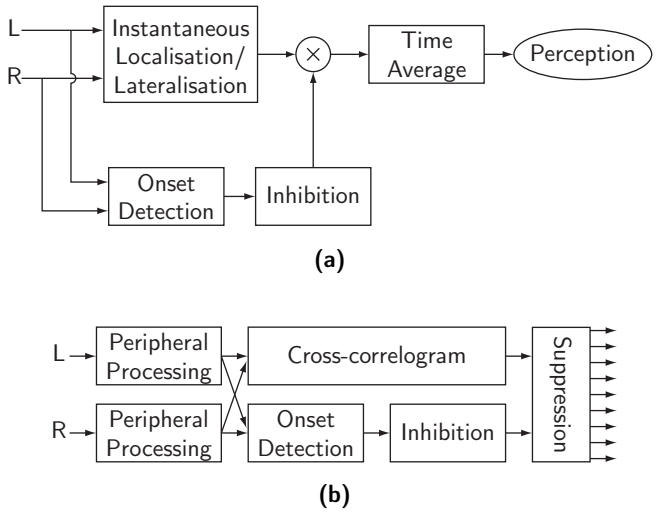


Figure 6.3: Modelling the precedence effect. **(a)** Schematic of Zurek's (1987) precedence model. **(b)** Martin's (1997) computational implementation of Zurek's (1987) model.

6.1.2 The Baseline Precedence Model

A precedence model is introduced into the baseline algorithm in order to enhance the local and global estimates of the target and interferer azimuths. The precedence model incorporated into the baseline algorithm is based on the work of Zurek (1987) and Martin (1997), the latter of which is a computational implementation of the former. Schematics of both models are given in Figure 6.3.

The upper path of the model in Figure 6.3(a) considers steady-state signals; localisation is achieved by a running average over the past and present and is formed by a combination of ILD and ITD (Zurek 1987). The lower path of the model takes effect when sharp onsets are present in the signal. When such an onset is detected, a brief period of inhibition is triggered that suppresses the contribution of the upper path for a period of about 5 ms after the onset. The inhibition takes place about 500 μs after the onset.

The implementation of the baseline precedence model (Palomäki et al. 2004b) is an adaptation of Martin's (1997) model (see Figure 6.3(b) and Section 6.2.1). Specifically, the model employs on onset-de-emphasising low-pass filter with an impulse response of the form:

$$h_{lp}(n) = Ane^{-n/\alpha_p} \tag{6.16}$$

where α_p is a time constant chosen to be the number of samples corresponding to 15 ms and A is set to give unity gain at DC. This is used to filter the Hilbert envelope ε_k to produce an "inhibitory signal". This inhibitory signal is then subtracted from the half-wave rectified gammatone filterbank fine structure. The process is summarised in

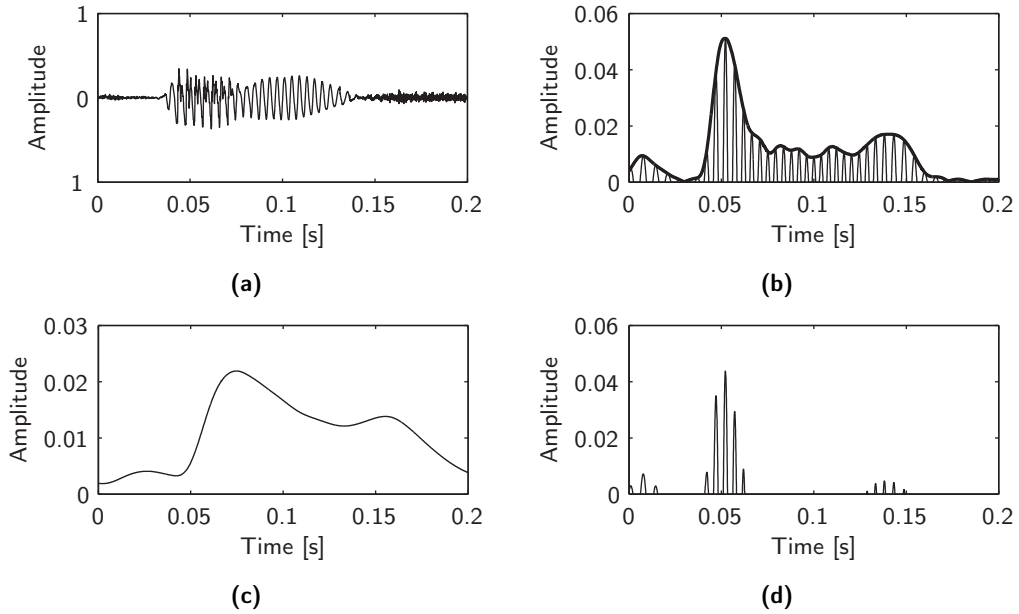


Figure 6.4: Examples of the processing in the baseline precedence model. **(a)** Input waveform (excerpt of male speech). **(b)** Half-wave rectified gammatone filter output (153 Hz frequency channel) showing the fine structure and Hilbert envelope. **(c)** The onset-de-emphasised low-pass filtered signal envelope. **(d)** The inhibitied fine structure.

the following way:

$$\mathbf{r}_k(i,n) = \max\Big(\mathbf{h}_k(i,n) - G\big(h_{lp}(n) * \varepsilon_k(i,n)\big), 0\Big) \tag{6.17}$$

where G is an inhibitory gain factor that is set to 1. The precedence–modelled fine structure $\mathbf{r}$ is used to obtain the cross-correlograms (see Equation 6.1). Examples of this processing are given in Figure 6.4.

Zurek (1987) notes that inhibited information is only used in localisation and that reverberation makes a significant contribution to the timbral and spatial characteristics of a perceived sound. The baseline algorithm reflects this by only using precedence–modelled information in the localisation aspect of the algorithm. Also note that this model can account for the monaural precedence effects discussed in Section 4.3.4, since the inhibition is applied separately for each ear.

6.2 Replacing the Precedence Model

This section details the incorporation of numerous computational precedence models with the baseline separation algorithm. In order to attempt to improve the performance of the baseline separation algorithm, each of a selection of the numerous computational precedence and binaural localisation models proposed in the literature was incorporated into the algorithm. Models proposed by Martin (1997) (Section 6.2.1), Faller &

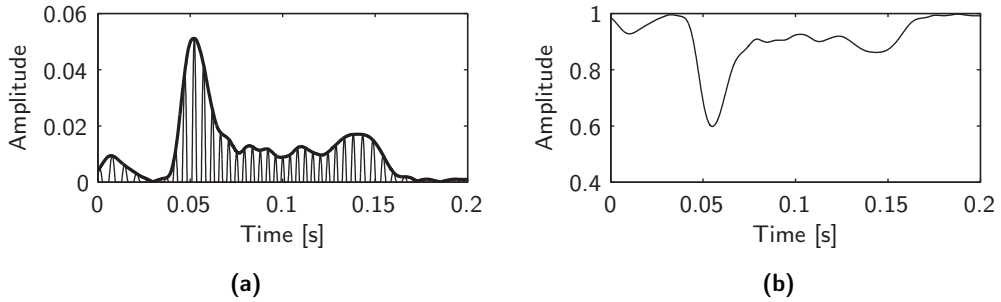


Figure 6.5: Examples of the processing in Martin's precedence model. **(a)** Half-wave rectified gammatone filter output as in Figure 6.4(b). **(b)** The resulting inhibitory signal.

Merimaa (2004) (Section 6.2.2), Lindemann (1986a,b) (Section 6.2.3) and Macpherson (1991) (Section 6.2.4) are presented. In each case, much of the baseline algorithm is retained, but the precedence and localisation—and in some cases parts of the peripheral processing—routines are replaced by those proposed by the model under test.

6.2.1 Martin's Model

Martin's (1997) work is the basis for the precedence model employed in the baseline algorithm and hence is an obvious first choice of model to incorporate and test. The perceptual theory behind the model has already been given in Section 6.1.2. Unfortunately, the paper is lacking some crucial details necessary to implement the model accurately. Specifically, Martin's paper lacks details regarding the filter to calculate the "excitation envelope" and about the numeric levels of the numerous signals that are calculated. However, there is only one conceptual difference between the baseline precedence model and Martin's model: the point at which the inhibition is applied (compare Figure 6.3(b) with Figure 6.1). In the baseline model, inhibition is applied to the fine structure before it is cross-correlated, whereas in Martin's model inhibition is applied to the running cross-correlation. Consequently, the implementation of Martin's model is heavily based upon the baseline precedence model. An example of the processing is shown in Figure 6.5.

In the implementation, firstly the "excitation envelope" $\mathbf{x}$ is calculated from the Hilbert envelope thus:

$$\mathbf{x}_k(i,n) = \boldsymbol{\varepsilon}_k(i,n) * h_{lp}(n) \tag{6.18}$$

where h_{lp} was given in Equation 6.16, except that in this case the time constant $\alpha_p = \alpha_m = 1.5$ ms. Following this, a mono excitation envelope $\mathbf{x}_{LR}$ is calculated:

$$\mathbf{x}_{LR}(i,n) = \frac{1}{2}\big(\mathbf{x}_L(i,n) + \mathbf{x}_R(i,n)\big) \tag{6.19}$$

and subsequently normalised independently for each frequency channel to be in the

range [0,1]. The inhibitory signal $\boldsymbol{\iota}$ is calculated from this excitation envelope thus:

$$\boldsymbol{\iota}(i,n) = \max\Big(1 - \big(G \cdot \mathbf{x}_{LR}(i,n)\big), 0\Big) \tag{6.20}$$

The inhibited running cross-correlation $\mathbf{c}_\iota$ is then calculated in the following way:

$$\mathbf{c}_\iota(i,n,\tau) = \boldsymbol{\iota}(i,n)\,\acute{\mathbf{c}}(i,n,\tau) \tag{6.21}$$

where

$$\acute{\mathbf{c}}(i,n,\tau) = \mathbf{h}_L\big(i, \max(n+\tau, n)\big)\mathbf{h}_R\big(i, \max(n-\tau, n)\big) \tag{6.22}$$

Finally, these cross-correlations are averaged over a three-frame rectangular window to produce the cross-correlograms:

$$\mathbf{c}(i,l,\tau) = \frac{1}{3M}\sum_{d=1}^{3M} \mathbf{c}_\iota\big(i, (l-1)L + d, \tau\big) \tag{6.23}$$

As with the following models, subsequent processing of the cross-correlograms, grouping and separation routines is identical to that described in Section 6.1. Note that unlike the baseline model, this model can not account for monaural precedence effects since the inhibitory signal is monophonic.

6.2.2 Faller & Merimaa's Model

The model proposed by Faller & Merimaa (2004) differs from other computational precedence models by suggesting that some precedence effects can be modelled by calculating Interaural Coherence (IC). Specifically, if a dichotic signal is coherent then this is a good indication that the obtained ITD and ILD correspond to the sound's true direction. IC $\boldsymbol{\chi}$ is calculated in each frequency band as the maximum value of the running normalised cross-correlation $\hat{\mathbf{c}}$:

$$\boldsymbol{\chi}(i,n) = \max_{\tau} \hat{\mathbf{c}}(i,n,\tau) \tag{6.24}$$

This gives a result in the interval [0,1], with a value of one indicating that the signals are perfectly coherent and hence that the elicited cues are indicative of the sound's true direction. It is therefore necessary to specify a threshold for cue selection. According to Faller & Merimaa, this is a trade-off between selecting reliable cues that correspond closely to free-field conditions and maximising the proportion of the input signals that contributes to localisation. They also note that this threshold is likely to adapt to the acoustical environment.

In terms of implementation, the first stage of the model is the peripheral auditory processing. Faller & Merimaa suggest the use of a model of neural transduction proposed by Bernstein et al. (1999). This model recreates the compression and

half-wave rectification that has been observed by numerous researchers in auditory physiology, but does not enhance onsets. The employed process is summarised as follows:

- Each Hilbert envelope output of the gammatone filterbank ε_k is compressed by raising it to the power 0.23 and then squared
- This envelope is then filtered with a fourth-order Finite Impulse Response (FIR) low-pass filter with a cut-off frequency of 425 Hz
- The resulting envelopes $\acute{\varepsilon}_k$ are half-wave rectified and then re-combined with the half-wave rectified gammatone filterbank output thus:

$$\mathbf{h}_k(i,n) = \frac{\acute{\varepsilon}_k(i,n)}{\varepsilon_k(i,n)} \max\big(\boldsymbol{\gamma}_k(i,n), 0\big) \tag{6.25}$$

  where $\mathbf{h}_k$ is the modelled IHC response and $\boldsymbol{\gamma}_k$ is the gammatone filter fine structure.

The cross-correlograms are calculated using the IHC–modelled data. As stated above, this model requires the calculation of normalised running cross-correlation, which is of the form

$$\hat{\mathbf{c}}(i,n,\tau) = \frac{\acute{\mathbf{c}}(i,n,\tau)}{\sqrt{\mathbf{a}_L(i,n,\tau)\mathbf{a}_R(i,n,\tau)}} \tag{6.26}$$

where

$$\acute{\mathbf{c}}(i,n,\tau) = \frac{1}{\alpha_f}\mathbf{h}_L\big(i,\max(n+\tau,n)\big)\mathbf{h}_R\big(i,\max(n-\tau,n)\big) + \left(1-\frac{1}{\alpha_f}\right)\acute{\mathbf{c}}(i,n-1,\tau), \tag{6.27}$$

$$\mathbf{a}_L(i,n,\tau) = \frac{1}{\alpha_f}\mathbf{h}_L^2\big(i,\max(n+\tau,n)\big) + \left(1-\frac{1}{\alpha_f}\right)\mathbf{a}_L(i,n-1,\tau), \tag{6.28}$$

$$\mathbf{a}_R(i,n,\tau) = \frac{1}{\alpha_f}\mathbf{h}_R^2\big(i,\max(n-\tau,n)\big) + \left(1-\frac{1}{\alpha_f}\right)\mathbf{a}_R(i,n-1,\tau) \tag{6.29}$$

and α_f is the time constant of the exponentially decaying window, chosen to be the number of samples corresponding to 10 ms. The cross-correlograms are calculated by averaging only the running normalised cross-correlations within a given frame for which the corresponding IC value $\boldsymbol{\chi}$ exceeds a threshold value Θ_χ:

$$\mathbf{c}(i,l,\tau) = \begin{cases} 0 & \text{if } \Psi = \varnothing \\ \dfrac{1}{|\Psi|}\displaystyle\sum_{d\in\Psi}\hat{\mathbf{c}}(i,d,\tau) & \text{otherwise} \end{cases} \tag{6.30}$$

where $\{\Psi \in n : (l-1)L+1 \le n \le lL, \boldsymbol{\chi}(i,n) \ge \Theta_\chi\}$, $\boldsymbol{\chi}$ was given in Equation 6.24, $\varnothing$ is the empty set and Θ_χ is chosen to be 0.5, corresponding to 2 simultaneous and

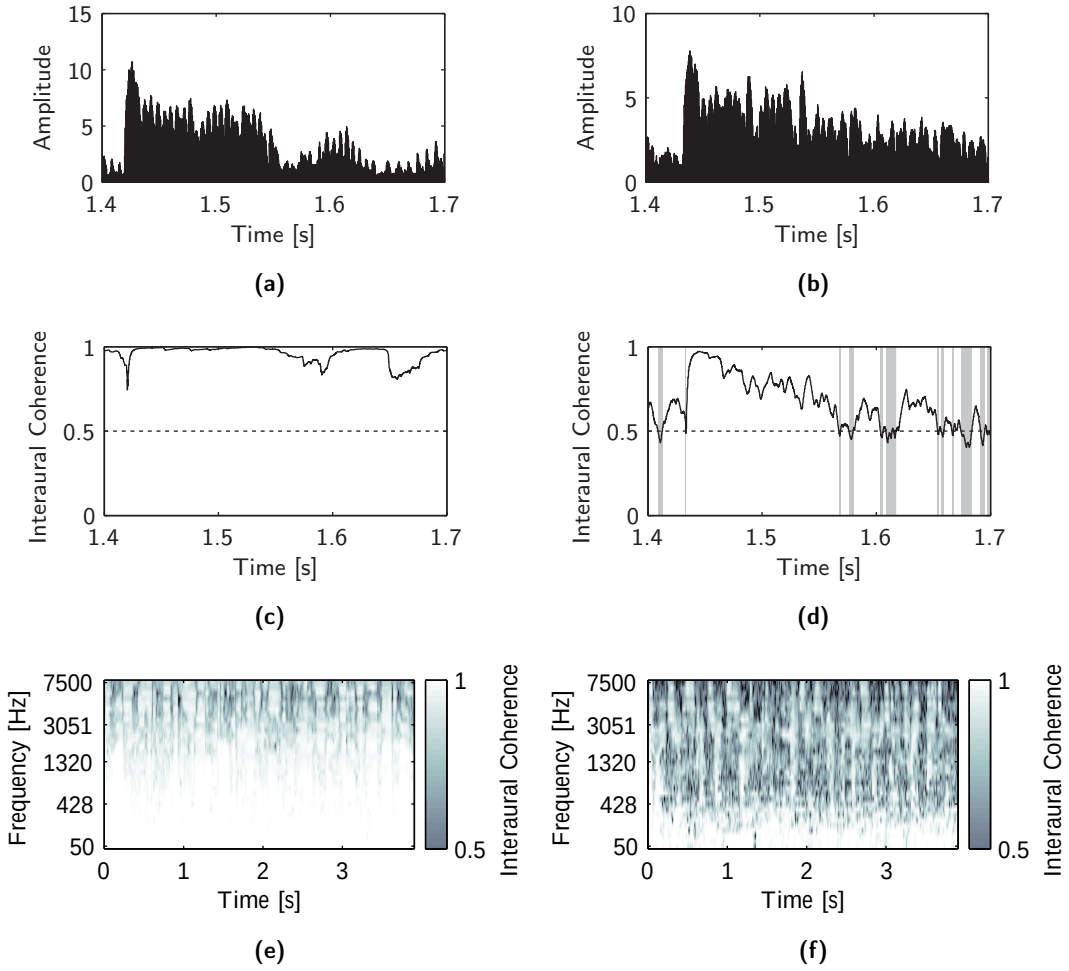


Figure 6.6: Examples of the processing in Faller & Merimaa's model for a mixture of male and female speech. **(a)** An excerpt of the IHC–modelled data in the 1.9 kHz frequency channel with $RT_{60} = 0$ s. **(b)** The IHC–modelled data in the same frequency channel with $RT_{60} = 0.89$ s. **(c)** The IC signal with $RT_{60} = 0$ s; the dashed line shows the IC threshold $\Theta_\chi = 0.5$. All regions contribute to localisation. **(d)** The IC signal with $RT_{60} = 0.89$ s. Greyed regions do not contribute to localisation. **(e)** The time–frequency IC for the entire signal with $RT_{60} = 0$ s. Black regions are below the IC threshold. **(f)** The time–frequency IC with $RT_{60} = 0.89$ s.

coherent onsets arising from 2 statistically–independent sound sources. Examples of this processing are given in Figure 6.6. Note that, because this model requires both ear signals in order to calculate IC, it can not account for monaural precedence effects.

6.2.3 Lindemann's Model

Lindemann's (1986a) model (see also Lindemann 1986b) can be considered as an extension of Jeffress' (1948) original cross-correlation theory of sound localisation. The model is extended with two components: "monaural detectors" and a "contralateral-inhibition mechanism" (an inhibition along the τ-axis). This inhibition is achieved through two components: a static inhibition component and a dynamic inhibition

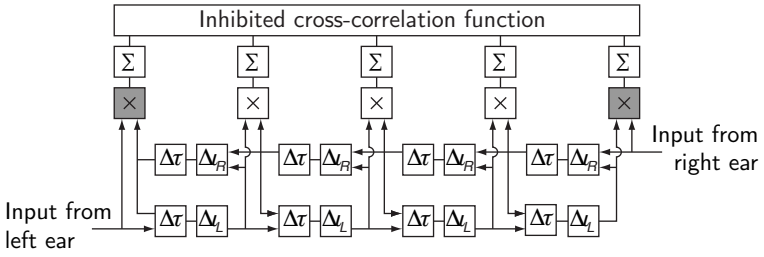


Figure 6.7: The architecture of Lindemann's binaural localisation model (Lindemann 1986a). Adapted from (Braasch 2005; Lindemann 1986a).

component, the latter of which is intended to simulate the precedence effect. Although intended for stationary signals, the cross-correlation–based architecture lends itself well to this application. However, the suitability of the model to non-stationary signals remains unclear.

The architecture of the localisation model is summarised in Figure 6.7. The inhibition is derived from the contralateral signals and also from previous calculations of the cross-correlation. Furthermore, the inhibition is triggered by peaks in the primary cross-correlation and decays with a time constant of 10 ms. Additionally, monaural detectors (indicated by the grey multiplication boxes at the beginning of each delay line in Figure 6.7) are included in order to lateralise the input even if only one ear signal is present and cross-correlation fails. The model can therefore account for monaural precedence effects.

In terms of implementation, the peripheral auditory processing of the baseline algorithm is retained since Lindemann states that the exact nature of the peripheral processing is inconsequential to the operation of the model. According to Lindemann, the first step is to normalise the binaural signals to have a maximum value of one. However, the input to the model is critical to its operation; this is discussed towards the end of the section (see 'The Operating Point'). Following this, the modified inputs to the model, $\acute{\mathbf{h}}_L$ and $\acute{\mathbf{h}}_R$, are defined thus:

$$\acute{\mathbf{h}}_L(i, n+1, \tau+1) = \begin{cases} \acute{\mathbf{h}}_L(i, n, \tau)\boldsymbol{\iota}_L(i, n, \tau) & -T \leq \tau \leq T-1 \\ \mathbf{h}_L(i, n+\tau) & \tau = T \end{cases} \tag{6.31}$$

$$\acute{\mathbf{h}}_R(i, n+1, \tau-1) = \begin{cases} \acute{\mathbf{h}}_R(i, n, \tau)\boldsymbol{\iota}_R(i, n, \tau) & -T+1 \leq \tau \leq T \\ \mathbf{h}_R(i, n+\tau) & \tau = -T \end{cases} \tag{6.32}$$

where T is the maximum lag in samples. Note here that the outputs of the peripheral processor $\mathbf{h}_L$ and $\mathbf{h}_R$ have had zeros placed between alternate samples in order to halve the sample period. The inhibitory components $\boldsymbol{\iota}_L$ and $\boldsymbol{\iota}_R$ are derived from the

contralateral signal in the following way:

$$\boldsymbol{\iota}_L(i,n,\tau) = \big(1 - \acute{\mathbf{h}}_R(i,n,\tau)\big)\big(1 - \boldsymbol{\Phi}(i,n-1,\tau)\big) \tag{6.33}$$

$$\boldsymbol{\iota}_R(i,n,\tau) = \big(1 - \acute{\mathbf{h}}_L(i,n,\tau)\big)\big(1 - \boldsymbol{\Phi}(i,n-1,\tau)\big) \tag{6.34}$$

Here, $\boldsymbol{\Phi}$ is the dynamic inhibitory component which is derived from the cross-correlation product $\acute{\mathbf{c}}$ in the following way:

$$\boldsymbol{\Phi}(i,n,\tau) = \acute{\mathbf{c}}(i,n-1,\tau) + \boldsymbol{\Phi}(i,n-1,\tau)e^{-T_d/\alpha_{\text{inh}}}\big(1 - \acute{\mathbf{c}}(i,n-1,\tau)\big) \tag{6.35}$$

where T_d is half the sample period and α_{inh} is the fade-off time constant (10 ms). The running cross-correlation is calculated as follows:

$$\acute{\mathbf{c}}(i,n,\tau) = \Big(p(\tau) + \big(1 - p(\tau)\big)\acute{\mathbf{h}}_R(i,n,\tau)\Big)\Big(p(-\tau) + \big(1 - p(-\tau)\big)\acute{\mathbf{h}}_L(i,n,\tau)\Big) \tag{6.36}$$

where p is the monaural sensitivity function such that $p(\tau) = 0.035e^{-(T+\tau)/6}$. The inhibited cross-correlation $\mathbf{c}_\iota$ is calculated from the running cross-correlation using an exponential window thus:

$$\mathbf{c}_\iota(i,n,\tau) = \big(1 - e^{-T_d/T_{\text{int}}}\big)\acute{\mathbf{c}}(i,n,\tau) + e^{-T_d/T_{\text{int}}}\mathbf{c}_\iota(i,n-1,\tau) \tag{6.37}$$

where T_{int} is the integration time constant (5 ms). The cross-correlograms are calculated by averaging the running cross-correlations over the frame:

$$\mathbf{c}(i,l,\tau) = \frac{1}{M}\sum_{d=1}^{M}\mathbf{c}_\iota(i,(l-1)L + d,\tau) \tag{6.38}$$

The Operating Point

One difficulty in Lindemann's (1986a) paper is the discussion of the 'operating point' or 'inhibition parameter' (c_{inh}). The parameter appears to be crucial for controlling the amount of inhibition. Although Lindemann states how it is derived, he does not discuss how it is implemented. Specifically, Lindemann states that:

> The operating point is described by the "inhibition parameter" c_{inh} that is derived from the input signal having the greater amplitude. For pure tones with the amplitudes A_r (right input signal) and A_l (left input signal) the inhibition parameter is
>
> $$c_{\text{inh}} = \max\{A_r, A_l\} \qquad \text{with } 0 \le c_{\text{inh}} \le 1$$
>
> For stationary noise signals c_{inh} was derived analogously, A_r and A_l being the root-mean-square (before half-wave rectification), multiplied by $\sqrt{2}$. The noise signals were clipped after the half-wave rectification to avoid

input signals greater than one.

(Lindemann 1986a)

Clearly, although the inhibition parameter is "derived", there must be a mechanism that aims to achieve a given inhibition parameter (c_{inh}) at the input to the model. Consequently, the input to the model $\mathbf{h}$ is derived in the following way, based on the above description and a target inhibition parameter c_{inh}:

$$\mathbf{h}_k(i,n) = \min\left(\max\left(\frac{c_{\text{inh}}}{c_\gamma(i)}\boldsymbol{\gamma}_k(i,n), 0\right), 1\right) \tag{6.39}$$

where

$$c_\gamma(i) = \max_k \sqrt{\frac{2}{\Lambda}\sum_{n=1}^{\Lambda}\boldsymbol{\gamma}_k^2(i,n)}, \tag{6.40}$$

$\boldsymbol{\gamma}$ is the output of the gammatone filterbank and Λ is the length of the input signal in samples. Lindemann states that the optimal value for $c_{\text{inh}} = 0.3$ and hence this value is employed in the investigation.

6.2.4 Macpherson's Model

Macpherson (1991) proposes a model for stereo imaging measurement. However, since the model is based on cross-correlation, it can be easily adapted for use in this work. The first stage of the model is the peripheral processing, however, there is insufficient information to accurately recreate this stage. Since this stage aims to recreate both the cochlear filtering and the half-wave rectification, adaptation and phase– and envelope–locking seen in auditory nerve responses, a combination of a gammatone filterbank and a Meddis IHC model are utilised in the peripheral processing.

The precedence modelling is introduced through the selection of "analysis points". Macpherson argues that performing a running cross-correlation for the entire signal length is inefficient. Therefore, a set of analysis points (samples) Ψ are chosen where local peaks occur across the left and right ear signals within the cross-correlation window M_c (2 ms, in samples) such that:

$$\Psi = \Psi_L \cap \Psi_R \tag{6.41}$$

where

$$\Psi_L = \big\{n : \big(\mathbf{h}_L(i,n) - \mathbf{h}_L(i,n-1)\big)\big(\mathbf{h}_L(i,n) - \mathbf{h}_L(i,n+1)\big) > 0\big\}, \tag{6.42}$$

$$\Psi_R = \Big\{n+\mu : \big(\mathbf{h}_R(i,n) - \mathbf{h}_R(i,n-1)\big)\big(\mathbf{h}_R(i,n) - \mathbf{h}_R(i,n+1)\big) > 0, \quad \mu \in \mathbb{Z}, \frac{-M_c}{2} \le \mu \le \frac{M_c}{2}, \mu \neq 0\Big\} \tag{6.43}$$

At high frequencies, even with the envelope–locking characteristics of the IHC model, peaks can occur very close together, creating significant overlap of the cross-correlation windows. To reduce this inefficiency, the input is divided into frames of length $M_c/2$ and only the last analysis point from each frame is selected.

The cross-correlation $\acute{\mathbf{c}}$ is calculated for each member of Ψ with the peak at the centre of the cross-correlation window. To simulate the precedence effect, an inhibited cross-correlation is calculated as a weighted average of cross-correlations that fall within the inhibition window 20 ms in length (two frames, in samples) after the initial analysis point. Unfortunately, Macpherson does not specify this weighting function, only stating that peaks that occur within 1–6 ms are suppressed. Consequently, the weighting window proposed by Martin (1997) is adapted and utilised and the inhibited cross-correlation is calculated in the following way:

$$\mathbf{c}_\iota(i,n,\tau) = \begin{cases} 0 & \text{if } \psi = \varnothing \\ \dfrac{1}{|\psi|}\displaystyle\sum_{d\in\psi} w_m(x-n)\acute{\mathbf{c}}(i,d,\tau) & \text{otherwise} \end{cases}, \qquad (n \in \Psi) \tag{6.44}$$

where $\{\psi \subset \Psi : n \leq \Psi \leq n + 2L\}$,

$$\acute{\mathbf{c}}(i,n,\tau) = \frac{1}{M_c+1}\sum_{d=n-\frac{M_c}{2}}^{n+\frac{M_c}{2}} \mathbf{h}_L\big(i,\max(d+\tau,d)\big)\mathbf{h}_R\big(i,\max(d-\tau,d)\big), \qquad (n \in \Psi), \tag{6.45}$$

$$w_m(n) = A\max\left(1 - G\frac{e}{\alpha_m}h_{lp}(n), 0\right), \tag{6.46}$$

h_{lp} was as in Martin's model (see Section 6.2.1, page 98), α_m was defined in Martin's model (set in samples to 1.5 ms), G is the inhibitory gain (set to 1) and A is set to give unity gain at DC. Lastly, these weighted cross-correlations are averaged across the duration of the frame to form the cross-correlograms thus:

$$\mathbf{c}(i,l,\tau) = \begin{cases} 0 & \text{if } \acute{\psi} = \varnothing \\ \dfrac{1}{|\acute{\psi}|}\displaystyle\sum_{d\in\acute{\psi}} \mathbf{c}_\iota(i,d,\tau) & \text{otherwise} \end{cases} \tag{6.47}$$

where $\{\acute{\psi} \subset \Psi : (l-1)L + 1 \leq \Psi \leq lL\}$. An example of this processing is given in Figure 6.8. Note that this processing strategy can account for monaural precedence effects.

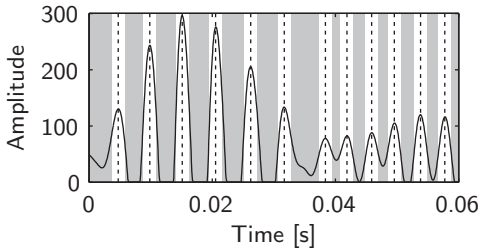


Figure 6.8: Example of the processing in Macpherson's precedence model showing the left ear signal, the analysis points (vertical dashed lines) and cross-correlation windows (in white) for the 200 Hz frequency channel. Grey regions do not contribute to localisation.

6.3 Experimental Procedure

The experimental procedure, mixture conditions, BRIRs and choice of metric were discussed in detail in Chapter 5. To summarise, the following mixture conditions were chosen in order to evaluate the algorithm:

- Target/interferer azimuthal separations of 10°, 20° and 40° (i.e. ±5°, ±10° and ±20° with respect to the frontal median plane), with the target on the left
- Target-to-Interferer Ratios (TIRs) of 0, 10 and 20 dB (RMS)
- The following interfering signals: white noise, male speech and a modern piece of rock music
- RT_{60}s of 0, 0.32, 0.47, 0.68 and 0.89 seconds

This range of conditions was employed to ensure that the performances (reported later) were representative of a range of realistic conditions offering a varying degree of difficulty. However, the research is not explicitly concerned with the performance of the algorithms in each of the mixture conditions. Hence, only RT_{60} will be compared in the results, with model performances reported as means calculated across the other variables.

6.4 Results and Discussion

The results from the study are given in Figure 6.9. The plot shows IBMR versus RT_{60} with the data averaged over all experimental conditions. The data are compared to "No Inhibition", i.e. the data obtained from the baseline algorithm except that the precedence model is bypassed by setting $G = 0$. Plotting the data obtained without precedence processing demonstrates the performance gain achieved by each of the precedence models.

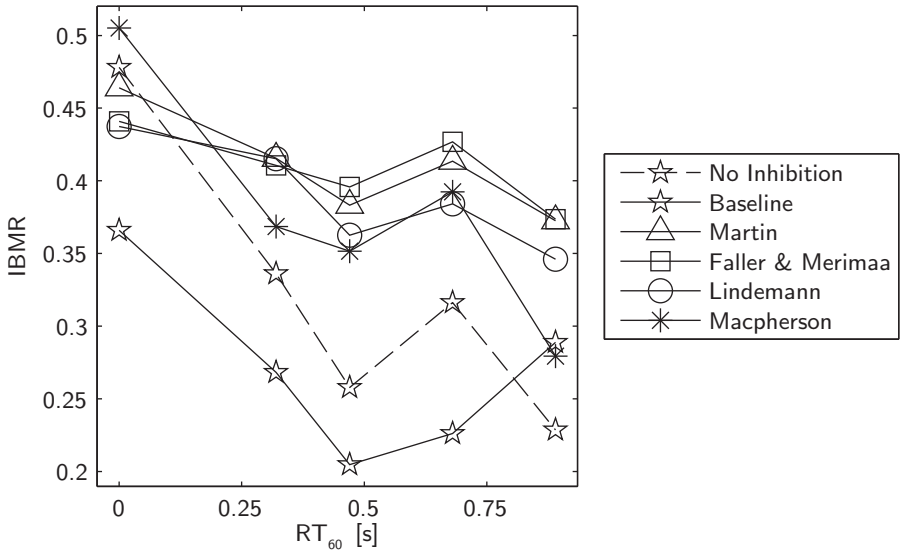


Figure 6.9: Mean model performances showing IBMR versus RT_{60}.

Further analysis of the data was conducted via a univariate ANalysis Of VAriance (ANOVA). The full ANOVA table is included in Appendix B. The analysis was carried out by treating each interfering stimulus as a trial and hence the ANOVA analysis does not include this as a variable. This was performed under the assumption that the interfering stimulus would have less of an effect than the other variables.

The ANOVA shows that the effect of the model is significant (sig. < 0.05) and has a large effect (partial $\eta^2 = 0.246$) on IBMR; it is second only to the effect due to the azimuthal separation. The mean performances of the models from the ANOVA is shown in Figure 6.10. The figure includes 95% confidence intervals from the ANOVA rather than from the raw data. The graph shows that the models of Martin, Faller & Merimaa, Lindemann, and Macpherson perform significantly better than the baseline and uninhibited models. Interestingly, the uninhibited model performs significantly better than the baseline model.

The performance of the models from the ANOVA, broken down by room, is plotted in Figure 6.11. As before, the figure includes 95% confidence intervals from the ANOVA. The figure reflects the observations from Figure 6.10, since the models of Martin, Faller & Merimaa, and Lindemann all appear to perform comparably in all of the rooms. The baseline model performs significantly worse than these models in rooms A–C. Macpherson's model performs comparably with these models in rooms X–C, but performs significantly worse in room D.

An important result is that whilst the uninhibited model performs comparably to most models in room X, the performance drops rapidly and is significantly worse than many of the models for rooms B–D. The baseline precedence model appears to provide no

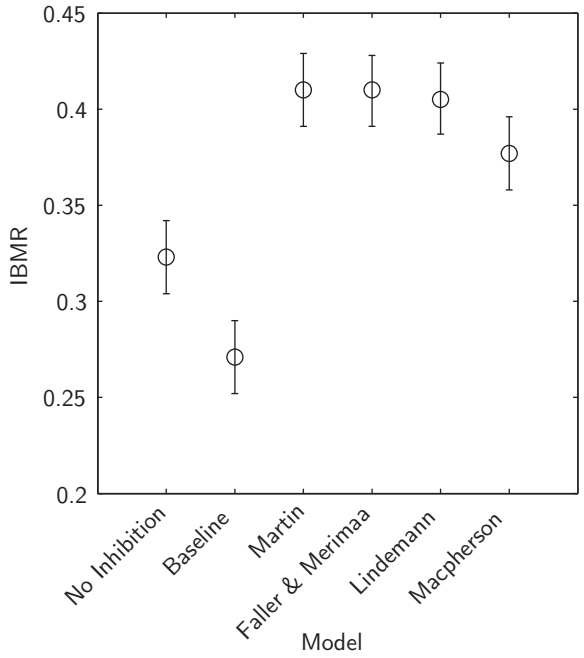


Figure 6.10: Mean performance of the precedence models from the ANOVA, with 95% confidence intervals.

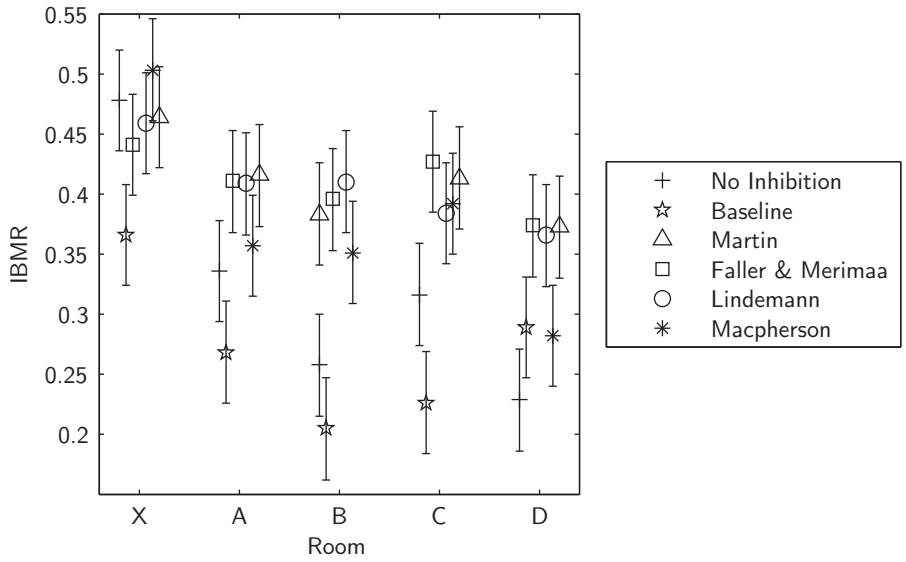


Figure 6.11: Mean performance of the precedence models from the ANOVA, broken down by room, with 95% confidence intervals.

performance gain until more reverberant conditions, and actually performs significantly worse than the uninhibited model in room X. This may be because, in less reverberant conditions, the baseline model is excessively removing information that would otherwise positively contribute to localisation. This suggests that the baseline model could be adapted in each room in order to improve performance. For example, setting $G = 0$ in room X would improve the performance to match the uninhibited model. Potentially, G could be increased as the reverberation increases. Setting $G = 1$ appears to have offered some improvement for room D.

6.5 Summary and Conclusions

This chapter aimed to answer the question: *7. Which approaches work best and are there any lessons to be learned for future development?* To investigate this, numerous computational precedence models were implemented and incorporated into a baseline separation algorithm. Of the models tested, the results show that the precedence models proposed by Martin (1997), Faller & Merimaa (2004), and Lindemann (1986a) work best and are a significant improvement on the baseline precedence model. Martin's model calculated an inhibitory signal based on onset data and multiplied this with the running cross-correlation. Faller & Merimaa's model calculated Interaural Coherence (IC) from the running normalised cross-correlation and used an IC threshold to specify cue selection. Lindemann's model is an extension of Jeffress' (1948) original cross-correlation theory of sound localisation. The model is extended with monaural detectors and a contralateral-inhibition mechanism.

These results indicate that a pyschoacoustic engineering approach has improved the reverberation–performance of a source separation algorithm, which partly answers the main research question for this thesis. However, it was also observed that a dynamic component may be necessary in order to optimise the performance of the precedence model. It was noted earlier that Faller & Merimaa (2004) state that setting the IC threshold in their model is a trade-off between selecting reliable cues that correspond closely to free-field conditions and maximising the proportion of the input signals that contributes to localisation. The results shown in this chapter reflect this and indicate that a dynamic component of the precedence models may be necessary in order to adapt the precedence processing to the acoustic conditions, thus maximising the separation performance of the algorithm. The following chapter will discuss dynamic computational precedence.

Chapter

7 Room–Specific Computational Precedence

In the previous chapter, an experiment was conducted that compared the separation performance achieved by incorporating numerous computational precedence models into a separation algorithm, against a baseline precedence model. These models were used to estimate the azimuths of two sound sources in order to separate the signals arising from each source. It was noted in the conclusions that the uninhibited baseline model performed well in the anechoic condition, but performed less favourably at higher RT_{60}s. Conversely, the baseline precedence model performed poorly at low RT_{60}s but performed well at higher RT_{60}s. Similarly, many of the precedence models appeared to perform less favourably in the anechoic condition but performed well in reverberant conditions. This suggests that in order to optimise the performance of the separation algorithm, the precedence model may need to adapt its processing to the acoustic conditions under which it is deployed. For example, the precedence model may need to be disengaged under anechoic conditions and the amount of inhibition increased as the acoustic conditions deteriorate. Dynamic processes in the precedence effect, and in particular the Clifton effect, have been observed in the psychoacoustic literature for many years. Hence, implementing such a feature retains the perceptual validity of the model. This chapter therefore aims to answer the research questions:

8. Can performance be further improved?

9. Are the results generalisable?

From the experiment detailed in Chapter 6, it is possible to directly compare the performance of the baseline algorithm with and without the precedence model (recall that the inhibitory gain was $G = 1$ for the former and $G = 0$ for the latter; see Equation 6.17, page 97). The logical follow-up to the previous experiment is therefore to test the baseline model with a range of values of G, in each of the rooms, in order to determine whether an optimal value exists for each room. In addition, it may also be possible to optimise the inhibitory time constant α_p, which affects the point at which the inhibition starts relative to the onset. Specifically, each room has a different early reflection pattern (highlighted by the range of Initial Time Delay Gap (ITDG) and C_{50} values). Although early reflections can be beneficial to human perception, they are still likely to provide contradictory localisation cues because of their different direction of

arrival and potential alteration of the spectral content. Hence it may be necessary to start inhibition before the first reflection, whilst keeping the value as high as possible in order to maximise the amount of signal that contributes to localisation.

Dynamic processes in the perceptual precedence effect will be discussed in Section 7.1. In order to address the first research question given above, Section 7.2 will detail an experiment that investigates a room–specific component of the baseline model. In response to the second question, Section 7.3 will investigate whether this room–specific component can be realised in the other precedence models detailed in Chapter 6. The results of the experiments will be compared and discussed in Section 7.4. Lastly, answers to the research questions will be concluded in Section 7.5.

7.1 Dynamic Procceses in the Precedence Effect

Dynamic processes in the precedence effect have been observed for many years. One effect that has been observed is the apparent 'build-up' up of the precedence effect (Thurlow & Parks 1961; Clifton & Freyman 1989; Freyman et al. 1991). This effect is observed when a listener is presented with a train of identical lead–lag clicks: the echo threshold is seen to raise by several milliseconds over the course of the train. Interestingly, the build-up has a finite duration that is related to the train rate—the echo threshold will reach a maximum after about 12 click pairs. Furthermore, the build-up is also affected by other stimulus parameters, including the number of lagging clicks (Yost & Guzman 1996) and whether the lead stimulus is presented from the left or the right (Clifton & Freyman 1989; Grantham 1996).

Another effect that has been observed is the 'break-down' of the precedence effect (Clifton 1987). Clifton reported that the precedence effect appeared to break-down when the stimuli were spatially reconfigured. Specifically, Clifton presented click trains to the listeners through two spatially–separate loudspeakers; the clicks were delayed by a few milliseconds in one loudspeaker and presented at a rate of about one per second. Under these conditions the listener always localised the sound as originating from the lead loudspeaker, as expected, due to the precedence effect. However, when the lead and lag loudspeakers were swapped, the listeners could temporarily localise both clicks separately until echo suppression re-engaged and localisation moved to the new lead loudspeaker. In a similar manner to the build-up effect, the duration over which listeners were able to localise both clicks was determined by the click rate, i.e. it took a fixed number of clicks (8–12) to re-engage echo suppression. These findings were subsequently confirmed by others (e.g. Freyman et al. 1991; Blauert & Col 1992; Blauert 1997) with a wider range of parameters, including loudspeaker quantity and stimulus type (e.g. noise bursts and band-pass-filtered clicks and noise bursts). This apparent breakdown of the precedence effect became known as the "Clifton effect".

Since these experiments, some authors have found that the precedence effect does not truly break down (e.g. Djelani & Blauert 2001a,b), but is actually temporarily stored. Specifically, consider the above situation of swapping the lead–lag stimulus. During the initial click train the precedence effect is seen to build up. When the lead and lag are swapped, listeners are temporarily able to localise both clicks until the precedence effect builds up again. Now, if the lead and lag are then swapped back to their original configuration, the initial built-up precedence effect will be reinstated. This original precedence effect can be remembered for about nine seconds. Beyond this time, a new build-up will occur.

Blauert (1997) concludes that the Clifton effect occurs when an "implausible" reflection pattern is heard. This may include, for example, an ITD that exceeds the maximum possible ITD in free-field conditions, implying that the time of arrival difference is due to a reflection. This implausible reflection pattern causes the precedence effect to breakdown whilst the listener rescans the room. This breakdown affects echo suppression, localisation, externalisation and fusion. The precedence effect then builds-up again in response to the new reflection pattern.

These observations indicate that the precedence effect is able to adapt to the acoustic conditions in which the listener is situated. This is an intuitive result; in a free-field, all auditory cues are important because none have been altered by reflections arriving from room boundaries or surfaces. Conversely, in any real room, the signal arriving at the ear is a summation of the direct signal and room reflections. The nature of these reflections is likely to vary dramatically according to the room dimensions, the absorption coefficients of the boundaries, the relative positions of source and listener and so forth. It is likely that a computational model of precedence will also need to factor in the room acoustics in this way, in order to maximise the effectiveness of the cue selection. Therefore, it would be useful to investigate whether a room–specific component of the precedence model is necessary and whether it can offer an improvement in the separation performance of the algorithm.

Unfortunately, like the precedence effect, there is little data on the neurophysiological mechanisms that are responsible for these dynamic processes, although Blauert (1997) and Litovsky et al. (1999a) agree that the effect is at least partially achieved by feedback from higher auditory systems to the peripheral auditory system through the centrifugal pathways. Consequently, the work discussed in this chapter does not intend on accurately modelling the mechanisms that achieve these dynamic processes, but is a first step towards implementing a model that represents a functional equivalent. A dynamic system would need to estimate optimal precedence parameters based on the input signal. This would require a significant amount of work (which was not possible within the time-scale of this project) and is discussed further in Section 8.3 (page 136). Instead, this chapter takes the first step by formally investigating whether

this dynamic component is present and if it offers any performance improvement. Since the algorithms tested in this chapter are not technically 'dynamic', they are called 'room–specific' computational precedence models.

7.2 Optimising the Baseline Precedence Model

8. Can performance be further improved?

As discussed in the introduction of this chapter, the results obtained in Chapter 6 suggest that it may be possible to improve the performance of the separation algorithm by optimising the inhibitory parameters of the precedence model: the inhibitory gain G and the inhibitory time constant α_p. A shown in the previous section, such a dynamic component of the precedence effect has been observed in the psychoacoustic literature for many years. Hence, it is likely that the optimal parameters will be room–dependent because each room has different acoustic conditions (e.g. different ITDGs, DRRs and RT_{60}s). As previously discussed, choosing the precedence model parameters is a trade-off between selecting reliable cues and maximising the amount of signal that contributes to localisation, and this trade-off is likely to be dependent upon the acoustics of the room. Therefore, this section details an investigation that will show whether a room–specific component in the baseline precedence model can improve separation performance and also show the extent of performance improvement offered by such a component. The experimental procedure is detailed in the following section and the results are presented and discussed in Section 7.2.2.

7.2.1 Experimental Procedure

The experimental procedure employed in this investigation was identical to that described in Section 6.3 (page 106) and Chapter 5 (page 74). To summarise, the model was tested with the following mixture conditions:

- Target/interferer azimuthal separations of 10°, 20° and 40° (i.e. ±5°, ±10° and ±20° with respect to the frontal median plane), with the target on the left
- Target-to-Interferer Ratios (TIRs) of 0, 10 and 20 dB (RMS)
- The following interfering signals: white noise, male speech and a modern piece of rock music
- RT_{60}s of 0, 0.32, 0.47, 0.68 and 0.89 seconds

In addition to these mixture parameters, a range of values for the inhibitory gain G and inhibitory time constant α_p were also tested. With $\alpha_p = 0$ or $G = 0$, no inhibition

will be triggered and the algorithm will simply cross-correlate the input. The time regions of the input signal that will be inhibited will be affected by varying α_p (i.e. how soon the inhibition starts after an onset); the strength of inhibition increases with G. Setting these values is a trade-off between selecting reliable regions of the input signal that exhibit minimal corruption by reverberation and maximising the proportion of the input signal that contributes to localisation. Specifically, the input could be highly inhibited with a small value of α_p and high G; this would yield a signal that is highly uncorrupted by reverberation, but bears little or no resemblance to the input and thus the separation result will be highly inaccurate. Additionally, increasing G will increase the likelihood of cross-correlation values dropping below the grouping threshold Θ_c. This is because increasing G increases the inhibition and reduces the value of the precedence–modelled fine structure, resulting in low values being output by the cross-correlation. This will result in the corresponding T–F unit being excluded at the output. A broad range of α_p values was used such that α_p = [0,25] ms; this range extends beyond the default value of 15 ms (specified in Chapter 6) and encompasses many of the precedence threshold values given in Table 4.1 (page 58). The range of G values used was G = [0,2], i.e. up to double the default value. The following tests were then performed on the model for all other mixture variables:

- Values of the inhibitory gain G were tested, with α_p fixed at its default value of 15 ms.
- Values of the inhibitory time constant α_p were tested, with G fixed at its default value of 1.
- Both parameters were optimised by testing for the optimal value of G, given the optimal value of α_p.

A summary of these values is presented in Table 7.1 (other models are listed in the table for use in Section 7.3). The results presented below will compare the 'static' precedence model (i.e. the baseline model as presented in Chapter 6, with the precedence parameters fixed to their default values) with the performance of the model achieved by independently optimising the inhibitory gain, the inhibitory time constant and both parameters together.

7.2.2 Results and Discussion

The results for this experiment are given in Figure 7.1. From these plots there are several important observations to make:

- Optimising the precedence parameters has resulted in a large increase in separation performance.

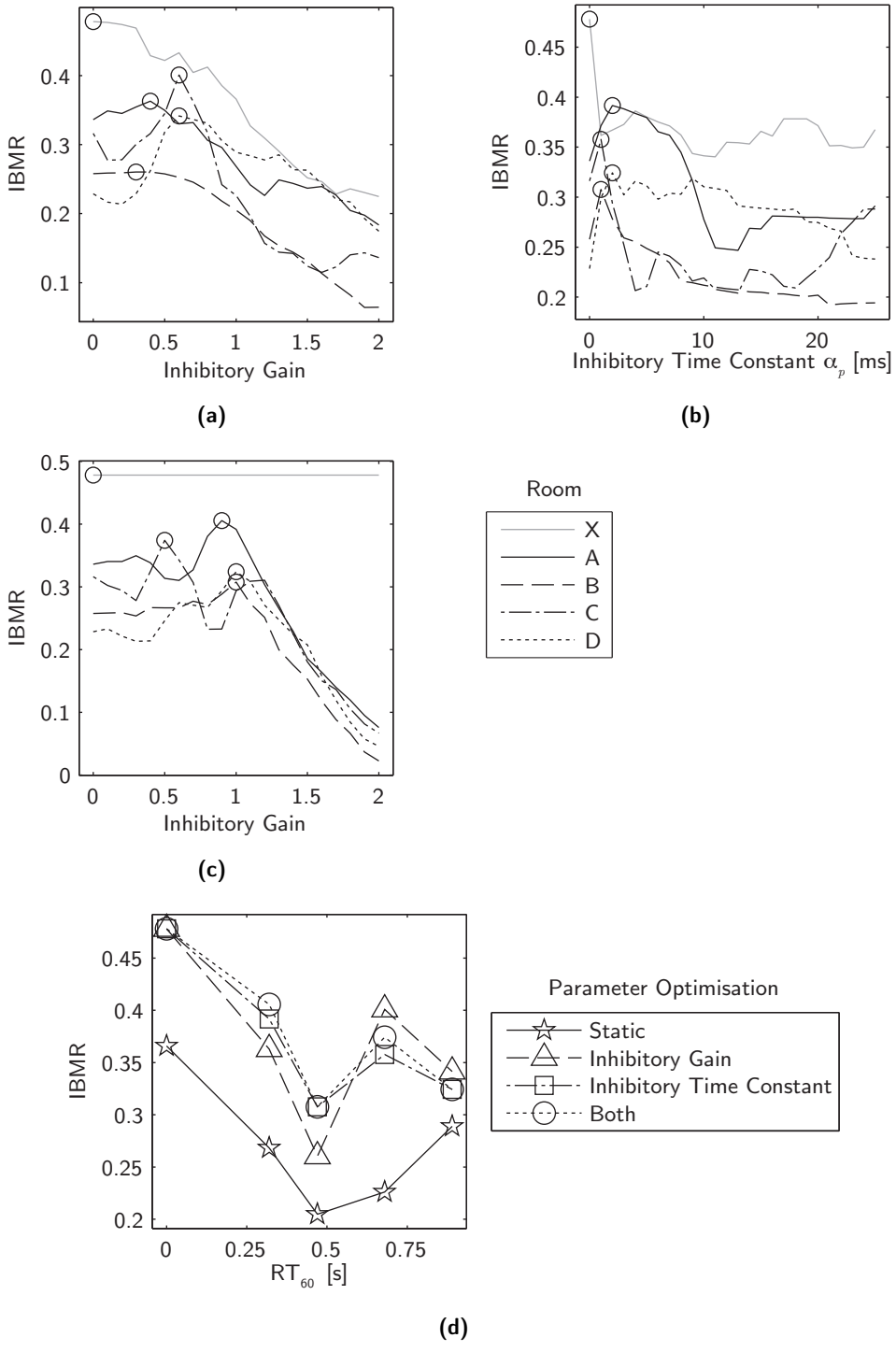


Figure 7.1: Optimising the baseline model. The highest point in plots (a), (b) and (c) are identified with a circle; this indicates the optimal parameter. **(a)** Optimising the inhibitory gain G. **(b)** Optimising the inhibitory time constant α_p. **(c)** Optimising G, given the optimal α_p. **(d)** Model performance given the optimal parameter values (obtained from the other plots) versus the 'static' case presented in Chapter 6.

Table 7.1: Precedence model parameters for each of the models under test. The 'default value' indicates the value assigned to the parameter in Chapter 6 and the value that the parameter is held at whilst the other parameter is varied; the 'range' indicates the range of values tested in order to optimise the parameter.

Precedence Model	Parameters	Default Value	Range	
			Min	Max
Baseline	Inhibitory Gain G	1.0	0.0	2.0
	Inhibitory Time Constant α_p [ms] (in samples)	15.0	0.0	25.0
Martin	Inhibitory Gain G	1.0	0.0	2.0
	Inhibitory Time Constant α_m [ms] (in samples)	1.5	0.0	25.0
Faller & Merimaa	IC Threshold Θ_χ	0.5	0.0	0.99
	Exponential Window Time Constant α_f [ms] (in samples)	10.0	0.0	25.0
Lindemann	Inhibition Parameter c_{inh}	0.3	0.05	1.0
	Fade-off Time Constant α_{inh} [ms] (in samples)	10.0	0.0	25.0
Macpherson	Inhibitory Gain G	1.0	0.0	2.0
	Inhibitory Time Constant α_m [ms] (in samples)	1.5	0.0	25.0

- The optimal inhibitory gain G and optimal inhibitory time constant α_p is different in every room.
- For the anechoic condition (Room X), optimal performance is obtained by setting $G = 0$ and/or $\alpha_p = 0$.
- Optimal values of α_p are generally small and much smaller than the default value of 15 ms.
- The optimal gain values appear to depend upon the inhibitory time constant, i.e. there is an interaction between the parameters. This is demonstrated by two points: firstly, the optimal values of G are different in Figures 7.1(a) and 7.1(c). Secondly, according to Figure 7.1(d), with both parameters optimised, optimising only the inhibitory time constant provides better performance at high RT_{60}s. This interaction is discussed further below.

It was stated in the previous section that choosing values of the inhibitory gain G and inhibitory time constant α_p is likely to be a trade-off between selecting reliable cues corresponding to free-field conditions, and maximising the amount of signal that contributes to localisation. The results seemed to support this, since the optimal combination of values of G and α_p are room–dependent and hence for each acoustic condition there is a single parameter value that offers the best performance. Furthermore, this procedure has led to a large increase in separation performance.

In the anechoic (free-field) condition, optimum performance is achieved by bypassing the precedence model by setting $G = \alpha_p = 0$. From this, it could be hypothesised that the optimal precedence parameter values are correlated to a corresponding acoustic parameter. For example, the value of the inhibitory time constant that achieves maximum performance could be related to the ITDG of the room in order for inhibition to start before the first reflection. As discussed earlier in the chapter, each room has a different early reflection pattern and these reflections are likely to provide contradictory localisation cues. Hence it may be necessary to start inhibition before these reflections, whilst keeping the value as high as possible in order to maximise the amount of signal that contributes to localisation. Similarly, the value of the inhibitory gain that achieves maximum performance could be related to the DRR of the room such that the strength of inhibition is related to the amount of reverberation. The model may need to maximise inhibition in order to suppress unreliable cues, whilst maximising the amount of signal that contributes to localisation. Quantifying these correlations is beyond the scope of this thesis, although some validation of this theory could be achieved if this effect is observed in other precedence models.

This hypothesis may be able to partially explain the interaction between the precedence parameters. The hypothesis suggests that the optimal precedence parameter values are correlated to a corresponding acoustical parameter. This interaction could be explained if the optimal precedence parameter values are in fact correlated to two (or more) acoustical parameters. However, quantifying these interactions would require a detailed statistical analysis and a controlled stimulus set in which individual acoustical parameters could be independently controlled; this is beyond the scope of this thesis.

7.3 Optimising other Precedence Models

9. Are the results generalisable?

The results presented in the previous section demonstrated that the baseline model can be dynamically optimised and that the optimal inhibitory gain and inhibitory time constant values are dependent upon the room under test. Furthermore, the results indicated that this procedure offered an improvement in separation performance. However, it is unclear whether these results are coincidental for this particular precedence model, or whether they reflect a wider necessity for this room–specific component amongst other computational precedence models. In order to answer this question, the other models implemented in the previous chapter were tested with a range of precedence parameter values. This investigation is detailed in this section.

In order to answer the above question, an identical investigation to that detailed in Section 7.2 was conducted. However, due to the differing operations and precedence parameters of the models, different parameter values were used. These values are also

detailed in Table 7.1.

7.3.1 Optimising Martin's model

Martin's model was tested in an identical manner to the baseline model, since the precedence parameters are identical. Note that since Martin (1997) suggests a default inhibitory time constant α_p of 1.5 ms, rather than the 15 ms of the baseline model, this was the held value used whilst the inhibitory gain was tested.

The results of the experiment are given in Figure 7.2. From these plots there are several important observations to make:

- Optimising the precedence model parameters has had a small effect on the separation performance and hence there is only a small increase in overall separation performance.
- Contrary to the baseline model, optimal performance in the anechoic condition is not achieved with small or zero values of G and/or α_m.
- As with the baseline model, there appears to be an interaction between the precedence model parameters.

As with the baseline model, the optimal precedence parameter values appear to be room–dependent. This supports the suggestion from the baseline model results that there is a requirement for a mechanism that adapts the precedence model to the specific acoustic characteristics of the room under test. Furthermore, like the baseline model, there appears to be an interaction between the precedence model parameters; this requires further quantification. However, there are some dissimilarities compared to the baseline results. Firstly, in the anechoic condition, the best performance is not achieved by setting G and/or α_m to zero. This undermines the aforementioned assertion that the precedence model should be bypassed in anechoic conditions. Secondly, the performance improvement achieved by optimising the precedence model parameters appears to be relatively small.

It is interesting to note the dissimilarities in performance between the baseline model and Martin's model, given the conceptual similarities. However, there are a number of differences between the models. The primary difference is that the baseline model applies inhibition before cross-correlation, whereas Martin's model applies the inhibition after cross-correlation. Despite this, it is difficult to draw specific conclusions about the relative merits of pre- versus post-cross-correlation inhibition. This is because of the other differences between the models, including: the cross-correlation algorithm and the procedure employed to calculate the inhibitory signal.

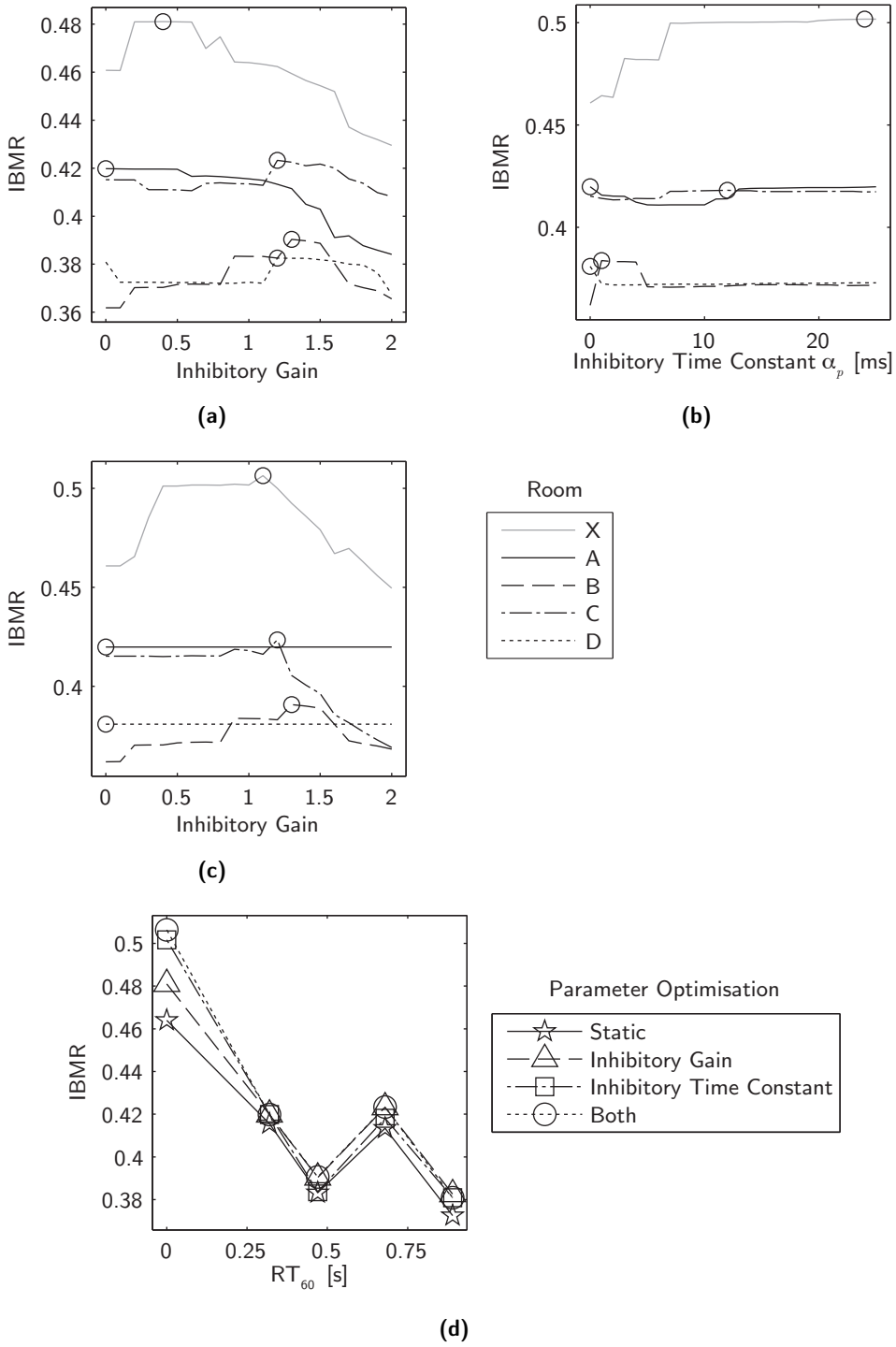


Figure 7.2: Optimising Martin's model. The highest point in plots (a), (b) and (c) are identified with a circle; this indicates the optimal parameter. **(a)** Optimising the inhibitory gain G. **(b)** Optimising the inhibitory time constant α_p. **(c)** Optimising G, given the optimal α_p. **(d)** Model performance given the optimal parameter values (obtained from the other plots) versus the 'static' case presented in Chapter 6.

7.3.2 Optimising Faller & Merimaa's model

Faller & Merimaa's model was tested by varying the IC threshold Θ_χ and the exponential window time constant α_f. The IC threshold determines the time–regions of the input signal that contribute to localisation. It is likely that in the anechoic condition, the IC will generally be higher because all of the cues are reliable; the threshold can therefore be set to a low value (0) in order to bypass cue selection. In more reverberant conditions, where the IC will show a greater degree of variation, optimal performance may be achieved by being more selective with cues and thus the IC threshold should be set higher. Values of the IC threshold were chosen such that $\Theta_\chi = [0, 0.99]$.

The exponential windowing of the model indirectly provides a form of inhibition. For example, a strong peak in the cross-correlation will mask any less coherent cross-correlations that follow. Increasing the time constant will increase this masking effect. However, the converse is also true: strong peaks in the cross-correlation may also be masked by long–running incoherent cross-correlations. Values of the exponential window time constant were chosen to be identical to the inhibitory time constant values chosen for the baseline model and Martin's model such that $\alpha_f = [0, 25]$ ms (in samples).

The results of the experiment are given in Figure 7.3. From these plots there are several important observations to make:

- The IC threshold Θ_χ can be set to a wide range of values in any room, it only appears to be important that it does not exceed approximately 0.6 in any room.
- Optimal performance in the anechoic condition is achieved by setting the IC threshold Θ_χ and exponential window time constant α_f to zero.
- Choosing the exponential window time constant α_f appears to be more important than choosing the IC, especially in the anechoic condition. The optimal value of α_f is different in each room, further supporting the necessity for an adaptive mechanism identified in the previous models.
- There appears to be a small interaction between the precedence parameters, exemplified by the anechoic data. However, this appears to have had a minimal effect on the overall results.
- Optimising the precedence parameter values has resulted in a larger improvement at low RT_{60}s and a much smaller effect at higher RT_{60}s.

As with previous models, the optimal precedence parameter values appear to be room–dependent, further supporting the necessity for a mechanism that adapts the precedence

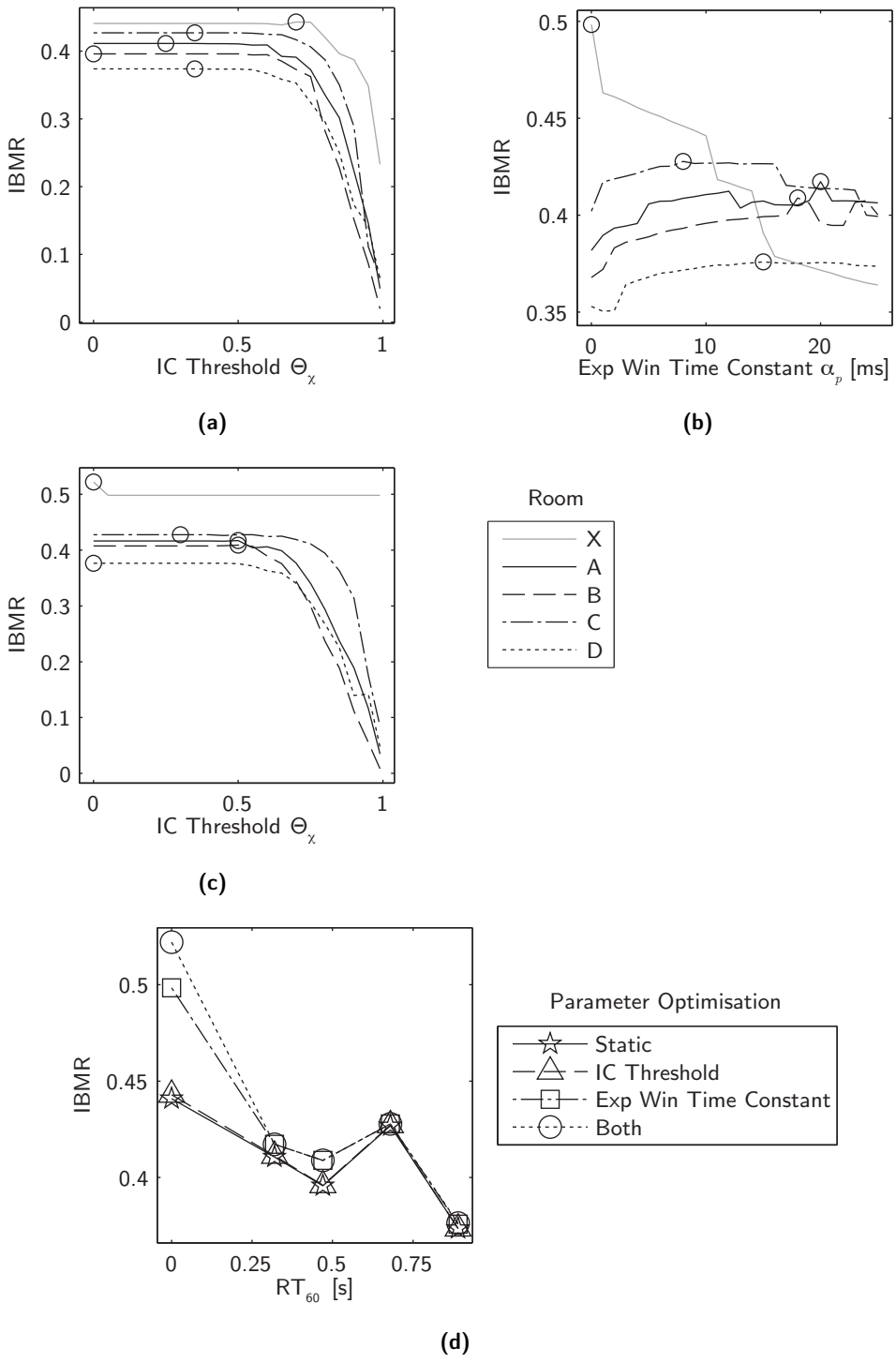


Figure 7.3: Optimising Faller & Merimaa's model. The highest point in plots (a), (b) and (c) are identified with a circle; this indicates the optimal parameter. **(a)** Optimising the IC threshold Θ_χ. **(b)** Optimising the exponential window time constant α_f. **(c)** Optimising Θ_χ, given the optimal α_f. **(d)** Model performance given the optimal parameter values (obtained from the other plots) versus the 'static' case presented in Chapter 6.

model to the acoustic characteristics of the room under test. However, the relationship between the IC threshold and IBMR raises some interesting points. In the anechoic condition, where the optimal exponential window time constant $\alpha_f = 0$ ms, there is a clear interaction with the IC threshold Θ_χ. In this case, the performance no longer reduces for high values of Θ_χ, but is higher for $\Theta_\chi = 0$. For other rooms, where the optimal $\alpha_f \neq 0$ ms, the performance is consistent for $\Theta_\chi \approx$ [0,0.6], but drops off rapidly for $\Theta_\chi > 0.6$. This finding appears to disagree with the assertion made by (Faller & Merimaa 2004) that the cue selection threshold should adapt to "each specific listening scenario", since the data suggests that a wide range of values provide optimal performance in all of the rooms. However, this experiment does not take into account varying numbers of sound sources, which may have a large impact on the optimal IC threshold. This is because IC is inversely proportional to the number of sound sources.

As with the baseline model, and contrary to Martin's model, optimal performance in the anechoic condition is achieved by setting $\Theta_\chi = \alpha_f = 0$. These values achieve a large performance gain in the anechoic condition; this is the largest gain across any of the rooms. The performance gain at the highest RT_{60}s is negligible. Despite this, the plots demonstrate that it is still necessary to choose the values of Θ_χ and α_f carefully in order to not impede the performance of the precedence model.

7.3.3 Optimising Lindemann's model

Lindemann's model was tested by varying the inhibition parameter c_{inh} and the fade-off time constant α_{inh}. The inhibition parameter determines the strength of inhibition. It is likely that in the anechoic condition, less inhibition—as determined by a lower value of c_{inh}—will provide the optimal performance; the reverse is likely to be true in reverberant conditions. In accordance with Lindemann's (1986a) model, values of the inhibition parameter were chosen such that $c_{\text{inh}} \in (0, 1]$. Note that the lowest value cannot be zero since this will mute the input.

The fade-off time constant α_{inh} of the model determines the time constant of the dynamic inhibitory component. Small values of α_{inh} will result in the inhibition being determined by the contralateral signal and only the most recent cross-correlations. Increasing α_{inh} will increase the duration of the dynamic inhibitory component and hence the duration over which the inhibition—caused by peaks in the cross-correlation—is applied. As with previous models, the fade-off time constant is chosen such that $\alpha_{\text{inh}} = [0, 25]$ ms (in samples).

The results of the experiment are given in Figure 7.4. From these plots there are several important observations to make:

- Optimising the precedence parameter values has resulted in a moderate and consistent increase in separation performance across all of the rooms.

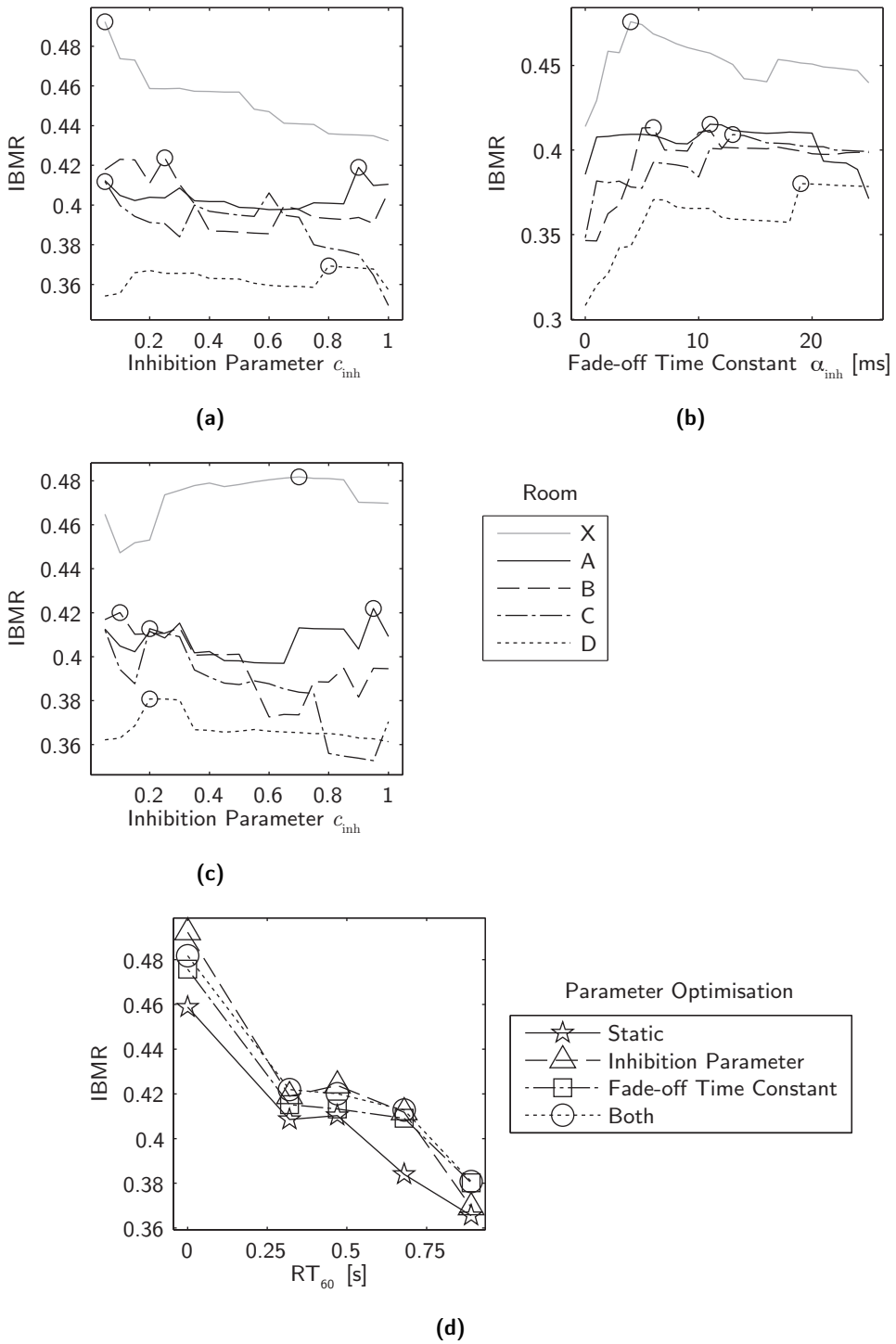


Figure 7.4: Optimising Lindemann's model. The highest point in plots (a), (b) and (c) are identified with a circle; this indicates the optimal parameter. **(a)** Optimising the inhibition parameter c_{inh}. **(b)** Optimising the fade-off time constant α_{inh}. **(c)** Optimising c_{inh}, given the optimal α_{inh}. **(d)** Model performance given the optimal parameter values (obtained from the other plots) versus the 'static' case presented in Chapter 6.

- There is an interaction between the precedence model parameters.
- There appears to be a minimum fade-off time constant that is necessary for effective separation.

The results show that the optimal parameter values vary across the different rooms, as with previous models. However, like Martin's model, optimal performance in the anechoic condition is not achieved by setting the inhibition parameter c_{inh} and/or fade-off time constant α_{inh} to zero. From Equation 6.35 (page 103), it can be seen that as $\alpha_{\text{inh}} \to 0$, then $\mathbf{\Phi}(i, n, \tau) \to \acute{\mathbf{c}}(i, n-1, \tau)$, where $\mathbf{\Phi}$ is the dynamic inhibitory component and $\acute{\mathbf{c}}$ is the running cross-correlation. It is likely that neglecting the other elements of the dynamic inhibitory component, especially the negative running cross-correlation term, is detrimental to the operation of the model, accounting for the poor performance at small values of α_{inh}.

7.3.4 Optimising Macpherson's model

Macpherson's model was tested in an identical manner to the baseline and Martin's model, with $\alpha_m = 1.5$ ms. This was possible because, despite the vastly different implementations, these models have identical parameters. However, it seems unlikely that varying these parameters will have a large effect on the performance of the model. This is because the precedence–based weighting procedure of the cross-correlations for individual peaks, relative to an initial peak, is performed for every peak. Therefore, the resulting effect may appear to be a sliding window, rather an active inhibitory process. Despite this, it is still possible that varying this window will impact upon the performance of the model and this possibility should be tested nonetheless.

The results of the experiment are given in Figure 7.5. From these plots there are several important observations to make:

- With the default inhibitory time constant α_m, varying the inhibitory gain G appears to have a minimal effect.
- With the optimised α_m, varying G appears to have a larger effect. As with previous models, this highlights an interaction between the parameters.
- Like previous models, the optimal precedence parameter values appear to be room–specific. However, like Martin's and Lindemann's models, optimal performance in the anechoic condition is not achieved by setting α_m and/or G to zero.
- Although the optimisation has resulted in only a small performance gain, it still appears to be important to choose the precedence parameter values carefully, since incorrect choice can be detrimental to the performance.

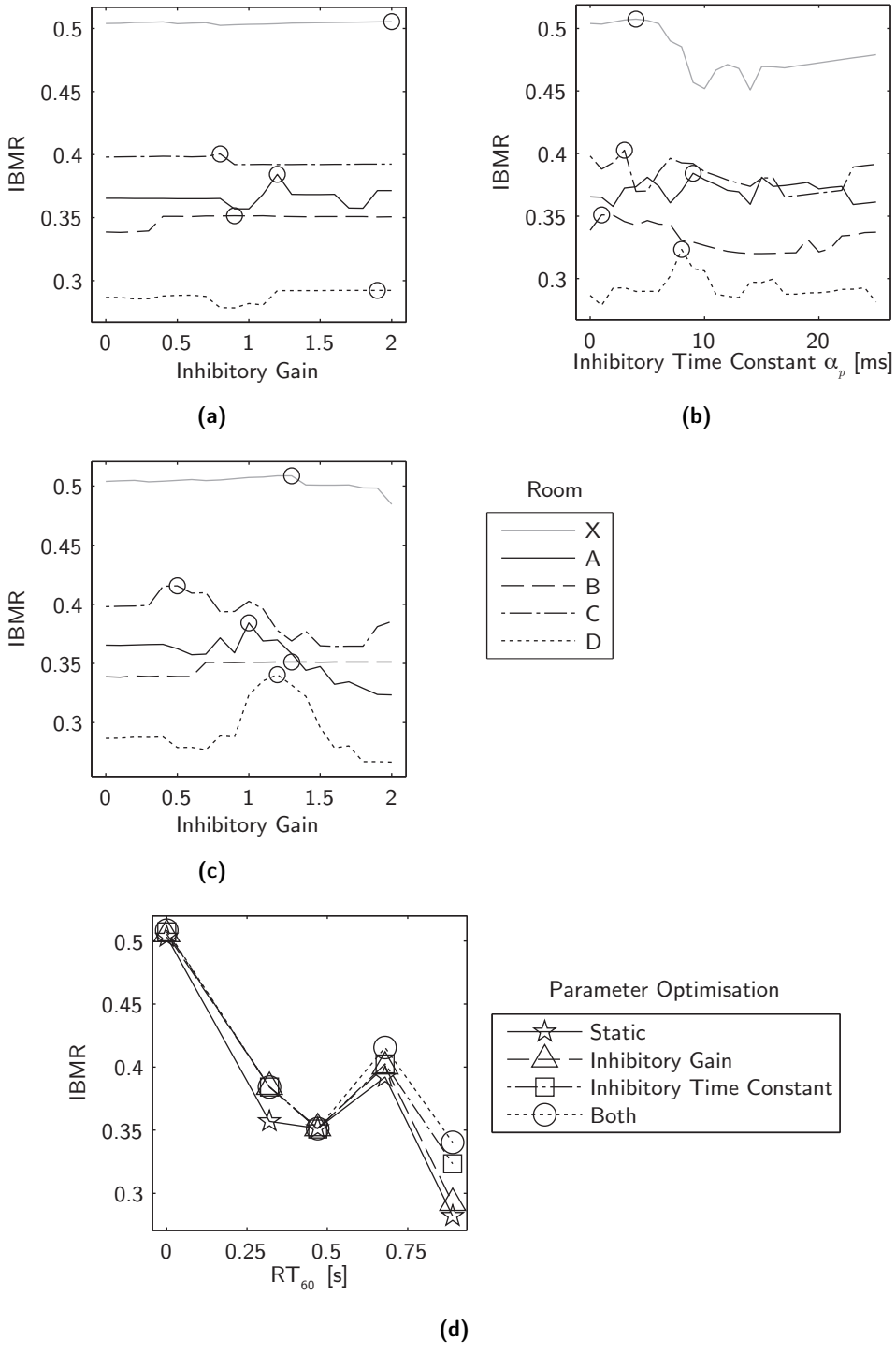


Figure 7.5: Optimising Macpherson's model. The highest point in plots (a), (b) and (c) are identified with a circle; this indicates the optimal parameter. **(a)** Optimising the inhibitory gain G. **(b)** Optimising the inhibitory time constant α_m. **(c)** Optimising G, given the optimal α_m. **(d)** Model performance given the optimal parameter values (obtained from the other plots) versus the 'static' case presented in Chapter 6.

These results seem to discount the assertion made above: that varying these parameters will have a minimal effect on the result. Although the variation in performance with each parameter is smaller compared to previous models, appropriate choice of parameter value can still lead to an IBMR gain of up to 0.05.

7.4 Results Comparison and Discussion

From the observations made in the previous sections, there are three that are of interest for comparative purposes:

- All models appear to benefit from individual adaptation of their parameters to the room under test. The optimal values of these parameters are unique to each room.
- For some models (baseline, Faller & Merimaa) it was shown that bypassing the precedence model, by setting its parameters to zero, provided the optimum performance in the anechoic condition. However, this was not true for all models (Martin, Lindemann, Macpherson).
- All models demonstrated an interaction between their precedence parameters.

It was stated in Section 7.2.2 that there is likely to be a correlation between each precedence parameter value that achieves optimal performance and a corresponding acoustic parameter. Although this has not been formally tested, it was suggested that it might be indicated if optimal performance in the anechoic condition is achieved by setting the precedence parameter values to zero. Interestingly, only the baseline model and Faller & Merimaa's model appear to support this specific hypothesis. However, although the other models do not demonstrate this specific effect, they do not discount the theory that there is a correlation between their precedence parameters and some acoustic parameter. This is because for every model and every room, there is a unique combination of precedence parameter values that achieve optimal performance. Therefore, there must be an interaction between the particular mechanisms of the precedence model and the acoustic features of each room.

The performance of each of the optimised models is compared in Figure 7.6. These results can be directly compared with the results given in the previous chapter for the static case (see Figure 6.9, page 107). There are several important observations that can be made by comparing these plots:

- Unlike the static models (with the exception of Macpherson's model), all optimised models out-perform the uninhibited model in all conditions.

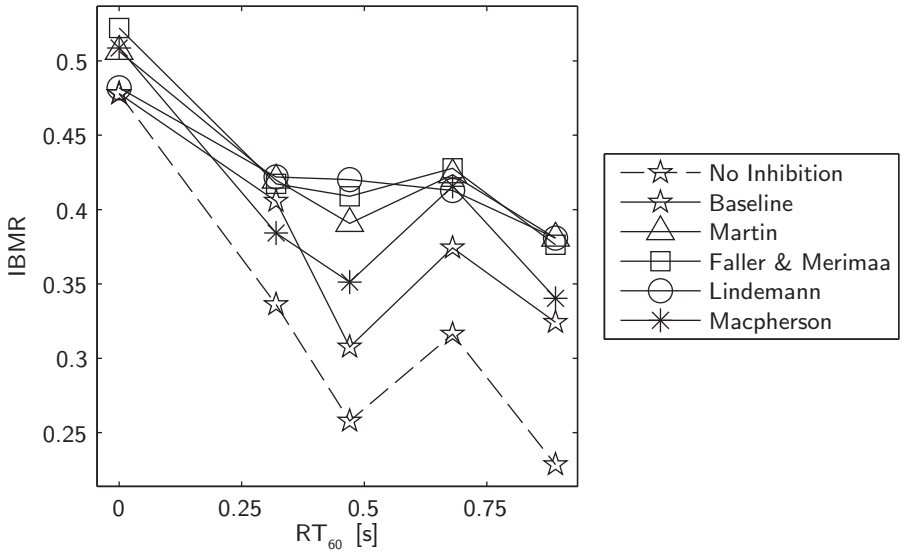


Figure 7.6: Mean optimised model performances.

- The performance of the optimised models in the anechoic condition is more similar than in the static case.
- Faller & Merimaa's model generally appears to perform best in the static and optimised cases.
- Optimising Lindemann's model has improved its performance relative to other models when compared to the static case; the relative performance of other models appears to have remained consistent.

In order to quantify these observations, the performance gain achieved by optimising the precedence model was calculated and is plotted in Figure 7.7. The plot shows that the optimisation procedure has provided the most improvement for the baseline model. For Martin's and Faller & Merimaa's models, the largest gain is in the anechoic condition. Conversely, for Lindemann's and Macpherson's models, the largest gain is at higher RT_{60}s. However, there are a large number of model/room combinations that achieve a very small or null performance gain.

It is perhaps most interesting that the baseline model demonstrates the most performance gain as a result of optimising its parameters. This may be because it is the only precedence model that processes the fine structure before cross-correlation. For example, consider a sample point n that is uncorrupted by reverberation, but a later point $n + d$ is highly corrupted by reverberation. In many of the precedence models the cross-correlogram is calculated from a series of cross-correlations, which in turn are calculated from windowed portions of the IHC data. If d is sufficiently small, the cross-correlation for point n will include point $n + d$, reducing the reliability

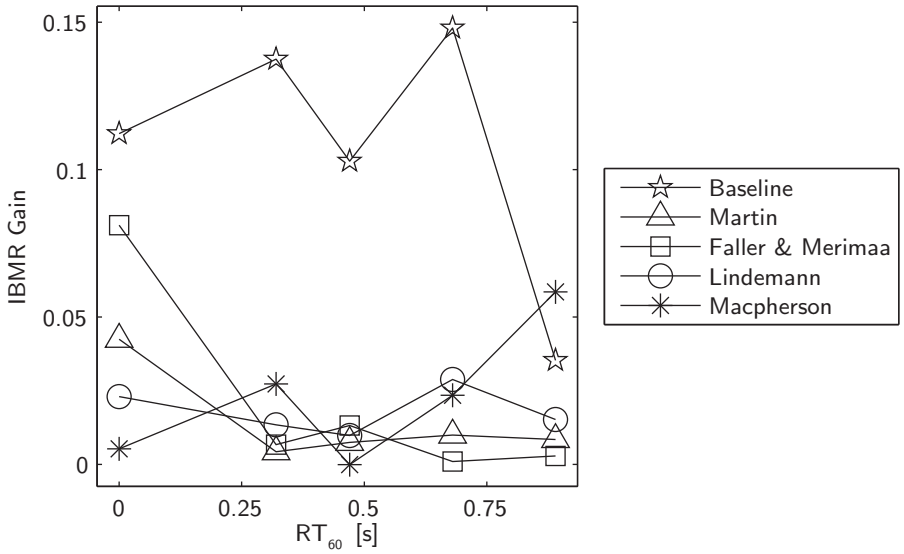


Figure 7.7: Performance gains arising from the model optimisations.

of the result. If the precedence model has detected that point $n + d$ is corrupted then the cross-correlation for $n + d$ will be excluded, but the cross-correlation for point n, which integrated $n + d$, will still be included. This reduces the reliability of the cross-correlogram. However, in the case of the baseline model, $n + d$ will be excluded at the input to the cross-correlation and will thus never be integrated into the cross-correlogram. Because this effectively broadens the time regions that contribute to localisation, with precedence processing engaged, the performance in the anechoic condition appears to be better than expected, whereas performance in more reverberant conditions is lower than expected. This is partially reflected in the results from Chapter 6, where all except the baseline precedence model perform comparably to the uninhibited model in the anechoic condition. Furthermore, it suggests that greater performance may be achievable at higher RT_{60}s.

If the above assertion—that pre-cross-correlation processing is more effective that post-cross-correlation processing—is true, why then does the baseline model perform poorest of all models? This may simply be due to the cross-correlation algorithm, and perhaps the IHC model. Consider Faller & Merimaa's model (which has a different IHC model to the baseline model): bypassing the precedence model (i.e. with $\Theta_\chi = \alpha_f = 0$)[1] still results in higher performance than the optimised baseline model for all rooms except Room A. From this it is clear that the choice of cross-correlation algorithm is just as important as the precedence model. This includes parameters such as the window length, window shape, and whether the cross-correlation is normalised.

[1]This is estimated from Figure 7.3(b) ($\alpha_f = 0$) given that, from Figure 7.3(a), the performance for $\Theta_\chi = 0.5$ is approximately equal to the performance for $\Theta_\chi = 0$.

Some work, not presented in this thesis, was conducted that attempted to relate the optimal precedence parameters to acoustical parameters of the room (Hummersone et al. 2010). This work considered only the baseline model and found that there was some correlation between the inhibitory gain G and DRR, and between the inhibitory time constant α_p and ITDG. However, further work failed to show similar correlations in the other precedence models, suggesting that further work is necessary in order to identify the reasons for these differences.

7.5 Summary and Conclusions

This chapter aimed to answer the following research questions:

8. Can performance be further improved?
9. Are the results generalisable?

8. Can performance be further improved?

This chapter has investigated whether it was possible to improve the separation performance achieved by the precedence–model–enhanced separation algorithms presented in Chapter 6. Specifically, Chapter 6 hypothesised a room–specific component of the precedence models, whereby inhibitory parameters of models could be adapted in each room in order to optimise the performance in the room. The subsequent investigation tested the baseline model with a range of mixture and inhibitory parameters in each room. The results showed that the performance achieved using the baseline model could be further improved. Furthermore, this improvement was shown to be achieved by the hypothesised room–specific component. This component is analogous to the perceptual Clifton effect, which appears to adapt the precedence effect to the acoustic environment in which the listener is located.

9. Are the results generalisable?

In order to test whether this room–specific component was specific to the baseline model or present in other precedence models, the investigation was repeated with the other precedence models. These investigations confirmed that such a room–specific component was present and offered further performance improvement. However, the overall performance improvement was shown to be less than that for the baseline model. It can be concluded from these results that, at least for this source separation algorithm, a model of the Clifton effect is a necessary part of a computational precedence model. The results also show that the choice of cross-correlation algorithm and cross-correlation parameters such as window length and shape are important considerations and must be chosen carefully.

Chapter

8 Summary and Conclusions

This chapter aims to answer the main research question given in Section 1.6. In order to answer the research question, several sub-questions were formulated that have been answered throughout this thesis. A summary of the thesis and the answers to the research sub-questions are given in Section 8.1; the main research question is answered at the end of the section. The original contributions of this thesis are given in Section 8.2. Future work arising from the research described in this thesis is given in Section 8.3.

8.1 Thesis Summary and Answers to Research Questions

This section will summarise each chapter of the thesis and answer the research sub-questions that each chapter addressed. The main research question will then be addressed.

8.1.1 Chapter 2: Auditory Scene Analysis

Before considering psychoacoustic engineering approaches to machine source separation in reverberant environments, it was necessary to have an understanding of the human physiological, perceptual and psychoacoustical mechanisms that accomplish this task. Therefore, Chapter 2 aimed to establish the mechanisms behind Auditory Scene Analysis (ASA): the theory that describes psychoacoustic sound source segregation. The chapter firstly dealt with the physiological mechanisms of the peripheral auditory system, since this processing is crucial to how ASA is performed. Specifically, three important observations were made. Firstly, the outer and middle ear provide directional filtering and match the impedance of the air to the impedance of the inner ear. Secondly, the inner ear filters the sound into numerous frequency bands by way of the cochlea and basilar membrane. Thirdly, the auditory nerve exhibits numerous interesting properties such as frequency selectivity, a kind of noise floor, non-linear compression and adaptation to steady stimuli. Following this, the mechanisms of ASA were presented. ASA has two stages. The first stage is segmentation, which is the process that breaks the sounds arriving at the ear into local time–frequency region. In the second stage, grouping, these segments are recombined into streams that represent each sound source.

This grouping takes place both simultaneously and sequentially in time. Furthermore, this grouping can either take place using primitive mechanisms that are innate or using learned schemas, such as those that are used to group components of speech.

8.1.2 Chapter 3: Computational Auditory Scene Analysis

Before considering psychoacoustic engineering approaches to machine source separation in reverberant environments, it was also necessary to have an understanding of current psychoacoustic engineering approaches to machine source separation. Therefore, Chapter 3 aimed to establish which techniques are commonly used for CASA and how they are implemented computationally.

A typical CASA system architecture contains the following stages: peripheral analysis, feature extraction, mid-level representations, scene organisation and resynthesis.

For *peripheral analysis*, the gammatone filterbank is commonly employed as the first stage of a cochlear model due to its correlation with physiological data and its computational efficiency. Following this, Meddis et al.'s (1990) model of the IHCs is a popular computational model of the conversion from basilar membrane displacement to neural activity. These data can then be used to extract cues such as periodicity and ITD. Neural activity data are then passed to *feature extraction* in order to reveal other acoustical cues such as AM, FM, onsets, offsets and cross-channel correlation. Thereafter, *mid-level representations* are created as an intermediate step between T–F units and groups. The segment is commonly used due to its perceptual relevance. A segment is a continuous region of the cochleagram and is created monaurally. Subsequently, *scene organisation* attempts to collect segments together to form streams such that a stream represents the sound originating from a single sound source. As in ASA, acoustical cues are used to inform grouping and grouping takes place both simultaneously and sequentially. The exact method of grouping often depends upon the acoustical cues that the system utilises in order to achieve separation. Finally, once the T–F mask has been calculated, *resynthesis* applies the T–F mask to the filterbank outputs in order to recreate the constituent signals of the mixture.

8.1.3 Chapter 4: CASA in Reverberant Environments

The chapter aimed to answer the following research sub-questions:

1. What are the problems posed by reverberation to human auditory perception in general?
2. What are the problems posed by reverberation to machine listening in general?
3. What are the human solutions to reverberation?

4. What are the machine listening solutions to reverberation, in particular in terms of source separation? How do machine listening solutions relate to human solutions?

5. Which reverberant source separation solution has most scope for improvement?

1. What are the problems posed by reverberation to human auditory perception in general?

It was established that reverberation poses several problems for human auditory perception. These problems include degradations in speech perception, source segregation and sound localisation. This is because reverberation blurs or destroys many cues—such as periodicity, the temporal and spectral envelopes and binaural cues—that humans rely on for these tasks.

2. What are the problems posed by reverberation to machine listening in general?

Similarly to the effects on human auditory perception, and for the same reasons, reverberation has deleterious effects on numerous aspects of machine listening including ASR, pitch tracking, binaural cues and onset and offset detection.

3. What are the human solutions to reverberation?

Humans have numerous mechanisms that are used in order to attempt to overcome the effects of reverberation. These mechanisms include: *utilising the slow temporal modulation of speech*, which occurs at rates below the envelope–filtering effect of reverberation; *the binaural advantage*, whereby listeners gain a significant advantage in many areas of perception by having two ears rather than one; *spectral envelope distortion compensation*, which counteracts the spectral distortion introduced by reverberation; and *precedence*, which weights the first few wavefronts of the direct sound over later wavefronts arriving as reflections from other surfaces.

4. What are the machine listening solutions to reverberation, in particular in terms of source separation? How do machine listening solutions relate to human solutions?

Several machine listening techniques can be utilised in order to reduce the deleterious effects of reverberation. *Dereverberation* removes reverberation before any further processing. This permits the use of existing algorithms that are untested in reverberation. This is an effective technique but its perceptual relevance remains unclear. *Spatial filtering* aims to enhance the target location and suppress sounds, including reverberation, arriving from other directions. However, this approach depends on the ability of the algorithm to locate the sound source, an ability that may be severely impeded by reverberation. Although it is possible that spatial filtering may be achieved via an EC mechanism, this link requires further research. *Utilising robust acoustic features* represents the signal using features that are robust to reverberation. Unfortunately, many of the approaches described in the literature are not usable in paradigms other than the one for which they were developed. Furthermore, this

technique has little or no perceptual relevance. *Reverberation masking* attempts to identify T–F regions that show minimal corruption by reverberation. This technique has questionable perceptual relevance and remains untested for CASA. *Precedence modelling* attempts to enhance source localisation estimates by modelling the perceptual precedence effect. The localisation data can then be used to inform grouping. This technique has perceptual relevance. *Utilisation of multiple cues* is motivated by the idea that if individual cues break down in reverberation, gathering data from many cues may achieve greater robustness to its effects. This approach has perceptual relevance since it is clear that humans use many acoustical cues in order to accomplish source segregation.

5. Which reverberant source separation solution has most scope for improvement?

Within the scope of the current investigation, modelling the precedence effect offers the most scope for improvement. There were four reasons for this: firstly, it is perceptually–relevant; secondly, it remains relatively untested for source separation; thirdly, there is a comprehensive existing body of work on computational precedence; and lastly, previous work has shown that with suitable processing, the reverberation–robustness of spatial cues can be improved. Furthermore, for other cues, it was shown that onsets and offsets are likely to be unreliable in reverberation, and pitch is only robust to reverberation if dereverberation processing is introduced, which has questionable perceptual relevance.

8.1.4 Chapter 5: Evaluating Source Separation in Reverberant Environments

The chapter aimed to answer research sub-question 6. Recall from the introduction that this question was:

6. How should the performance of different approaches to the chosen solution be evaluated? What signals? What metrics?

However, the question was adapted in the chapter in light of the findings of Chapter 4:

6′. How should the performance of separation algorithms incorporating different precedence models be evaluated? What signals? What metrics?

The algorithms were tested in a range of mixture conditions that incorporate a range of source–target azimuthal separations, TIRs, interferer signals and RT_{60}s, using a metric that facilitates meaningful comparison between different models and across different acoustic conditions. Specifically, the target signal was female speech; the interfering signals were male speech, music and noise. The Binaural Room Impulse Responses (BRIRs) were captured in real rooms. A novel metric meets the above criterion—Ideal Binary Mask Ratio (IBMR)—by comparing the calculated binary mask with the IBM. The metric is robust to the contribution of convolutional distortion to the output

because it compares the pattern of the calculated and ideal masks without weighting the contribution of each unit according to its local SNR.

8.1.5 Chapter 6: Modelling Precedence for Source Separation

The chapter aimed to answer the following research sub-question:

7. Which approaches work best and are there any lessons to be learned for future development?

A study was conducted that compared several computational precedence models and their impact on the performance of a baseline separation algorithm. The baseline algorithm included a precedence model, which was replaced with the other precedence models during the investigation. Of the models tested, the results showed that precedence models proposed by Martin (1997), Faller & Merimaa (2004), and Lindemann (1986a) work best and are a significant improvement on the baseline precedence model. Martin's model calculated an inhibitory signal based on onset data and multiplied this with the running cross-correlation. Faller & Merimaa's model calculated Interaural Coherence (IC) from the running normalised cross-correlation and used an IC threshold to specify cue selection. Lindemann's model is an extension of Jeffress' (1948) original cross-correlation theory of sound localisation. The model is extended with monaural detectors and a contralateral-inhibition mechanism. However, the results also indicated that it may be beneficial to adapt parameters of the precedence models to each room under test.

8.1.6 Chapter 7: Room–Specific Computational Precedence

The chapter aimed to answer the following research sub-questions:

8. Can performance be further improved?

9. Are the results generalisable?

8. Can performance be further improved?

The chapter investigated whether it was possible to improve the separation performance achieved by the precedence–model–enhanced separation algorithms presented in Chapter 6. Specifically, Chapter 6 hypothesised a room–specific component of the precedence models, whereby inhibitory parameters of models could be adapted in each room in order to optimise the performance. The subsequent investigation tested the baseline model with a range of mixture and inhibitory parameters in each room. The results showed that the performance achieved using the baseline model

could be further improved. Furthermore, this improvement was shown to be achieved by the hypothesised room–specific component. This component is analogous to the perceptual Clifton effect, which appears to adapt the precedence effect to the acoustic environment in which the listener is located.

9. Are the results generalisable?

In order to test whether this room–specific component was specific to the baseline model or present in other precedence models, the previous investigation was repeated with the other precedence models. These investigations confirmed that such a room–specific component was present and offered further performance improvement. However, the overall performance improvement was shown to be less than that for the baseline model. It can be concluded from these results that, at least for this source separation algorithm, that a model of the Clifton effect is a necessary part of a computational precedence model. The results also show that the choice of cross-correlation algorithm and cross-correlation parameters such as window length and shape are important considerations and must be chosen carefully.

8.1.7 Answer to the Main Research Question

The answers to the research sub-questions have been given in the preceding sections. The main research question can now be answered. This research question was:

> Can the reverberation–performance of existing psychoacoustic engineering approaches to machine source separation be improved?

The data presented in this thesis shows, at least by modelling the precedence effect, that yes, the reverberation–performance of existing psychoacoustic engineering approaches to machine source separation can be improved. However, the results also indicate that more performance improvement may be possible by modelling dynamic processes of the precedence effect.

8.2 Contributions to Knowledge

The research documented within this thesis has resulted in several innovations that make a contribution to knowledge. This section lists these innovations.

A Novel Metric for Assessing Separation Performance

During the development of the testing procedure, a novel metric (IBMR) was developed that is suitable for any separation algorithm that attempts to calculate the IBM. Tackling the issues posed by reverberation is a key research goal for many areas of

signal processing, including source separation (Jin & Wang 2009). A metric that is robust to reverberation is a key requisite for such research. Until this research, little consideration had been given to developing a metric that facilitated a meaningful comparison of algorithms when convolutional distortions were introduced. As discussed in Section 8.1.4, IBMR facilitates meaningful and direct comparison of separation algorithms, in particular in situations where acoustic conditions can not be held constant, or where it is important that the results should not be skewed by a particular set of acoustic conditions.

Comparing Computational Precedence Models

Work on computational precedence models has typically served one of two purposes: 1. to mimic psychoacoustic data obtained through experimentation; these models have seldom considered application. 2. a processing block for an algorithm that serves a specific application; in these cases the precedence effect has often provided justification for the processing scheme, but there has been little consideration of the multitude of computational precedence models and processing schemes proposed in the literature. This research has demonstrated some effective processing techniques, and in particular shown that a model based on IC (Faller & Merimaa 2004) can offer an improvement in separation performance of as much as 35%.

Towards a Computational Clifton Model

As stated in Chapter 7, the necessity for a computational Clifton model had been suggested in the literature, but not formally demonstrated. This research was the first to formally demonstrate the necessity for, and subsequently the potential advantage of, such a model. The research is a first step towards a computational Clifton model, which would be the first of its kind.

Incorporating Precedence into CASA

To date, few CASA models have included a precedence model in their peripheral processing or feature extraction stages. This research has demonstrated that including a precedence model can dramatically improve separation performance for a system based on spatial cues. Furthermore, in order to build a complete CASA model, a precedence model is a necessary constituent. This research is a first step towards such a component that retains perceptual relevance whilst actively improving separation performance.

8.3 Future Work

The research described in this thesis suggests three areas for further work. These areas are described in this section.

Quantification of Acoustical/Precedence Model Correlations and their Interactions

The BRIRs employed in this research were captured from real rooms; comparing rooms compares changes in numerous acoustical parameters, including ITDG, RT_{60} and DRR. In order to more precisely identify the correlations and interactions between these parameters and the optimal precedence model parameters, a set of controlled BRIRs needs to be created such that the acoustical parameters of the responses can be finely and independently controlled. This would lead to a series of mappings relating each acoustic parameter to its corresponding optimal precedence parameter(s). Furthermore, results obtained in Chapter 7 suggest that the precedence parameters are not independent of each other, but demonstrate a level of interaction. Therefore, further tests need to be conducted in which the performance achieved by additional combinations of precedence values is assessed, in order to quantify this interaction. The outcomes of these tests are likely to reveal improved performance in all of the models.

Quantification of the Contribution of Individual Precedence Processing Techniques

This research has considered numerous precedence models that each suggest different precedence processing techniques. Whilst the comparison of these models is useful and has highlighted some effective techniques, further work needs to be conducted in order to quantify the contribution of these techniques to the overall performance of the model. Specifically, there are numerous other differences between the models, including a variety of time constants, cross-correlation window lengths and shapes, cross-correlation algorithms, and IHC models. In order to more accurately compare the techniques, these differences need to be eliminated so that each technique can be isolated. The performance achieved by these techniques and their parameters can then be quantified. From this, precedence processing techniques could be combined in order to create a precedence model that might out-perform previously proposed models.

Build a Self-adapting Precedence Model

This future work is leading towards the eventual goal of a self-optimising precedence model and separation algorithm that can measure the acoustical properties of the room and optimise its precedence parameters accordingly. This adaptive processing is analogous to the perceptual Clifton effect. This self-optimising model has two pre-requisites: firstly, it requires the mappings described above to relate each acoustic parameter to its corresponding optimal precedence model parameter(s); secondly, it requires a method for blindly extracting the acoustic parameters of the room. A large body of work already exists in this field and producing algorithms to identify these parameters should be readily achievable.

Appendix

A Rooms used to Capture the BRIRs

This appendix describes each of the rooms used to obtain the BRIRs utilised in this research that were presented in Chapter 5. Each of the rooms is described in a corresponding section of this appendix. The octave-band RT_{60}s for each room are given in Table A.1. In all of the following cases, dimensions are in millimetres and the head height is measured from the floor to the centre of the ear canal; the diagrams are to scale. In all diagrams the arc on which the speaker was placed is shown and loudspeakers are shown at $-90°$, $0°$ and $90°$.

Room A

Room A was a typical medium-sized office that seats 8 people but had surprisingly small RT_{60} for its size. The room layout and dimensions are given in Figure A.1.

Room B

Room B was a medium–small class room. Despite the small shoebox shape, the construction of the room gave it a relatively long RT_{60} for its size. The room layout and dimensions are given in Figure A.2.

Room C

Room C was a large cinema–style lecture theatre that seats 418 people. However, the abundance of soft seating and the low ceiling of the area around the lectern resulted in a relatively small RT_{60} for the room's size. The room layout and dimensions are given in Figure A.3.

Room D

Room D was a typical medium–large sized seminar and presentation space with a very high ceiling. The room layout and dimensions are given in Figure A.4.

Table A.1: Octave-band and overall room RT_{60}s (in seconds).

Room	Octave-band Centre Frequency [Hz]							Overall
	125	**250**	**500**	**1k**	**2k**	**4k**	**8k**	
A	0.56	0.33	0.36	0.27	0.29	0.27	0.23	0.32
B	0.89	0.60	0.47	0.46	0.60	0.70	0.61	0.47
C	0.93	0.97	0.70	0.67	0.64	0.54	0.40	0.68
D	0.94	0.88	0.94	0.83	0.77	0.64	0.48	0.89

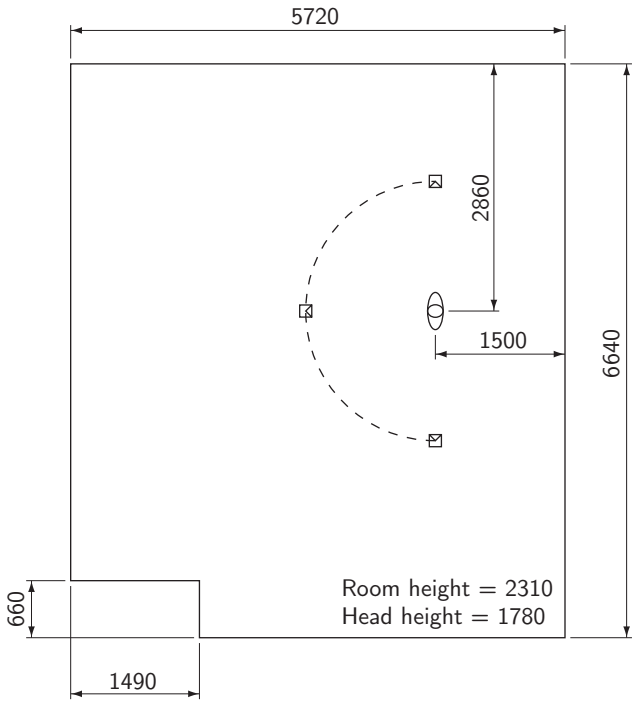


Figure A.1: Room A plan elevation and HATS location.

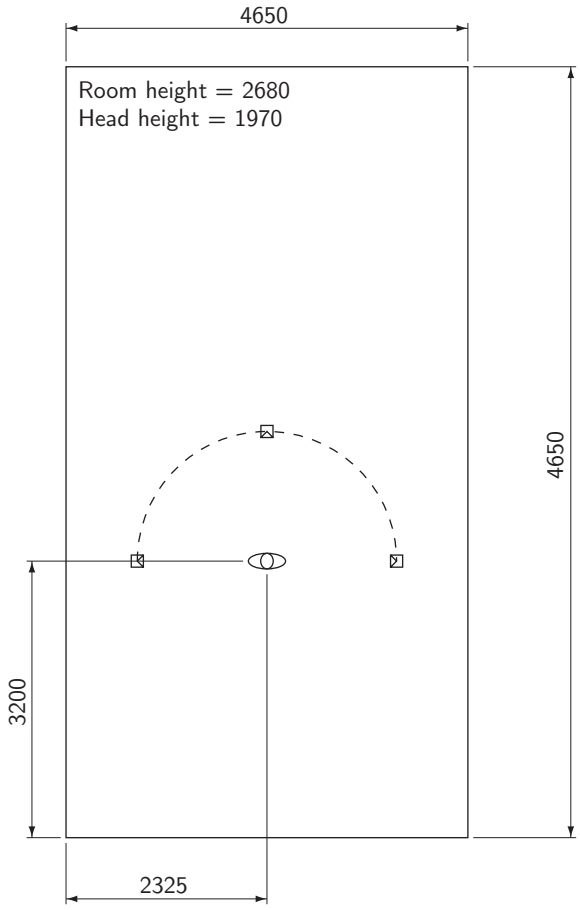


Figure A.2: Room B plan elevation and HATS location.

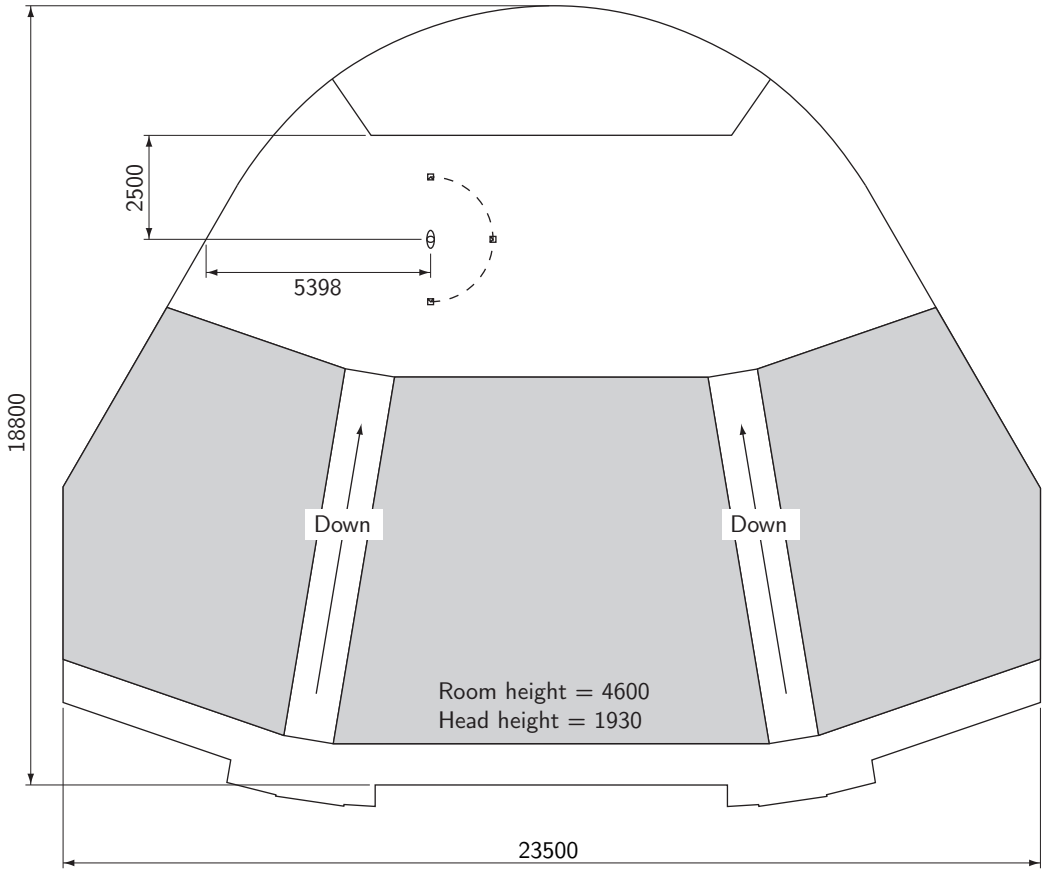


Figure A.3: Room C plan elevation and HATS location. The shaded area denotes banked seating; the room height is the height of the room at the HATS position

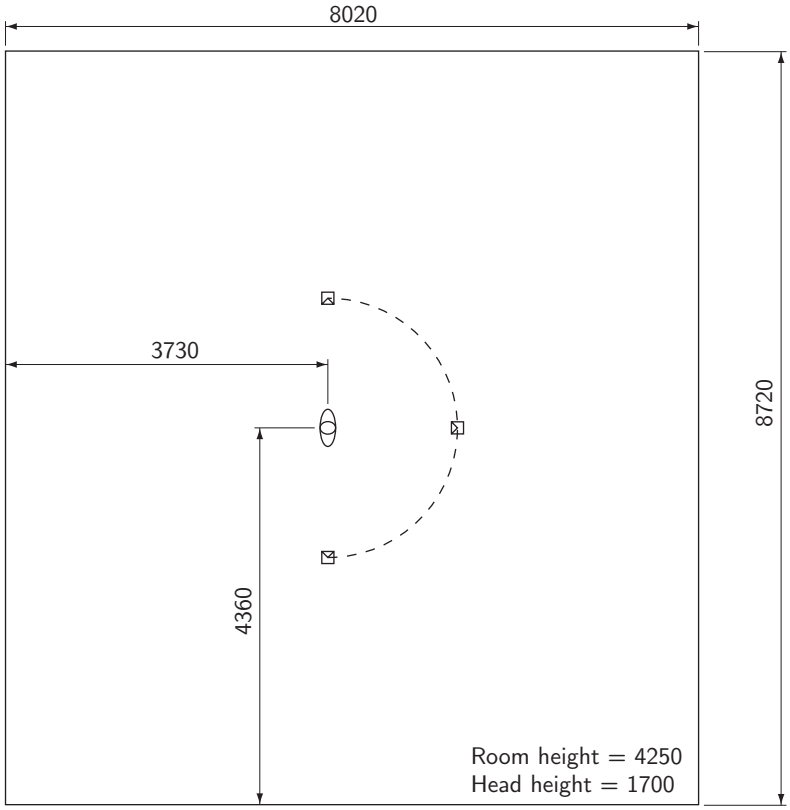


Figure A.4: Room D plan elevation and HATS location.

Appendix

B Additional Data for Chapter 6

Table B.1: Univariate ANOVA with IBMR as the dependent variable calculated over the interferer stimulus.

Source	**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**	**Partial η^2**
Corrected Model	14.597[a]	269	0.054	4.330	0.000	0.683
Intercept	108.507	1	108.507	8657.497	0.000	0.941
RT_{60}	1.733	4	0.433	34.563	0.000	0.204
Azi. Sep.	3.544	2	1.772	141.402	0.000	0.344
Model	2.210	5	0.442	35.271	0.000	0.246
TIR	0.713	2	0.357	28.447	0.000	0.095
RT_{60}* Azi. Sep.	0.468	8	0.059	4.668	0.000	0.065
RT_{60}* Model	0.758	20	0.038	3.025	0.000	0.101
RT_{60}* TIR	0.071	8	0.009	0.708	0.684	0.010
Azi. Sep. * Model	0.753	10	0.075	6.008	0.000	0.100
Azi. Sep. * TIR	0.184	4	0.046	3.668	0.006	0.026
Model * TIR	0.254	10	0.025	2.026	0.029	0.036
RT_{60}* Azi. Sep. * Model	1.082	40	0.027	2.158	0.000	0.138
RT_{60}* Azi. Sep. * TIR	0.155	16	0.010	0.772	0.718	0.022
RT_{60}* Model * TIR	0.512	40	0.013	1.022	0.437	0.070
Azi. Sep. * Model * TIR	0.828	20	0.041	3.301	0.000	0.109
RT_{60}* Azi. Sep. * Model * TIR	1.331	80	0.017	1.328	0.038	0.164
Error	6.768	540	0.013			
Total	129.871	810				
Corrected Total	21.365	809				

a. $R^2 = 0.683$ (Adjusted $R^2 = 0.525$)

Acronyms

AM	Amplitude Modulation
ANOVA	ANalysis Of VAriance
ASA	Auditory Scene Analysis
ASR	Automatic Speech Recognition
BMLD	Binaural Masking Level Difference
BRIR	Binaural Room Impulse Response
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CMN	Cepstral Mean Normalisation
DC	Direct Current
DFT	Discrete Fourier Transform
DRR	Direct-to-Reverberant Ratio
EC	Equalisation–Cancellation
ERB	Equivalent Rectangular Bandwidth
***F*0**	Fundamental Frequency
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FM	Frequency Modulation
HATS	Head And Torso Simulator
HMM	Hidden Markov Model
HRIR	Head–Related Impulse Response
HRTF	Head–Related Transfer Function
IBM	Ideal Binary Mask
IBMR	Ideal Binary Mask Ratio
IC	Interaural Coherence
ICA	Independent Component Analysis
IDFT	Inverse Discrete Fourier Transform
IF	Instantaneous Frequency
IHC	Inner Hair Cell
ILD	Interaural Level Difference
IMTF	Inverse Modulation Transfer Function
IPD	Interaural Phase Difference
ITD	Interaural Time Difference
ITDG	Initial Time Delay Gap
LP	Linear Prediction

MFCC	Mel–Frequency Cepstral Coefficient
MTF	Modulation Transfer Function
PLP	Perceptual Linear Prediction
RASTA	RelAtive SpecTrAl
RMS	Root Mean Square
RSNR	Reveberant-Signal-to-Noise Ratio
RT_{60}	Reverberation Time (to −60 dB)
SINR	Signal-to-Ideal-Noise Ratio
SNR	Signal-to-Noise Ratio
SRT	Speech Reception Threshold
STI	Speech Transmission Index
T–F	Time–Frequency
TIR	Target-to-Interferer Ratio

Mathematical Symbols

Time–Frequency Matrices

$\mathbf{a}$	Correlogram
$\hat{\mathbf{a}}$	Normalised correlogram
$\mathbf{c}$	Cross-correlogram
$\hat{\mathbf{c}}$	Normalised cross-correlation
$\acute{\mathbf{c}}$	Running cross-correlation
$\mathbf{c}_\iota$	Inhibited cross-correlation
$\mathbf{h}$	Inner Hair Cell (IHC)–modelled data
$\acute{\mathbf{h}}$	Modified IHC–modelled data
$\mathbf{k}$	Cross-channel correlation
$\hat{\mathbf{k}}$	Normalised cross-channel correlation
$\mathbf{m}$	Binary mask
$\mathbf{m}_{\text{ibm}}$	Ideal Binary Mask (IBM)
$\mathbf{OIR}$	Output-to-Input Energy Ratio in Roman & Wang's (2004) model
$\mathbf{q}$	Local peaks of $\mathbf{c}$
$\mathbf{r}$	Precedence–modelled fine structure
$\mathbf{s}$	Skeleton cross-correlogram
$\mathbf{u}$	Auditory nerve firing rate
$\acute{\mathbf{u}}$	Auditory energy
$\mathbf{x}$	Excitation envelope
$\mathbf{y}$	Post-cancellation signal residue in Roman & Wang's (2004) model
$\mathbf{z}$	Signal input in Roman & Wang's (2004) model
$\boldsymbol{\gamma}$	Gammatone filterbank output / basilar membrane displacement
$\boldsymbol{\varepsilon}$	Hilbert envelopes
$\acute{\boldsymbol{\varepsilon}}$	Smoothed Hilbert envelopes
$\boldsymbol{\zeta}$	ILD Template
$\boldsymbol{\iota}$	Inhibitory signal
$\boldsymbol{\Xi}$	Running energy average
$\boldsymbol{\Phi}$	Dynamic inhibitory signal
$\boldsymbol{\chi}$	Interaural Coherence (IC)

Other Matrices

G_o	Differentiating and smoothing kernel for onset/offset detection
G'_o	Differentiating and smoothing function for onset/offset detection
G_{FM}	Frequency Modulation (FM) kernel
FM	FM operation

Vectors

$\bar{\mathbf{a}}$	Pooled correlogram
$\bar{\mathbf{c}}$	Pooled cross-correlogram
c_γ	Derived inhibition parameter
E	Equivalent Rectangular Bandwidth (ERB)–rate scale
gt	Gammatone filter time response
$\tilde{gt}$	Phase–corrected gammatone filter time response
GT	Gammatone filter frequency response
h_{lp}	Onset de-emphasising low-pass filter
o	Differentiated and smoothed x
p	Monaural sensitivity function in Lindemann's model
P	Instantaneous pressure of an impulse response
$\bar{\mathbf{s}}$	Pooled skeleton cross-correlogram
s	Estimated target signal
s_t	Target signal
s_r	Reverberated target signal
s_{ibm}	Signal resynthesised from the mixture using the IBM
w	Generic window function
w_m	Window function function used in Macpherson's model
x, x_1, x_2	Generic time-domain signal
$\tilde{x}$	Analytic of x
x_{pc}	Power cepstrum of x
X	Z-transform of x
y_1, y_2	Signals with different fundamental frequencies
z	Mixture of y_1 and y_2
β	Spectral energy normalisation factor
ε	Hilbert envelope
θ	Unit step function
Θ_i	Frequency–dependent grouping thresholds in Roman & Wang's (2004) model
Π	Frequency–dependent multiplier for converting between Interaural Time Difference (ITD) and azimuth
σ	Standard deviation

Constants

A	Set to give unity gain at DC
b	Gammatone filter bandwidth parameter: $b = 1.019\,\text{ERB}$
$\mathring{\mathbf{c}}$	Cross-correlogram peak
c_0	Speed of sound: 344 m s^{-1}
c_{inh}	Target inhibition parameter: (0,1]
C_{te}	Clarity index (dB)
f_0	Gammatone filter centre frequency
G	Inhibitory gain factor
I	Number of frequency channels: 32
L	Frame length: 10 ms (in samples)

M	Generic window length
M_c	Cross-correlation window length in Macpherson's model: 2 ms
N	Gammatone filter order: 4
Q	Autocorrelation sharpening factor
r	Head radius: 0.093 m
T	Maximum cross-correlation lag: 1 ms (in samples)
t_c	Time lead to align gammatone filter phase at $t = 0$
U, V	Parameters for determining two separate fundamental frequencies in a mixture
α_f	Time constant in Faller & Merimaa's model: 10 ms (in samples)
α_{inh}	Fade-off time constant in Lindemann's model: 10 ms (in samples)
α_m	Time constant in Martin's and Macpherson's model: 1.5 ms (in samples)
α_p	Time constant in the baseline model: 15 ms (in samples)
α_s	Time constant to smooth the Hilbert envelope: 8 ms (in samples)
σ_f	Standard deviation, in frequency, for the FM kernel (Hz)
σ_t	Standard deviation, in time, for the FM kernel (seconds)
Θ_c	Cross-correlogram grouping threshold: -160 dB
Θ_{ibm}	IBM threshold: 0 dB
Θ_m	Mask threshold: [0,1)
Θ_r	Rate threshold: -11 dB
Θ_χ	IC threshold: [0,1)
Λ	The length of the input signal in samples
ϕ_1, ϕ_2	Largest peaks in the pooled skeleton cross-correlogram
ϕ_t	Target azimuth
ϕ_n	Interferer azimuth

Sets

$\mathbb{N}$	Natural numbers
$\mathbb{Z}$	Integers
$\varnothing$	Empty
ψ_ϕ	Peaks in the pooled skeleton cross-correlogram
$\Psi, \psi, \acute{\psi}$	Precedence–model–specific sets

Indices

d	Summation index
f	Frequency (Hz)
i	Frequency channel
k	Ear
l	Frame
n	Sample
t	Time (seconds)
v	Frequency channels above 2.8 kHz
τ	Correlation lag
ϕ	Azimuth (°)

ω	Angular frequency (rad s^{-1})

Operators

$*$	Convolution
$\|\ldots\|$	Modulus for real or complex numbers; cardinality (size) for sets
$\cap$	Set intersection
$\subset$	Subset
$\wedge$	Binary logical AND
$\oplus$	Binary logical XOR
DFT	Discrete Fourier Transform
H	Hilbert transform
IDFT	Inverse Discrete Fourier Transform

References

Aikawa, K., Singer, H., Kawahara, H. & Tohkura, Y. (1996), Cepstral representation of speech motivated by time–frequency masking: An application to speech recognition, *The Journal of the Acoustical Society of America* 100, 1, 603–614.

Albeck, Y. (2003), Sound localization and binaural processing, in M.A. Arbib (ed.) *The Handbook of Brain Theory and Neural Networks*, Cambridge, MA: MIT Press, 1061–1064.

Allen, J.B. (1973), Speech dereverberation, *The Journal of the Acoustical Society of America* 53, 1, 322.

Aoki, M. & Furuya, K. (2002), Real-time source separation based on sound localization in a reverberant environment, in *Neural Networks for Signal Processing, Proceedings of the 12th IEEE Workshop on*, 475–484.

Atal, B.S. (1972), Automatic speaker recognition based on pitch contours, *The Journal of the Acoustical Society of America* 52, 6B, 1687–1697.

Audience Inc. (2008), Press release: Audience introduces industry-first voice processor based on human hearing system and begins sampling to mobile handset manufacturers, `http://www.audience.com/news_20080211b.html`, accessed: 23/04/08.

Avendano, C. & Hermansky, H. (1996), Study on the dereverberation of speech based on temporal envelope filtering, in *Spoken Language Processing (ICSLP), The Fourth International Conference on*, volume 2, 889–892.

Banks, D. (1993), Localisation and separation of simultaneous voices with two microphones, *Communications, Speech and Vision, IEE Proceedings I* 140, 4, 229–234.

Barker, J., Ma, N., Coy, A. & Cooke, M. (2010), Speech fragment decoding techniques for simultaneous speaker identification and speech recognition, *Computer Speech and Language* 24, 1, 94–111.

Bell, A.J. & Sejnowski, T. (1995), An information–maximisation approach to blind separation and blind deconvolution, *Neural Computation* 7, 6, 1129–1159.

Bernstein, L.R., Van de Par, S. & Trahiotis, C. (1999), The normalized interaural correlation: Accounting for NoSπ thresholds obtained with gaussian and "low-noise" masking noise, *The Journal of the Acoustical Society of America* 106, 2, 870–876.

Blauert, J. & Col, J.P. (1992), A study of temporal effects in spatial hearing, in Y. Cazals, L. Demany & K. Horner (eds.) *Auditory Psychology and Perception*, Oxford: Pergamon Press, 531–538.

Blauert, J. (1997), *Spatial Hearing: The psychophysics of human sound localization*, revised edition, Cambridge, MA: MIT Press.

Bodden, M. (1993), Modelling human sound-source localization and the cocktail party effect, *Acta Acustica* 1, 43–55.

Bogart, B.P., Healy, M.J.R. & Tukey, J.W. (1963), The quefrency analysis of time-series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking, in M. Rosenblatt (ed.) *Time Series Analysis*, New York: Wiley, 209.

Bolt, R.H. & MacDonald, A.D. (1949), Theory of speech masking by reverberation, *The Journal of the Acoustical Society of America* 21, 6, 577–580.

Braasch, J. (2005), Modelling of binaural hearing, in J. Blauert (ed.) *Communication Acoustics*, Berlin Heidelberg: Springer-Verlag, 75–108.

Brand, A., Behrend, O., Marquardt, T., McAlpine, D. & Grothe, B. (2002), Precise inhibition is essential for microsecond interaural time difference coding, *Nature* 417, 543–547.

Brandstein, M.S. (1999), Time-delay estimation of reverberated speech exploiting harmonic structure, *The Journal of the Acoustical Society of America* 105, 5, 2914–2919.

Bregman, A.S. (1990), *Auditory Scene Analysis*, Cambridge, MA: MIT Press.

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

——— (1994b), Perceptual grouping of musical sounds: a computational model, *Journal of New Music Research* 23, 2, 107–132.

Brown, G.J. & Palomäki, K.J. (2006), Reverberation, in D. Wang & G.J. Brown (eds.) *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Hoboken, NJ: John Wiley & Sons, 209–250.

BS EN ISO 226 (2003), Normal equal–loudness–level contours.

BS EN ISO 3382 (2000), Measurement of the reverberation time of rooms with reference to other acoustical parameters.

Carlson, B. & Clements, M. (1991), A computationally compact divergence measure for speech processing, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13, 12, 1255–1260.

Cherry, E.C. (1957), *On Human Communication*, Cambridge, MA: MIT Press.

Childers, D., Skinner, D. & Kemerait, R. (1977), The cepstrum: A guide to processing, *Proceedings of the IEEE* 65, 10, 1428–1443.

Christensen, H., Ma, N., Wrigley, S. & Barker, J. (2007), Integrating pitch and localisation cues at a speech fragment level, in *Proceedings of Interspeech*, Antwerp, Belgium.

Christensen, H., Ma, N., Wrigley, S.N. & Barker, J. (2009), A speech fragment approach to localising multiple speakers in reverberant environments, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 4593–4596.

Clifton, R.K. & Freyman, R.L. (1989), Effect of click rate and delay on breakdown of the precedence effect, *Perception & Psychophysics* 46, 2, 139–145.

Clifton, R.K. (1987), Breakdown of echo suppression in the precedence effect, *The Journal of the Acoustical Society of America* 82, 5, 1834–1835.

Cooke, M. (1991), *Modelling auditory processing and organisation*, Ph.D. thesis, University of Sheffield.

Couvreur, L. & Couvreur, C. (2004), Blind model selection for automatic speech recognition in reverberant environments, *The Journal of VLSI Signal Processing* 36, 2–3, 189–203.

Coy, A. & Barker, J. (2007), An automatic speech recognition system based on the scene analysis account of auditory perception, *Speech Communication* 49, 5, 384–401.

Culling, J.F., Hodder, K.I. & Toh, C.Y. (2003), Effects of reverberation on perceptual segregation of competing voices, *The Journal of the Acoustical Society of America* 114, 5, 2871–2876.

Culling, J.F. & Lewis, H.G. (2010), Trading of intensity and interaural coherence in dichotic pitch stimuli, *The Journal of the Acoustical Society of America* 128, 4, 1908–1914.

Culling, J.F. & Summerfield, Q. (1995), Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay, *The Journal of the Acoustical Society of America* 98, 2, 785–797.

Culling, J.F., Summerfield, Q. & Marshall, D.H. (1994), Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels, *Speech Communication* 14, 1, 71–95.

Darwin, C.J. & Hukin, R.W. (2000), Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention, *The Journal of the Acoustical Society of America* 108, 1, 335–342.

Darwin, C.J. & Sutherland, N.S. (1984), Grouping frequency components of vowels: when is a harmonic not a harmonic?, *Quarterly Journal of Experimental Psychology* 36A, 193–208.

De Boer, E. & Kruidenier, C. (1990), On ringing limits of the auditory periphery, *Biological Cybernetics* 63, 6, 433–442.

De Cheveigne, A. (1991), Speech $F0$ extraction based on licklider's pitch perception model, in *Proceedings of the 12th International Conference of Phonetic Science*, Aix-en-Provence, 218–221.

——— (2006), Multiple $F0$ estimation, in D. Wang & G.J. Brown (eds.) *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Hoboken, NJ: John Wiley & Sons, 45–79.

De Cheveigne, A. & Kawahara, H. (2002), YIN, a fundamental frequency estimator for speech and music, *The Journal of the Acoustical Society of America* 111, 4, 1917–1930.

Deutsch, D. (1975), Two-channel listening to musical scales, *The Journal of the Acoustical Society of America* 57, 5, 1156–1160.

Devore, S. & Shinn-Cunningham, B. (2003), Perceptual consequences of including reverberation in spatial auditory displays, in *Proceedings of the International Conference on Auditory Displays*, Boston, MA, 75–78.

Ditech Networks (2008), Voice quality audit report, `http://www.ditechnetworks.com/auditreport.html`, accessed: 23/04/08.

Dizon, R.M., Litovsky, R.Y. & Colburn, H.S. (1997), Positional dependence on localization dominance in the median-sagittal plane, *The Journal of the Acoustical Society of America* 101, 5, 3106.

Dizon, R.M. & Litovsky, R.Y. (2004), Localization dominance in the median-sagittal plane: Effect of stimulus duration, *The Journal of the Acoustical Society of America* 115, 6, 3142–3155.

Djelani, T. & Blauert, J. (2001a), Investigation into the build-up and breakdown of the precedence effect, *Acta Acustica United with Acustica* 87, 253Ð261.

——— (2001b), Some new aspects of the build-up and breakdown of the precedence effect, in D.J. Breebart, A.V.J.M. Houtsma, A. Kohlrausch, V.F. Prijs & R. Schoonhoven (eds.) *Psychological and Physiological Bases of Auditory Function*, Aachen: Shaker Maastricht, 200–207.

Drennan, W.R., Gatehouse, S. & Lever, C. (2003), Perceptual segregation of competing speech sounds: The role of spatial location, *The Journal of the Acoustical Society of America* 114, 4, 2178–2189.

Drullman, R., Festen, J.M. & Plomp, R. (1994a), Effect of reducing slow temporal modulations on speech reception, *The Journal of the Acoustical Society of America* 95, 5, 2670–2680.

——— (1994b), Effect of temporal envelope smearing on speech reception, *The Journal of the Acoustical Society of America* 95, 2, 1053–1064.

Dudley, H. (1939), Remaking speech, *The Journal of the Acoustical Society of America* 11, 2, 169–177.

Durlach, N.I. (1963), Equalization and cancellation theory of binaural masking-level difference, *The Journal of the Acoustical Society of America* 35, 1206–1218.

——— (1972), Binaural signal detection: Equalization and cancellation theory, in J. Tobias (ed.) *Foundations of Modern Auditory Theory*, volume 2, New York: Academic, 369–462.

Durlach, N.I. & Colburn, H.S. (1978), Binaural phenomena, in E.C. Carterette & M.P. Friedman (eds.) *Handbook of Perception, Vol. IV*, New York: Academic Press, 365–466.

Ebata, M., Sone, T. & Nimura, T. (1968), On the perception of direction of echo, *The Journal of the Acoustical Society of America* 44, 2, 542–547.

EBU SQAM (1988), Sound quality assessment material for subjective listening tests, Tech. 3253-E, European Broadcasting Union, `http://tech.ebu.ch/publications/sqamcd`.

Edmonds, B.A. & Culling, J.F. (2005), The spatial unmasking of speech: evidence for within-channel processing of interaural time delay, *The Journal of the Acoustical Society of America* 117, 5, 3069–3078.

Ellis, D.P.W. (1996), *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, Massachusetts Institute of Technology.

Eneman, K., Duchateau, J., Moonen, M., Van Compernolle, D. & Van Hamme, H. (2003), Assessment of dereverberation algorithms for large vocabulary speech recognition systems, in *8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 1–4.

Faller, C. & Merimaa, J. (2004), Source localization in complex listening situations: Selection of binaural cues based on interaural coherence, *The Journal of the Acoustical Society of America* 116, 5, 3075–3089.

Feng, A.S. & Jones, D.L. (2006), Localization-based grouping, in D. Wang & G.J. Brown (eds.) *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Hoboken, NJ: John Wiley & Sons, 187–207.

Fincham, L. (1985), Refinements in the impulse testing of loudspeakers, *Journal of the Audio Engineering Society* 33, 3, 133–140.

Freyman, R.L., Clifton, R.K. & Litovsky, R.Y. (1991), Dynamic processes in the precedence effect, *The Journal of the Acoustical Society of America* 90, 4, 874–884.

Gardner, B. & Martin, K. (1994), HRTF measurements of a KEMAR dummy-head microphone, Technical Report 280, MIT Media Lab, `http://sound.media.mit.edu/resources/KEMAR.html`.

Gaskell, H. (1983), The precedence effect, *Hearing Research* 12, 3, 277–303.

Gelfand, S.A. & Silman, S. (1979), Effects of small room reverberation upon the recognition of some consonant features, *The Journal of the Acoustical Society of America* 66, 1, 22–29.

Giguère, C. & Abel, S.M. (1993), Sound localization: Effects of reverberation time, speaker array, stimulus frequency, and stimulus rise/decay, *The Journal of the Acoustical Society of America* 94, 2, 769–776.

Gillespie, B. & Atlas, L. (2002), Acoustic diversity for improved speech recognition in reverberant environments, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, 557–560.

Giuliani, D., Omologo, M. & Svaizer, P. (1996), Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation, in *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia, PA, 1329–1332.

Glasberg, B.R. & Moore, B.C.J. (1990), Derivation of auditory filter shapes from notched-noise data, *Hearing Research* 47, 1–2, 103–138.

Godsmark, D. & Brown, G.J. (1999), A blackboard architecture for computational auditory scene analysis, *Speech Communication* 27, 3–4, 351–366.

Gölzer, H. & Kleinschmidt, M. (2003), Importance of early and late reflections for automatic speech recognition in reverberant environments, in *Proceedings of Elektronische Sprachsignalverarbeitung (ESSV)*.

Grantham, D.W. (1996), Left–right asymmetry in the buildup of echo suppression in normal-hearing adults, *The Journal of the Acoustical Society of America* 99, 2, 1118–1123.

Greenwood, D.D. (1961), Critical bandwidth and the frequency coordinates of the basilar membrane, *The Journal of the Acoustical Society of America* 33, 10, 1344–1356.

Haas, H. (1951), Über den einfluss eines einfachechos auf die hörsamkeit von sprache (On the influence of a single echo on the intelligibility of speech), *Acustica* 1, 48–58.

Haggard, M. (1974), Selectivity for distortions and words in speech perception, *British Journal of Psychology* 65, 1, 69–83.

Hartmann, W.M. (1983), Localization of sound in rooms, *The Journal of the Acoustical Society of America* 74, 5, 1380–1391.

——— (1998), *Signals, Sound and Sensation*, New York: Springer.

Hatziantoniou, P.D. & Mourjopoulos, J.N. (2004), Real-time room equalization based on complex smoothing: Robustness results, in *Proceedings of the 116th Audio Engineering Society Convention*, Berlin, paper 6070.

Helmholtz, H. (1885), *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, second English edition, New York: Dover Publishers, translated by Alexander J. Ellis from the fourth German edition.

Hermansky, H. & Morgan, N. (1994), RASTA processing of speech, *IEEE Transactions on Speech and Audio Processing* 2, 4, 578–589.

Hermansky, H. (1990), Perceptual linear predictive (PLP) analysis of speech, *The Journal of the Acoustical Society of America* 87, 4, 1738–1752.

Holdsworth, J., Nimmo-Smith, I., Patterson, R. & Rice, P. (1988), Implementing a gamma tone filter bank, Technical report, MRC Applied Psychology Unit, Cambridge.

Hoover, A., Jean-Baptiste, G., Jiang, X., Flynn, P.J., Bunke, H., Goldgof, D.B., Bowyer, K., Eggert, D.W., Fitzgibbon, A. & Fisher, R.B. (1996), An experimental comparison of range image segmentation algorithms, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18, 7, 673–689.

Houtgast, T. & Steeneken, H.J.M. (1985), A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria, *The Journal of the Acoustical Society of America* 77, 3, 1069–1077.

——— (1973), The modulation transfer function in room acoustics as a predictor of speech intelligibility, *Acustica* 28, 66–73.

Hu, G. & Wang, D. (2007), Auditory segmentation based on onset and offset analysis, *IEEE Transactions on Audio, Speech, and Language Processing* 15, 2, 396–405.

——— (2004a), Auditory segmentation based on event detection, in *Proceedings of ISCA tutorial and research workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea.

——— (2004b), Monaural speech segregation based on pitch tracking and amplitude modulation, *IEEE Transactions on Neural Networks* 15, 5, 1135–1150.

Hummersone, C., Mason, R. & Brookes, T. (2010), Dynamic precedence effect modeling for source separation in reverberant environments, *IEEE Transactions on Audio, Speech, and Language Processing* 18, 7, 1867–1871.

——— (2011), Ideal Binary Mask Ratio: a novel metric for assessing binary–mask–based sound source separation algorithms, *IEEE Transactions on Audio, Speech, and Language Processing* (accepted).

Jeffress, L. (1948), A place theory of sound localization, *Journal of Comparative Psychology* 41, 1, 35–39.

Jin, Z. & Wang, D. (2009), A supervised learning approach to monaural segregation of reverberant speech, *IEEE Transactions on Audio, Speech, and Language Processing* 17, 4, 625–638.

Jurafsky, D. & Martin, J.H. (2009), *Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition*, second edition, Upper Saddle River, NJ: Pearson Education.

King, A.J. & Hutchings, M.E. (1987), Spatial response properties of acoustically responsive neurons in the superior collicus of the ferret: a map of auditory space, *Journal of Neurophysiology* 57, 2, 596–624.

Kingsbury, B.E.D. (1998), *Perceptually inspired signal-processing strategies for robust speech recognition in Reverberant environments*, Ph.D. thesis, University of California.

Kingsbury, B.E.D., Morgan, N. & Greenberg, S. (1998), Robust speech recognition using the modulation spectrogram, *Speech Communication* 25, 1–3, 117–132.

Koenig, W. (1950), Subjective effects in binaural hearing, *The Journal of the Acoustical Society of America* 22, 1, 61–62.

Kollmeier, B. & Koch, R. (1994), Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction, *The Journal of the Acoustical Society of America* 95, 3, 1593–1602.

Kubovy, M. (1981), Concurrent-pitch segregation and the theory of indispensable attributes, in M. Kubovy & J.R. Pomerantz (eds.) *Perceptual Organization*, Hillsdale, NJ: Erlbaum, 55–98.

Kuhn, G.F. (1977), Model for the interaural time differences in the azimuthal plane, *The Journal of the Acoustical Society of America* 62, 1, 157–167.

Kumaresan, R. & Rao, A. (1999), Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications, *The Journal of the Acoustical Society of America* 105, 3, 1912–1924.

Langhans, T. & Strube, H. (1982), Speech enhancement by nonlinear multiband envelope filtering, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 7, 156–159.

Lee, T.W., Lewicki, M.S., Girolami, M. & Sejnowski, T. (1999), Blind source separation of more sources than mixtures using overcomplete representations, *Signal Processing Letters, IEEE* 6, 4, 87–90.

Li, N. & Loizou, P.C. (2008), Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction, *The Journal of the Acoustical Society of America* 123, 3, 1673–1682.

Li, Y. & Wang, D. (2009), On the optimality of ideal binary time–frequency masks, *Speech Communication* 51, 3, 230–239.

Libbey, B. & Rogers, P.H. (2004), The effect of overlap-masking on binaural reverberant word intelligibility, *The Journal of the Acoustical Society of America* 116, 5, 3141–3151.

Libbey, B.W. & Rogers, P.H. (2000), Human capabilities of dereverberation, *The Journal of the Acoustical Society of America* 107, 5, 2822–2822.

Licklider, J.C.R. (1951), A duplex theory of pitch perception, *Experimentia* 7, 4, 128–133.

Lindemann, W. (1986a), Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals, *The Journal of the Acoustical Society of America* 80, 6, 1608–1622.

——— (1986b), Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front, *The Journal of the Acoustical Society of America* 80, 6, 1623–1630.

Litovsky, R.Y., Colburn, H.S., Yost, W.A. & Guzman, S.J. (1999a), The precedence effect, *The Journal of the Acoustical Society of America* 106, 4, 1633–1654.

Litovsky, R.Y., Dizon, R.M. & Colburn, H. (1999b), Studies of the precedence effect in the median-sagittal and azimuthal planes in a virtual acoustic space, submitted to *The Journal of the Acoustical Society of America* (unpublished).

Litovsky, R.Y., Hawley, M.L. & Colburn, H.S. (1997a), Measurement of precedence in monaural listeners, Meeting of the American Speech and Hearing Association.

Litovsky, R.Y., Rakerd, B., Yin, T.C.T. & Hartmann, W.M. (1997b), Psychophysical and physiological evidence for a precedence effect in the median sagittal plane, *Journal of Neurophysiology* 77, 4, 2223–2226.

Liu, C., Wheeler, B.C., O'Brien, J., Bilger, R.C., Lansing, C.R. & Feng, A.S. (2000), Localization of multiple sound sources with two microphones, *The Journal of the Acoustical Society of America* 108, 4, 1888–1905.

Liu, C., Wheeler, B.C., O'Brien, J., Lansing, C.R., Bilger, R.C., Jones, D.L. & Feng, A.S. (2001), A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers, *The Journal of the Acoustical Society of America* 110, 6, 3218–3231.

Liu, F.H., Stern, R.M., Huang, X. & Acero, A. (1993), Efficient cepstral normalization for robust speech recognition, in *Proceedings of the Sixth ARPA Workshop on Human Language Technology*, Plainsboro, NJ, 69–74.

Lochner, J.P.A. & Burger, J.F. (1958), The subjective masking of short time delayed echoes, their primary sounds, and their contribution to the intelligibility of speech, *Acustica* 8, 1–10.

——— (1964), The influence of reflections on auditorium acoustics, *Journal of Sound and Vibration* 1, 4, 426–448.

Lockwood, M.E., Jones, D.L., Bilger, R.C., Lansing, C.R., O'Brien, J., Wheeler, B.C. & Feng, A.S. (2004), Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms, *The Journal of the Acoustical Society of America* 115, 1, 379–391.

Lockwood, M.E., Jones, D.L., Su, Q. & Miles, R.N. (2003), Beamforming with collocated microphone arrays, *The Journal of the Acoustical Society of America* 114, 4, 2451.

Lyon, R. (1983), A computational model of binaural localization and separation, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 8, 1148–1151.

Ma, N., Green, P., Barker, J. & Coy, A. (2007), Exploiting correlogram structure for robust speech recognition with multiple speech sources, *Speech Communication* 49, 12, 874–891.

Macpherson, E.A. (1991), A computer model of binaural localization for stereo imaging measurement, *Journal of the Audio Engineering Society* 39, 9, 604–622.

Makhoul, J. (1975), Linear prediction: A tutorial review, *Proceedings of the IEEE* 63, 4, 561–580.

——— (1976), Correction to "Linear prediction: A tutorial review", *Proceedings of the IEEE* 64, 4, 285.

Mandel, M., Bressler, S., Shinn-Cunningham, B. & Ellis, D. (2010), Evaluating source separation algorithms with reverberant speech, *IEEE Transactions on Audio, Speech, and Language Processing* 18, 7, 1872–1883.

Marr, D. (1982), *Vision: a computational investigation into the human representation and processing of visual information*, New York: W. H. Freeman.

Martin, K. (1997), Echo suppression in a computational model of the precedence effect, in *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on*, New Paltz, NY.

Meddis, R. (1986), Simulation of mechanical to neural transduction in the auditory receptor, *The Journal of the Acoustical Society of America* 79, 3, 702–711.

——— (1988), Simulation of auditory–neural transduction: Further studies, *The Journal of the Acoustical Society of America* 83, 3, 1056–1063.

Meddis, R., Hewitt, M.J. & Shackleton, T.M. (1990), Implementation details of a computation model of the inner hair-cell auditory-nerve synapse, *The Journal of the Acoustical Society of America* 87, 4, 1813–1816.

Mellinger, D.K. (1991), *Event formation and separation of musical sound*, Ph.D. thesis, Stanford University.

Moore, B.C.J. (2004), *An Introduction to the Psychology of Hearing*, fifth edition, London: Academic Press.

Morgan, D., George, E., Lee, L. & Kay, S. (1997), Cochannel speaker separation by harmonic enhancement and suppression, *IEEE Transactions on Speech and Audio Processing* 5, 5, 407–424.

Nakatani, T., Juang, B.H., Kinoshita, K. & Miyoshi, M. (2005), Harmonicity based dereverberation with maximum a posteriori estimation, in *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on*, 94–97.

Nakatani, T., Miyoshi, M. & Kinoshita, K. (2004), One microphone blind dereverberation based on quasi-periodicity of speech signals, in S. Thrun, L.K. Saul & B. Schölkopf (eds.) *Advances in Neural Informaton Processing Systems 16*, Cambridge, MA: MIT Press, 1417–1424.

Nakatani, T. & Okuno, H.G. (1999), Harmonic sound stream segregation using localization and its application to speech stream segregation, *Speech Communication* 27, 209–222.

Oppenheim, A. & Schafer, R. (1968), Nonlinear filtering of multiplied and convolved signals, *Proceedings of the IEEE* 56, 8, 1264–1291.

——— (1999), *Discrete-Time Signal Processing*, second edition, Upper Saddle River, NJ: Prentice-Hall.

Palomäki, K.J., Brown, G.J. & Barker, J.P. (2004a), Techniques for handling convolutional distortion with "missing data" automatic speech recognition, *Speech Communication* 43, 1–2, 123–142.

Palomäki, K.J., Brown, G.J. & Wang, D. (2004b), A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation, *Speech Communication* 43, 4, 361–378.

Palomäki, K., Brown, G. & Barker, J. (2002), Missing data speech recognition in reverberant conditions, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, 65–68.

Park, H.M. & Stern, R. (2007), Missing feature speech recognition using dereverberation and echo suppression in reverberant environments, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, 381–384.

Parsons, T.W. (1976), Separation of speech from interfering speech by means of harmonic selection, *The Journal of the Acoustical Society of America* 60, 4, 911–918.

Patterson, R., Nimmo-Smith, I., Holdsworth, J. & Rice, P. (1987), An efficient auditory filterbank based on the gammatone function, Technical report, MRC Applied Psychology Unit, Cambridge.

Pickles, J.O. (2008), *An introduction to the physiology of hearing*, third edition, London: Academic Press.

Plomp, R. (1976), Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise), *Acustica* 34, 200–211.

Repp, B.H. (1987), On the possible role of auditory short-term adaptation in perception of the prevocalic [m]–[n] contrast, *The Journal of the Acoustical Society of America* 82, 5, 1525–1538.

Riley, M.D. (1989), *Speech Time-Frquency Representations*, Boston: Kluwer Academic.

Roman, N. & Wang, D. (2004), Binaural sound segregation for multisource reverberant environments, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, 373–376.

——— (2005), A pitch–based model for separation of reverberant speech, in *Proceedings of Interspeech*, Lisbon.

Roman, N., Wang, D. & Brown, G.J. (2003), Speech segregation based on sound localization, *The Journal of the Acoustical Society of America* 114, 4, 2236–2252.

Rouat, J., Liu, Y.C. & Morissette, D. (1997), A pitch determination and voiced/unvoiced decision algorithm for noisy speech, *Speech Communication* 21, 3, 191–207.

Schlang, M. (1989), An auditory based approach for echo suppression with modulation filtering, in *Proceedings of Eurospeech*, Paris, 661–664.

Schreiner, C.E. & Langner, G. (1988), Periodicity coding in the inferior collicus of the cat. II. Topographical organization, *Journal of Neurophysiology* 60, 6, 1823–1840.

Schroeder, M.R. (1968), Period histogram and product spectrum: New methods for fundamental-frequency measurement, *The Journal of the Acoustical Society of America* 43, 4, 829–834.

Schubert, E.D. & Wernick, J. (1969), Envelope versus microstructure in the fusion of dichotic signals, *The Journal of the Acoustical Society of America* 45, 6, 1525–1531.

Seneff, S. (1984), Pitch and spectral estimation of speech based on auditory synchrony model, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 9, 45–48.

Shamma, S.A., Vranic, S. & Wiser, P. (1992), Spectral gradient columns in primary auditory cortex: physiological and psychoacoustical correlates, in Y. Cazals, L. Demany & K. Homer (eds.) *Advances in the Biosciences*, volume 83, Oxford: Pergamon Press.

Shamsoddini, A. & Denbigh, P.N. (2001), A sound segregation algorithm for reverberant conditions, *Speech Communication* 33, 3, 179–196.

Shao, Y. & Wang, D. (2006), Model-based sequential organization in cochannel speech, *IEEE Transactions on Audio, Speech, and Language Processing* 14, 1, 289–298.

Shinn-Cunningham, B.G., Zurek, P.M. & Durlach, N.I. (1993), Adjustment and discrimination measurements of the precedence effect, *The Journal of the Acoustical Society of America* 93, 5, 2923–2932.

Slaney, M. & Lyon, R. (1990), A perceptual pitch detector, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, 357–360.

Slaney, M., Naar, D. & Lyon, R. (1994), Auditory model inversion for sound separation, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, 77–80.

Steeneken, H.J.M. & Houtgast, T. (1980), A physical method for measuring speech-transmission quality, *The Journal of the Acoustical Society of America* 67, 1, 318–326.

Stevens, S.S., Volkmann, J. & Newman, E.B. (1937), A scale for the measurement of the psychological magnitude pitch, *The Journal of the Acoustical Society of America* 8, 3, 185–190.

Suga, N. & Manabe, T. (1982), Neural basis of amplitude-spectrum representation in auditory cortex of mustached bat, *Journal of Neurophysiology* 47, 2, 225–255.

Summerfield, Q., Sidwell, A. & Nelson, T. (1987), Auditory enhancement of changes in spectral amplitude, *The Journal of the Acoustical Society of America* 81, 3, 700–708.

Thurlow, W.R. & Parks, T.E. (1961), Precedence–suppression effects for two click sources, *Perceptual & Motor Skills* 13, 7–12.

Tohyama, M., Lyon, R. & Koike, T. (1993), Source waveform recovery in a reverberant space by cepstrum dereverberation, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, 157–160.

Unoki, M. & Hosorogiya, T. (2007), Robust and accurate $F0$ estimation for reverberant speech by utilizing complex cepstrum analysis, *The Journal of the Acoustical Society of America* 122, 5, 3021.

Van Eeghem, J., Tohyama, M. & Koike, T. (1999), Blind dereverberation using short-time cepstrum frame subtraction, *The Journal of the Acoustical Society of America* 105, 2, 978.

Van Noorden, L. (1975), *Temporal coherence in the perception of tone sequences*, Ph.D. thesis, Institute of Perception Research, Eindhoven.

Wallach, H., Newman, E.B. & Rosenzweig, M.R. (1949), The precedence effect in sound localization, *The American Journal of Psychology* 62, 3, 315–336.

Wang, D. (2005), On ideal binary mask as the computational goal of auditory scene analysis, in P. Divenyi (ed.) *Speech separation by humans and machines*, Norwell, MA: Kluwer Academic, 181–197.

——— (2006), Feature-based speech segregation, in D. Wang & G.J. Brown (eds.) *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Hoboken, NJ: John Wiley & Sons, 81–114.

Wang, D. & Brown, G.J. (1999), Separation of speech from interfering sounds based on oscillatory correlation, *IEEE Transactions on Neural Networks* 10, 3, 684–697.

——— (2006), Fundamentals of computational auditory scene analysis, in D. Wang & G.J. Brown (eds.) *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Hoboken, NJ: John Wiley & Sons, 1–44.

Wang, K. & Shamma, S.A. (1995), Spectral shape analysis in the central auditory system, *IEEE Transactions on Speech and Audio Processing* 3, 5, 382–395.

Watkins, A.J. (1991), Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion, *The Journal of the Acoustical Society of America* 90, 6, 2942–2955.

——— (1999), The influence of early reflections on the identification and lateralization of vowels, *The Journal of the Acoustical Society of America* 106, 5, 2933–2944.

——— (2005), Perceptual compensation for effects of echo and of reverberation on speech identification, *Acta Acustica United With Acustica* 91, 892–901.

Weintraub, M. (1985), *A theory and computational model of auditory monaural sound separation*, Ph.D. thesis, Stanford University.

Winter, S., Sawada, H., Araki, S. & Makino, S. (2004), Overcomplete BSS for convolutive mixtures based on hierarchical clustering, in *Proceedings of the Fifth International Conference on Independent Component Analysis*, Granada, 652–660.

Wittkop, T., Albani, S., Hohmann, V., Peissig, J., Woods, W.S. & Kollmeier, B. (1997), Speech processing for hearing aids: noise reduction motivated by models of binaural interaction, *Acustica* 83, 684–699.

Woodruff, J. & Wang, D. (2010), Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization, *IEEE Transactions on Audio, Speech, and Language Processing* 18, 7, 1856–1866.

Wu, M. & Wang, D. (2003), A one-microphone algorithm for reverberant speech enhancement, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, 892–895.

Wu, M., Wang, D. & Brown, G.J. (2003), A multipitch tracking algorithm for noisy speech, *IEEE Transactions on Speech and Audio Processing* 11, 3, 229–241.

Yang, C., Soong, F.K. & Lee, T. (2007), Static and dynamic spectral features: Their noise robustness and optimal weights for ASR, *IEEE Transactions on Audio, Speech, and Language Processing* 15, 3, 1087–1097.

Yang, X. & Grantham, D.W. (1997a), Cross-spectral and temporal factors in the precedence effect: Discrimination suppression of the lag sound in free-fields, *The Journal of the Acoustical Society of America* 102, 5, 2973–2983.

——— (1997b), Echo suppression and discrimination suppression aspects of the precedence effect, *Perception & Psychophysics* 59, 7, 1108–1117.

Yegnanarayana, B. & Murthy, P. (2000), Enhancement of reverberant speech using LP residual signal, *IEEE Transactions on Speech and Audio Processing* 8, 3, 267–281.

Yost, W.A. & Guzman, S.J. (1996), Auditory processing of sound sources: Is there an echo in here?, *Current Directions in Psychological Science* 5, 4, 125–131.

Zurek, P.M. (1980), The precedence effect and its possible role in the avoidance of interaural ambiguities, *The Journal of the Acoustical Society of America* 67, 3, 952–964.

——— (1987), The precedence effect, in W.A. Yost & G. Gourevitch (eds.) *Directional Hearing*, New York: Springer-Verlag, 85–105.

Zurek, P.M., Freyman, R.L. & Balakrishnan, U. (2004), Auditory target detection in reverberation, *The Journal of the Acoustical Society of America* 115, 1609–1620.