# An investigation into improving speech intelligibility using binaural signal processing

Gary Anthony Spittle

PhD

Electronics

July 2009

# Abstract

This thesis sets out to improve the intelligibility of a target speech sound source when presented with simultaneous masking sounds. A summary of the human hearing system and methods for spatialising sounds is provided as background to the problem. A detailed review of relevant research in auditory masking, auditory continuity and speech intelligibility is discussed. Angular separation and sound object differentiation through amplitude modification are used to enhance a target speech sound. A novel method is developed for achieving this using only the binaural signals received at the ears of a listener. This new approach is termed an auditory lens. Psychoacoustic evaluation of the auditory lens processing has shown comparable intelligibility scores to direct spatialisation techniques which require prior knowledge of sound source spectral content and direction. The success of the auditory lens has led to a number of potential further research projects that will take the processing system closer to a real wearable product.

# Table of contents

# List of Figures

# List of Tables

## Acknowledgements

Firstly, I would like to thank Tony Tew for his continuous support and encouragement as my supervisor. I also wish to acknowledge the assistance of Marcelo Rodriguez and Omid Khorremy who helped run some of the listening experiments. I would also like to thank Ian Walker who was one of the people that convinced me to take on this challenge at the very beginning and has proven very useful when it came to the statistical side of things.

I am extremely grateful for the patience of my children, Ollie, Joe and Sam, who weren't there at the beginning, but are very much a part of this now. Finally, none of this would have been possible without the immense perseverance, motivation and support of my wife, Gill. This thesis is dedicated to her.

# Declaration

I declare that this thesis is entirely my own work and all contributions have been explicitly stated or referenced where appropriate.

# Chapter 1 Introduction

## *1.1 Background of the research*

Hearing aids of some form have been available for many years. The Eriksholm museum in Denmark has tracked the development of hearing aids through time. It is believed they started off as a crude trumpet shaped device in around 1800. Various shapes and sizes of hearing trumpet then followed ranging from the discreet to the large, depending on the severity of hearing loss. The first electric hearing aids were developed at the turn of the 20$^{th}$ century. These were very large and impractical and had reduced to "desktop" size by around 1930. Size remained an important issue with the goal being at least to make the devices portable, if not discreet. It wasn't until the mid 1950s that the first over the ear (OTE) devices were developed. These were significantly smaller and were discreetly positioned with the majority of the device behind the pinna. Size continued to reduce with the introduction of in the ear (ITE) devices, such that by the end of the 1980s in-the-ear-canal (ITEC) form factors were available. As their name suggests, these tiny devices can be placed in the ear canal and are totally hidden from view.

However, size and aesthetics are not the only considerations in a hearing aid. Hearing loss takes many forms and an aid which simply provides amplification, most suited to conductive hearing loss, can create difficulties with other types of deficit. Sensorineural hearing loss accounts for the majority of hearing problems and is due to damage in the inner ear or hearing nerve. For this form of loss hearing aids are the primary means of overcoming its effects.

With the advent of transistor-based hearing aids, the move towards more sophisticated signal processing began. Equalisation to match frequency selective loss could be provided along with compression to improve listening comfort for recruitment sufferers in the presence of loud sounds.

By the middle of the 1990s digital technology for hearing aids was beginning to become more accessible, despite a high price tag. Digital signal processing made possible a much more flexible device that could be finely tuned for each listener. A modern digital hearing aid will typically have algorithms for automatic gain control, noise reduction, equalisation and environment settings. These allow the hearing aid to adapt to the acoustic environment the user is in. Some devices also use directional microphones to assist with the isolation of a particular sound. We have come a long way since hearing aids indiscriminately amplified both wanted and interfering sounds. Yet, even the use of more advanced processing has often resulted in ultimate nonuse of the device (Kochkin, 2000). For example, in 1980, research in Finland revealed that 23% of people interviewed had stopped or almost stopped using their hearing aid after two years (Sorri *et al.*, 1984). More recently, a survey in 2008 showed a substantial improvement, with 13% no longer using their hearing aids after five years (Gimsing, 2008). According to a survey carried out by the Royal National Institute for Deaf People (RNID) in 2005 (RNID, 2005) there were approximately 9 million people in the UK with a hearing loss, more than two thirds of whom were over 60 years of age. Even nonuse by a small percentage of this group translates to a very large number of people who are in need of a hearing aid, but cannot find one which provides the benefits they seek.

Digital hearing aids, with their higher audio quality and better performance in noisy or acoustically difficult conditions seem to be addressing the situation, in that people are willing to pay in the region of £1000 to £2000 for a high quality device. Even so, for any hearing aid there comes a point where the acoustic environment becomes too challenging for it to help its wearer. A more recent RNID survey (RNID, 2007) revealed that even when wearing a hearing aid the most common issue reported is difficulty hearing in noisy situations.

Modern digital hearing aids use advances in microphone design and signal processing algorithms to assist with the reduction of interfering sounds. The auditory environment is constantly analysed to determine the characteristic of the background noise. Signals that are temporally and spectrally stationary, such as an air conditioning fan or road noise, tend to be of little interest and can be

removed. Speech, on the other hand, changes dynamically over time. Consequently, interfering speech sources are harder to attenuate, although a directional microphone that allows the hearing aid to focus to some extent on sounds in front of the user rather than to the side or behind may help.

Digital hearing aids are also benefiting from advances in other areas of technology, for example, the mobile phone. Due to the complexity of the mobile phone, convergence of the two technologies towards a single solution is still some way off. However, with the growing popularity of Bluetooth wireless devices, audio processed in a phone can be transferred wirelessly to a headset. This allows for the potential of distributed spatial processing (Roy and Vetterli, 2007). The advantages of doing this are not restricted to normal hearing users, as the integration of Bluetooth into hearing aids is seen as a key feature for the future (Oticon, 2008a). It will become practical to perform increasingly complex signal processing remote from the size-critical component fitted in or behind the ear, whilst maintaining power consumption at around 1mA and a battery life in excess of 100 hours (Oticon, 2008b). As mobile devices expand their functionality from simply making phone calls towards the storage and playback of music using 3D sound, the links between this and bilateral hearing aids (hearing assistance provided for both ears) become ever stronger. This blurring of the distinction between normal hearing and hearing impaired listeners is much to be welcomed.

## *1.2 Objective of the research*

The benefits of hearing with two ears compared to just one are well documented (e.g. Moore, 1997). They include a marked improvement in the ease with which speech can be understood in the presence of interfering sounds. This research is motivated by the desire to raise the intelligibility of speech in such conditions by exploiting the attributes of human binaural hearing. It is anticipated that as portable listening devices become more widely accepted, for example in the form of headsets, the way will be open to providing binaural hearing assistance for

anyone, with or without a hearing impairment, who needs to communicate in a noisy environment.

## 1.3 Statement of hypothesis

The hypothesis of this research is stated as:

*The intelligibility of a target speech source, in a binaural signal contaminated by spatially distinct interfering sounds, may be increased using an algorithm suitable for implementation in a hearing aid.*

The hypothesis depends on the veracity of several supplementary statements:

*It is possible to manipulate one sound source differently from another in a binaural mixture.*
The target speech source and the interfering sound sources must be distinguishable, even if not entirely separable, so that processing can be applied to them selectively.

*It is possible to improve the intelligibility of the target speech source without prior knowledge of the signals in the binaural mixture.*
It is assumed that the algorithm is supplied with the binaural mixture signal only, as would be the case when it is implemented as part of a portable or wearable device, such as a hearing aid.

*It is possible to develop an algorithm which exhibits a signal latency and computational burden commensurate with its use in a low power portable or wearable device, such as a hearing aid.*

Vision provides powerful cues for increasing the intelligibility of speech in noisy conditions. A delay in the acoustic signal reaching the listener may disrupt these cues. Preferably the delay should be as short as possible ideally remaining under

10ms to avoid the wearer hearing a delay in their own voice (Stone and Moore, 1999). Due to size and power constraints, devices such as portable music players, mobile phones and digital hearing aids have limited processing power. Both the limits on latency and computational burden place a requirement of relative simplicity on the algorithm to be developed.

## 1.4 Thesis structure

The thesis is organised into a series of chapters. Chapters 2 and 3 are of an introductory nature. The first of these starts with a detailed overview of the human hearing system. In particular, it considers the localisation cues used by the auditory system to determine the direction and distance of a sound source from a listener. Chapter 3 presents various methods for artificially spatialising sound sources. This includes surround sound systems and sound field reconstruction methods such as ambisonics. Head-related transfer functions are introduced along with the mathematical techniques needed for applying them to a mono sound to give the perceptual illusion of a spatialised sound.

Chapter 4 heads up three chapters that give an extensive review of the literature directly relevant to the goal of the investigation. It considers factors, such as auditory grouping, masking and spatial separation, which affect speech intelligibility in the presence of an interfering sound. Chapter 5 discusses the factors that influence the auditory continuity of a target sound in the presence of simultaneous interfering sounds. The continuity illusion is explained with reference to a variety of target sounds. Chapter 6 considers the influence of multiple interfering sound sources on the intelligibility of a target speech sound. In particular, the overlap of spectro-temporal components in the sounds is discussed.

The development of a practical binaural processing scheme for improving speech intelligibility in interference is presented in Chapter 7. The chapter begins with the respatialisation of a single sound source and migrates to an investigation of different respatialisation methods when multiple sound sources are involved. This leads to a novel approach based upon the concept of a binaural auditory

lens. The auditory lens algorithm is validated using a formal listening experiment, which is described in Chapter 8. The results of the experiment are analysed in Chapter 9. Overall conclusions and suggestions for further work are presented in Chapter 10.

# Chapter 2 Human hearing

## *2.1 Directions of sound*

Humans with normal hearing are able to hear sounds from all around them. In order to assist the description of hearing directions it is useful to refer to a defined set of planes and reference positions. These are shown in Figure 2-1. Directions are referenced as ($\theta$, $\phi$) representing azimuth and elevation respectively. Therefore, a direction of (0, 0) refers to a point at 0° elevation and 0° azimuth, which is located directly in front of a listener at ear level.

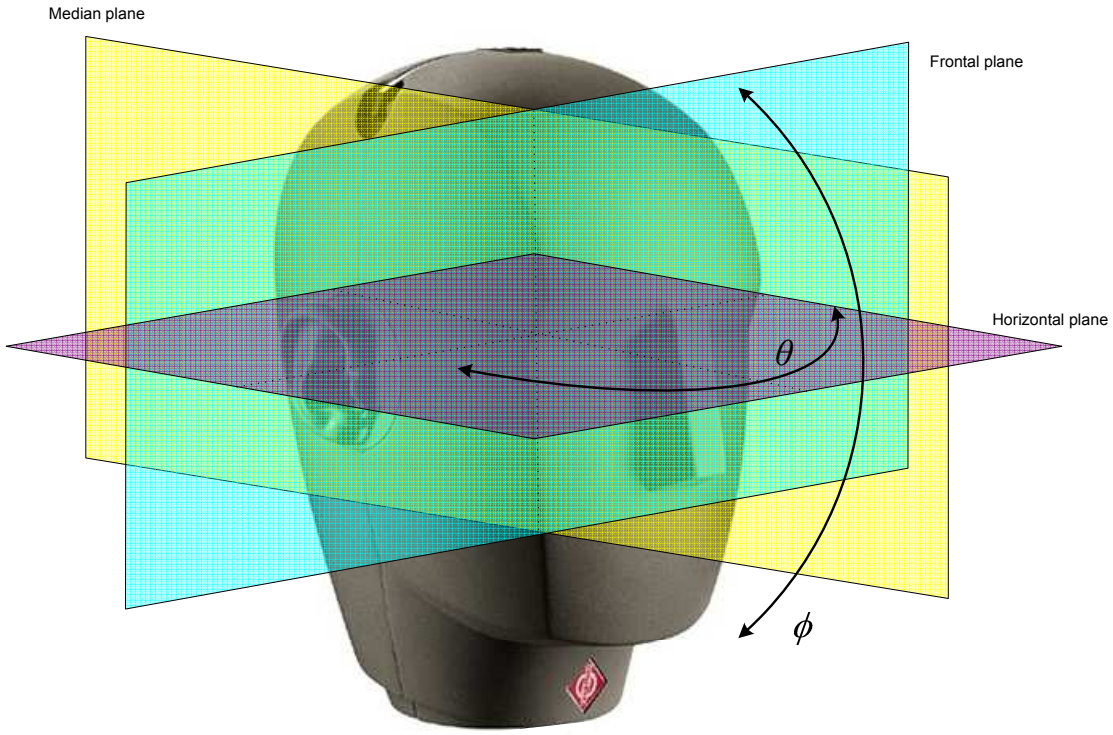



**Figure 2-1: Diagram illustrating the frontal, median and horizontal planes with reference to a head. Theta represents azimuth and represents elevation, with (0, 0) referenced as the point directly in front of a listener at the height of the interaural axis.**

## 2.2 Ear Physiology

The hearing system is split into two sections, the *peripheral* and the *internal* auditory systems. The intricate details of human hearing have been well documented by, for example Moore (1997). This section aims to provide a brief overview of the physical processes involved.

### 2.2.1 The peripheral auditory system

The peripheral auditory system is usually split into two parts, the *outer* and *middle ear*, the main components of which are shown in Figure 2-2.

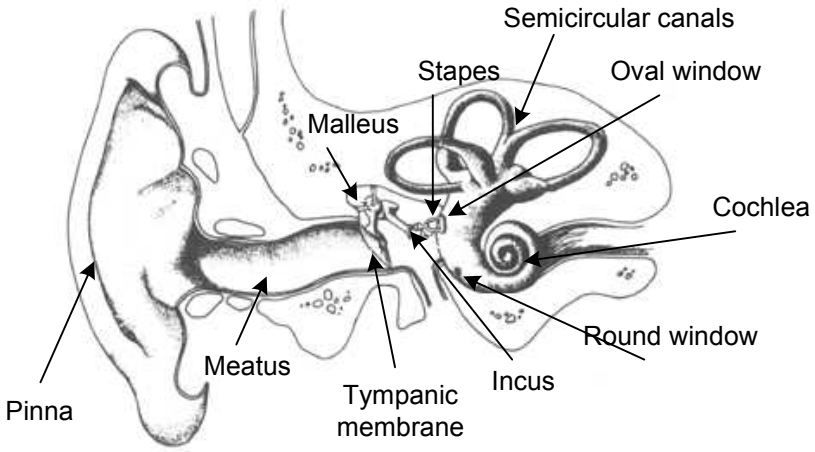



**Figure 2-2: The peripheral hearing system, adapted from Moore (1997 – p18).**

The outer ear is the part we see, and consists of the *pinna*, which comprises the folds of cartilage on the outside of the head, and the *ear canal* or *meatus*, which is the tube entering the head. An average ear canal is 25 mm long and has a diameter of 7-8 mm.

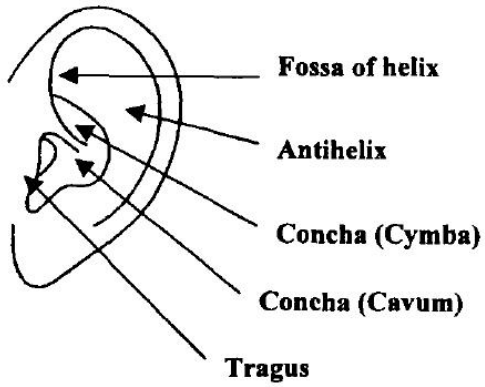



**Figure 2-3: The regions of the pinna, taken from Tan and Gan (2000)**

The pinna acts as an auditory filter and provides essential cues for localising sound sources. The regions of the pinna are shown in Figure 2-3. It directs the sound waves into the ear canal towards the *eardrum*, or *tympanic membrane*. The eardrum is oval in shape. It is 10-11 mm along its longest dimension and 8.5 - 9 mm along its shortest, and it is about 0.1 mm thick. The sound waves cause the eardrum to vibrate. The vibrations are transferred through the middle ear to the *oval window* by the *ossicles*. These are the three smallest bones in the body and are known as the *hammer*, *anvil* and *stirrup*, or *malleus*, *incus* and *stapes*. The middle ear converts the sound from air pressure differences to waves in the fluid of the *cochlea* and acts as an impedance matching device between the two.

### 2.2.2 The internal auditory system

The cochlea is probably the most complex part of the inner ear. It is a fluid-filled spiral, with two chambers running along its length. The two chambers, the *scala vestibuli* and *scala tympani*, are separated by the *Reissner* and *basilar membranes*, and are connected by a small hole at the apex end of the cochlea called the *helicotrema*. This region is filled with a fluid called *perilymph*. The base end of the cochlea has two windows, the previously mentioned oval window and the *round window*. This is shown in Figure 2-4.

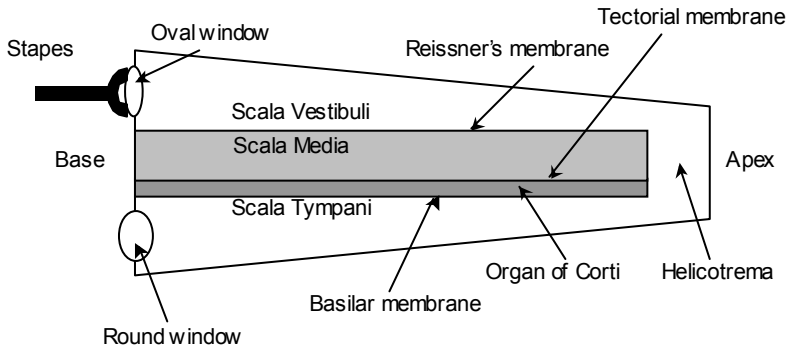



**Figure 2-4: Simplified diagram of an unravelled cochlea.**

As the stapes presses against the oval window, the fluid is set in motion thus causing a wave to travel along the basilar membrane and the round window to be pushed outwards. The basilar membrane moves in sympathy with the intensity and frequency of the signals presented at the oval window. The basilar membrane is stiff and narrow at the base and wider and flexible at the apex. Due to its structure, low frequencies cause maximum displacement of the basilar membrane at the apex and high frequencies at the base.

The *tectorial membrane* bisects the area between the Reissner and basilar membranes. The region between the Reissner and tectorial membranes is known as the *scala media*, and contains a fluid called *endolymph*. The region between the tectorial and basilar membranes contains the *organ of Corti*, which primarily consists of *outer* and *inner hair cells*. It is the inner hair cells that transform the physical movement of the basilar membrane into neural activity, which is then passed on to the brain.

## 2.3 Masking

The structure of the hearing system gives rise to a phenomenon known as *masking*. The cochlea responds to vibrations at the oval window by displacing the basilar membrane as described above. For a sine wave, the maximum displacement occurs at the position along the basilar membrane that corresponds to the frequency of the sine wave. However, neighbouring areas corresponding

to other frequencies are also displaced, such that there is a region of movement rather than a single narrow section. The neural firings to the brain for the maximum displacement cause us to hear a single frequency. However, this causes other low amplitude signals to be hidden under the region of movements caused by a louder signal at a nearby frequency. This is known as *simultaneous masking* and is illustrated below in Figure 2-5.

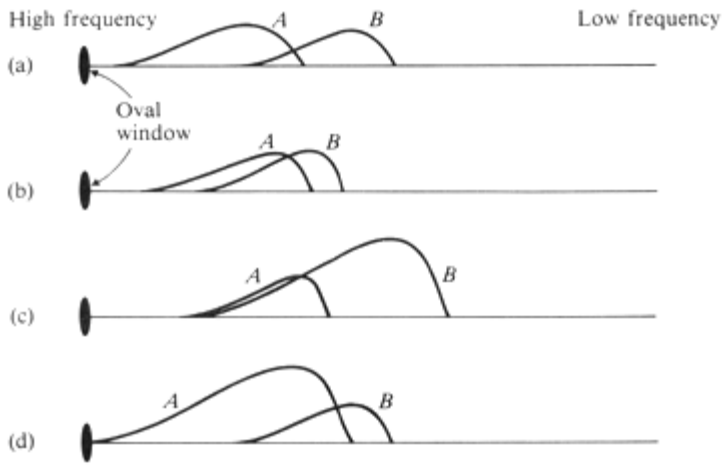



**Figure 2-5: Basilar membrane displacement for two tones A and B, tone A has a higher frequency than tone B. (a) the tones barely overlap, (b) tone B masks A, more than A masks B, (c) B almost totally masks A, (d) A partially masks B. Taken from Rossing (1990)**

The masking of low-level signals is not confined to sounds presented simultaneously; there exists another type of masking called *temporal masking*. Experiments have shown that we suffer from *forward* and *backward masking*, see for example Moore (1993). Forward masking occurs when we fail to hear a low-level sound if a much louder sound ends just before the target sound starts. This can typically occur with a gap of up to 20-30 ms between the sounds (Rossing, 1990). Backward masking is even more extraordinary in that we can fail to hear a low level sound if it is closely *followed* by a much louder sound that starts up to 10 ms later (Rossing, 1990). Moore (1997) offers three reasons why forward masking may occur:

1. The response of the basilar membrane to the loud masking signal continues for some time after the signal has ended. This is known as

ringing and would prevent a quieter signal closely following the masker from being heard. It is more noticeable at low frequencies, which have a longer ringing duration.

2. The masker produces short term fatigue in the auditory nerve which reduces the response to low-level signals after the masker.
3. The masker persists at a higher level in the auditory system after the signal has ended.

Backward masking occurs due to the louder masking sound propagating more quickly through the hearing system and overtaking the preceding lower-level signal (Rossing, 1990).

A feature of the hearing system that becomes evident through masking is the concept of the *critical band*. Moore (1997) provides a detailed discussion on the experiments which have investigated this phenomenon. The audible threshold of a tone is measured when masked by a narrowband noise centred at the frequency of the tone. It has been shown that increasing the bandwidth of the noise masker increases its ability to mask the tone. However, there comes a point where further increases in bandwidth do not cause additional masking. The auditory system acts as though it contains a bank of bandpass filters with overlapping pass-bands. These are known as the *auditory filters*. It is thought that, when listening to a target signal in a noisy background, we only make use of the auditory filter that has a centre frequency nearest to the target signal. Therefore, as the bandwidth of the noise increases, more noise will pass through the auditory filter. Once the bandwidth of the noise exceeds the width of the filter, no additional masking of the tone within that filter occurs.

This section only provides an overview of the types of auditory masking that occur. The components of the masking phenomenon that are essential to this research are considered in more detail in Chapter 4.

## *2.4 Binaural hearing*

The main components of the hearing system have been described in relation to the operation of a single ear. The binaural processes (that is, those involving both ears) that provide us with information about our auditory environment will now be discussed.

### 2.4.1 Sound localisation

The ability of humans to determine the location of a sound source has been well researched. The reader is directed towards Blauert (1997) for a detailed account of the processes involved. The general concept of sound localisation is based upon the analysis and comparison of the signals arriving at the two ears. As previously discussed, the brain is presented with temporal and spectral information about the signals. There is no direct information concerning the spatial location of a sound source, unlike vision, where the spatial location of an image is translated onto a specific location on the retina. The location of a sound source can be split into two components, its direction and distance. The auditory cues used to determine these components are discussed next.

### 2.4.2 Directional cues

Several acoustic cues are used in determining the direction of a sound source. For a binaural hearing system, these cues are:

- Interaural time difference (Section 2.4.2.1)
- Interaural intensity difference (Section 2.4.2.3)
- Spectral alteration (Section 2.4.2.4)

For a monaural hearing system there are no interaural differences. It is still possible, however, to localise a sound source to some degree of accuracy using a single ear, by using just the spectral information at that ear.

### 2.4.2.1 Interaural time difference

The fact that humans have two ears allows us to detect differences in the sounds arriving at each ear. In air, the pressure changes which make up sound travel at a near-constant speed of approximately 340 m/s. Figure 2-6 describes the head as a circle of radius $r$. If a sound $S$ originates from a point source outside the median plane in direction $\theta$, the signal will arrive at one ear before the other, thus providing an interaural time difference (ITD). The ear nearest to a sound source is referred to as *ipsilateral* and the other ear as *contralateral*. If the distance of the sound source from the listener is much greater than the radius of the head, the waves arriving at the listener can be approximated as being parallel.

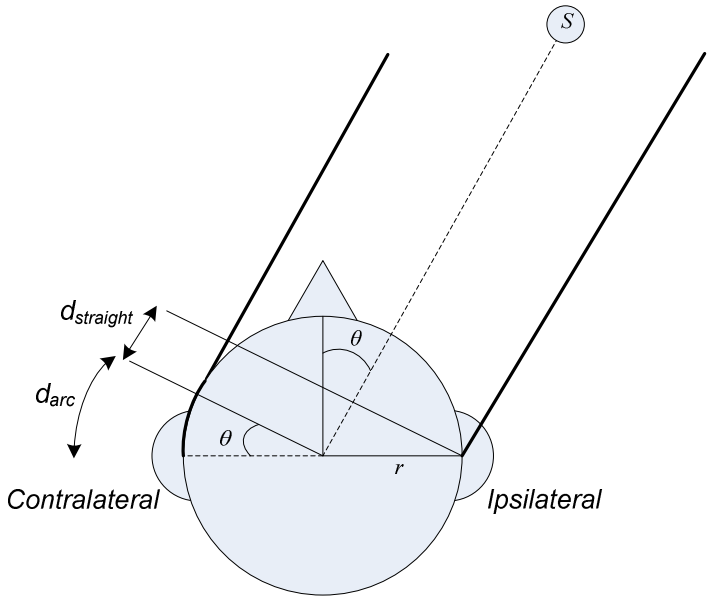



**Figure 2-6: The path difference between the near (*ipsilateral*) and far (*contralateral*) ears for a distant sound source *S* to one side of a listener.**

Figure 2-6 shows that the total path difference consists of the arc $d_{arc}$ and the line $d_{\text{straight}}$ or:

$$d_{\text{total}} = d_{\text{arc}} + d_{\text{straight}}$$ **Eq. 2-1**

where

$$d_{\text{arc}} = r\theta$$ **Eq. 2-2**

and

$$d_{\text{straight}} = r.\sin\theta$$ **Eq. 2-3**

which gives

$$d_{\text{total}} = r(\theta + \sin\theta)$$ **Eq. 2-4**

By converting the path difference into a difference in time of arrival $\Delta T$ of a sound at each ear, we obtain the popular Woodworth formula (Woodworth and Schlosberg, 1962):

$$\Delta T(\theta) \approx \frac{r(\theta + \sin\theta)}{c}$$ **Eq. 2-5**

where $c$ is the speed of sound in air.

For example, if a sound source lies at $\theta = 90°$ azimuth in the horizontal plane for a listener with a head diameter of 18 cm, the approximate difference in time of arrival at the ears can be calculated as:

$$d_{\text{total}} = 0.09\left[\frac{\pi}{2} + \sin\left(\frac{\pi}{2}\right)\right]$$ **Eq. 2-6**

$$d_{\text{total}} = 0.231\ m$$

$$\text{ITD} \approx 680\ \mu s$$

When a sinusoidal signal is considered, the ITD is more specifically detected as a phase difference between the left and the right signals. This method of direction detection is only applicable to signals with a half wavelength $\lambda/2$ greater than the path length between the ears. Otherwise ambiguities may arise in multiples of $2\pi$ over the exact phase difference. This is shown in Figure 2-7.

Using the calculations in Equation 2-6 the maximum unambiguous frequency limit for ITD, $f_{\text{ITD}}$, is given as:

$$f_{\text{ITD}} = \frac{c}{\lambda} \qquad \textbf{Eq. 2-7}$$

Therefore, for a head radius of 9 cm, ITD is only applicable for relatively low frequencies up to approximately 1.5 kHz. However, the difference in the time of arrival of the temporal envelope of the signal is used at higher frequencies, and this is discussed in Section 2.4.2.2.

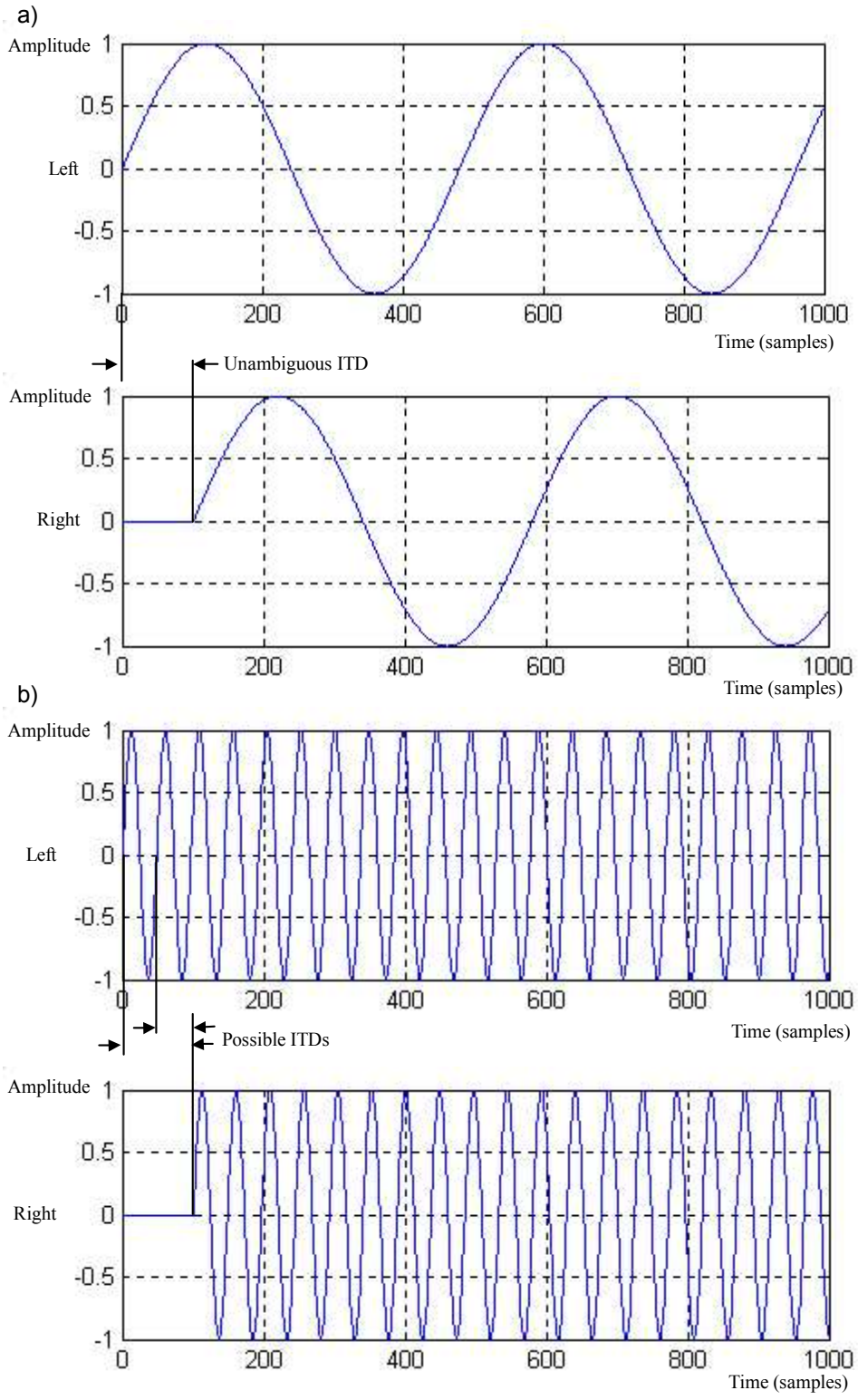



**Figure 2-7: Signal phase differences between the ears, (a) with unambiguous ITD, (b) with two possibilities for the ITD.**

Also, there exists an infinite set of imaginary surfaces, on each of which the ITD is constant. These surfaces are approximately conical and are given the term

*cones of confusion*. One such cone is shown in Figure 2-8. Over this surface ITD can provide no information with which to determine source direction further. To resolve this ambiguity the hearing system requires additional directional information. This will be described in subsequent sections.

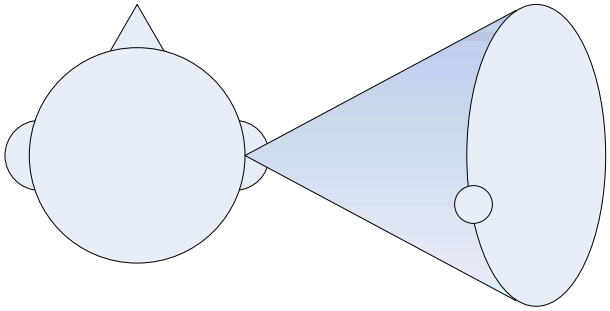


**Figure 2-8: Conceptualisation of a cone of confusion. All points on the surface of the cone have approximately the same ITD. Points on circular cross sections of the cone have approximately the same IID.**

Kuhn (1977) carried out numerous measurements that show how ITD varies with frequency. Theory indicates that this is due to the varying interaction between the head and the wavefront, with diffraction occurring at low frequencies and creeping waves forming around the head at high frequencies. Kuhn shows that the variation in measured ITD values can be expressed using a non-dimensional parameter, Π, by using a normalising factor given in Eq. 2-8. That is, the normalising factor can be used to approximate an ITD as shown in Eq. 2-9.

$$\mathrm{ITD}_{\mathrm{norm}} = \left(\frac{r}{c}\right)\sin\theta_{inc} \qquad \textbf{Eq. 2-8}$$

$$\mathrm{ITD} = \prod \mathrm{ITD}_{\mathrm{norm}} \qquad \textbf{Eq. 2-9}$$

Where, $r$ is the radius of the head, $c$ is the speed of sound and $\theta_{inc}$ is the angle of incidence. Conclusions drawn from the theoretical and measured results include:

- The ITD is frequency independent below approximately 500 Hz and can be reduced to $\Pi_L = 3$.

- The ITD is frequency independent above approximately 3 kHz and can be reduced to $\Pi_H = 2$.
- The ITD has a minimum between 1.4 kHz and 1.6 kHz for angles of incidence $\theta \leq 60°$.

For example, for an angle of incidence of 30°, a head radius of 9 cm and a frequency of 300 Hz, Equation 2-9 gives an ITD of 400 μs.

#### 2.4.2.2 Interaural envelope difference

As mentioned in Section 2.4.2.1, for signals that do not contain any frequencies below about 1.6 kHz it is still possible for the human hearing system to make use of interaural time differences. The auditory system ignores the interaural delay of the fine structure of the signals, and uses the interaural envelope difference (IED), Blauert (1997) and Henning (1974).

Zurek (1993) came to the following conclusions concerning the use of IEDs:

- At low frequencies (less than about 1500 Hz), the interaural carrier delay is dominant.
- At high frequencies (above about 1500 Hz) the interaural envelope delay is dominant.
- When whole-waveform interaural delays are presented simultaneously in both low and high frequency regions, the contribution from the low frequencies dominates.

He suggests therefore that it is the low frequency interaural carrier delay that dominates over high frequency envelope delay. This dominance is not absolute, as it is determined by the envelope shape, confirmed by Blauert (1997) in his summary of experiments involving IEDs at low frequencies.

However, Middlebrooks and Green (1990) found that the IEDs do not provide a significant contribution to localisation judgments for high frequencies. They

suggest that this is because other localisation cues will be used if they are available in preference to the IED cue. Furthermore, in a typical free-field listening scenario, the average modulation depth is too small or the modulation frequency is too high for the IED to be utilised.

#### 2.4.2.3 Interaural intensity difference

In addition to detecting time or phase differences, the human auditory system can detect intensity differences between the signals at each ear. The *interaural intensity difference* (IID) is more perceptible for high frequency signals. This is because the head acts as an obstacle and increasingly blocks signals as their wavelength gets shorter, whereas signals with a longer wavelength than the diameter of the head tend to diffract around it. This is illustrated in Figure 2-9.

Using the formula from Eq. 2-10 it is shown that the frequency, $f_{\mathrm{IID}}$, at which IIDs become effective for detecting the direction of a sound, is approximately 1.9 kHz, using an average head diameter of 18 cm as the minimum wavelength $\lambda$.

$$f_{\mathrm{IID}} = \frac{c}{\lambda} = \frac{340}{0.18} \approx 1.9 \text{ kHz} \qquad \textbf{Eq. 2-10}$$

Therefore, it is only mid to high frequencies that are affected by the acoustic shadowing effects of the head.

There is not a sharp transition between using ITD or IID cues within the human auditory system. Work by Kuhn (1977) (see Section 2.4.2.1) has shown that there is a minimum value in ITD between 1.4 kHz and 1.6 kHz, which renders its use as a localisation cue as quite poor. In addition, the IID cue is still weak below about 2 to 3 kHz. Therefore, localisation is generally poor between approximately 1.4 kHz and 3 kHz, over which range the auditory system switches from using ITD to IID, (Hartmann, 1999). In general, for wideband signals, the ITD tends to be the dominant cue (Macpherson and Middlebrooks (2002) and Wightman and Kistler (1992)).

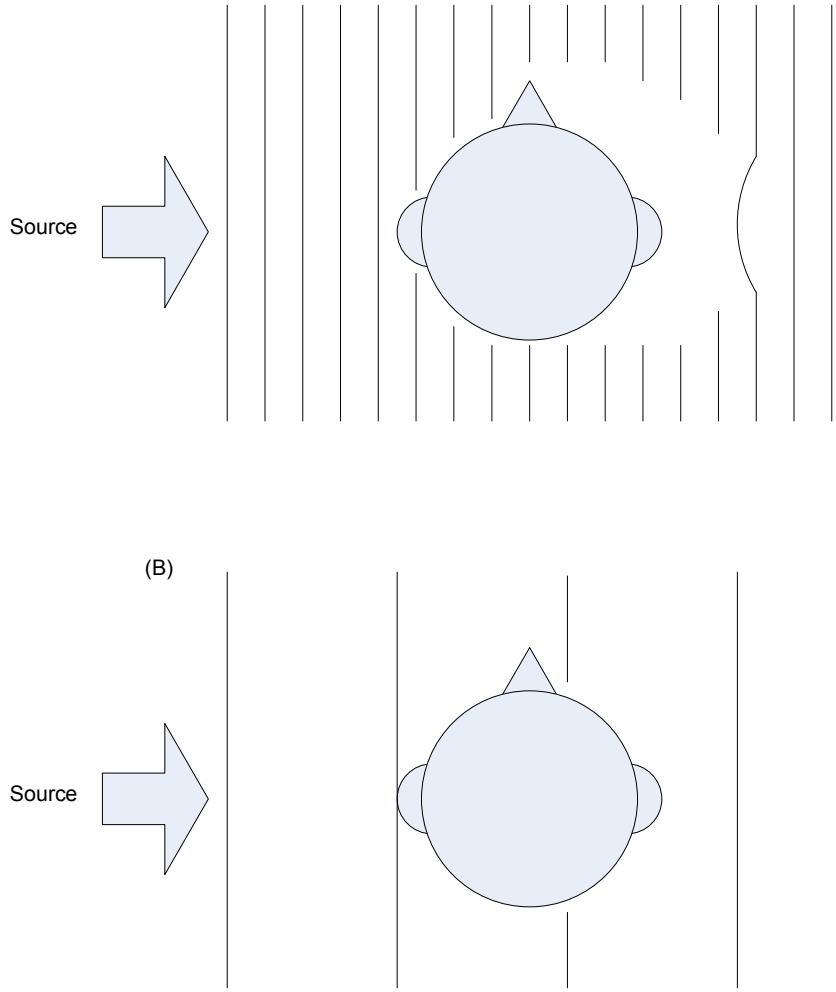



**Figure 2-9: Pictorial representation of the effect on intensity difference due to (a) head shadowing at high frequencies and (b) diffraction at low frequencies.**

An important property of the IID is its contribution to resolving the cone of confusion. As will be discussed in more detail in Section 2.4.3.2, the IID also provides a distance cue for close sources, which allows the cone to be reduced to a single circle of possible locations.

The IID cue is frequency dependent, which is mostly due to decreasing diffraction around the head with increasing frequency. Results from Middlebrooks *et al*. (1989) indicate that at high frequencies the IID varies with both azimuth and elevation of the sound source. They also found that at some frequencies there are asymmetries between the two ears for particular subjects which provide substantial IIDs for sounds located on the median plane.

Generally, however, the IID is approximately constant for sound sources that are equidistant from each ear.

When considering the IID variance across listeners it is necessary to take into account the absolute loudness levels at each ear and how they can be modeled. Moore and Glasberg (2007) describe a model for determining binaural loudness. They have found that the diotic to monaural loudness ratio is less than 2. This is due to a loud input at one ear inhibiting the internal response to a weaker sound at the other ear, which is referred to as "contralateral binaural inhibition". The level difference required for equal loudness of diotic and monaural sounds has been calculated by Moore and Glasberg and the results are shown in Figure 2-10.

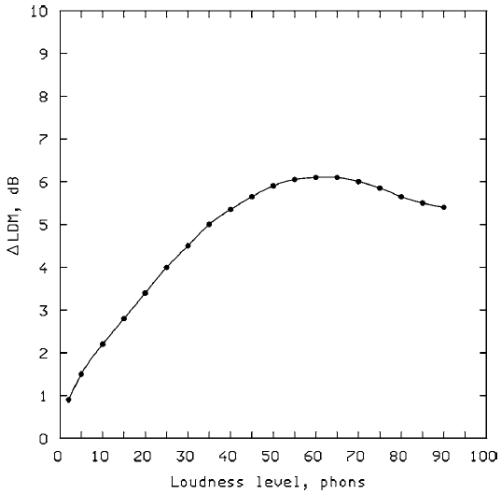



**Figure 2-10: The level difference ΔLDM required for equal loudness of a diotic and monaural 1 kHz tone. ΔLDM is plotted as a function of the level of the diotic tone in phons, which is equal at 1 kHz to its level in $dB_{SPL}$. Extracted from Moore and Glasberg (2007).**

The results show that the ratio is not 2 and that it is non-linear. It varies with loudness level with a peak at around 60-70 dB SPL. Moore and Glasberg (2007) provide an example of how the data are obtained; "…a diotic tone at 40 dB SPL gave a loudness level of 40 phons, and a monaural tone of 45.3 dB was required to give equal loudness."

Binaural loudness, therefore, is not a simple doubling of monaural loudness and is a function of signal level. In fact, work by Sivonen (2007) has shown that

human perception of loudness becomes more complicated when direction of arrival and reverberation are considered. The wide variation across the subject's used in listening experiments makes it difficult to produce a generic model of binaural loudness.

#### 2.4.2.4 Spectral cues

In addition to the temporal and wideband amplitude differences between the signals arriving at the ears, the auditory system is also provided with complex spectral information to aid localisation. The acoustic signals are not fed to the eardrum along straight tubes and human heads are not perfect spheres suspended in space, as is sometimes assumed in simple theoretical models of binaural hearing. If this were the case, the brain would be presented with signals that have undergone relatively simple and symmetrical spectral colouration. However, before reaching the eardrum real signals are filtered by obstacles in their path, such as our bodies, head, facial features and the elaborate shape of the pinnae. This section discusses the influence of these spectral cues and how they are used to disambiguate the cues provided by binaural signal differences alone.

Blauert (1997) in his extensive work on spatial hearing emphasises the importance of the pinna in localising sounds. The pinna distorts the incident sound signals depending on the location of the sound source. Due to the shape and size of the cavities within the pinna (see Figure 2-3) it has a greater impact on signals with shorter wavelengths corresponding to frequencies mostly above 6 kHz (Moore, 1997, p. 229). Despite this restricted bandwidth of effectiveness, the significance of the role of the pinna in sound source localisation is confirmed in experiments carried out by Musicant and Butler (1984). They found that the ability of listeners to localise 4 kHz highpass-filtered noise was significantly degraded when the pinna cues were partially removed by occlusion of the external ears. However, 4 kHz lowpass-filtered noise was not further degraded when the pinna cues were removed, thus proving that the pinna cues are mostly created by frequencies above 4 kHz.

Batteau (1967) has modelled the effects of the pinna. The cavities of the pinna can be simulated by using delay lines and reflection coefficient attenuators. An example is given in Figure 2-11, taken from Wright *et al*. (1974).

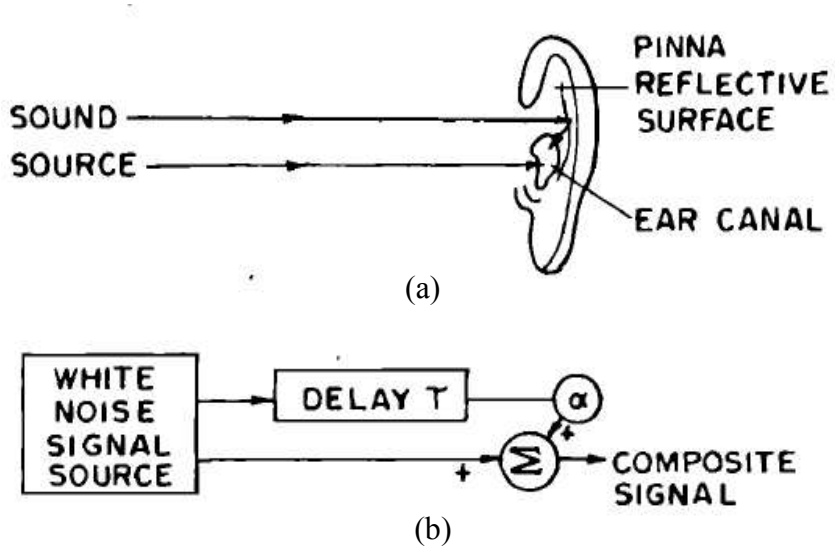



**Figure 2-11: (a) an example of a pinna reflection, showing a direct and indirect path from a remote sound source to the ear canal. (b) illustrates a model of the reflection using a white noise source with a delay line and attenuation factor, taken from Wright *et al.* (1974).**

The effect of this reflection is to cause peaks and troughs in the frequency response due to constructive and destructive interference. The general effect of this is termed the *pinna notch*. In practice there will be many reflections combined to produce a complex spectral mapping based on azimuth and elevation. Teranishi and Shaw (1968) have modelled the external properties of the human ear and found important resonances used for sound localisation. Due to the dimensions of the pinna, the prominent spectral notches occur at frequencies above 4-5 kHz (Raykar *et al*., 2005).

Butler and Humanski (1992) describe how it was widely believed that spectral cues provided by the pinna were largely responsible for localising sound in the vertical plane and that binaural difference cues were responsible for localising sound in the horizontal plane. In previous experiments, Humanski and Butler (1988) demonstrated that the localisation of a sound in the vertical plane is as

accurate with just the ipsilateral ear as it was with both ears. Their results imply that monaural spectral cues influence location judgements in the vertical plane. In more recent work (Butler and Humanski, 1992), they set out to investigate how well listeners could localise lowpass-filtered noise in the vertical plane (i.e. with the pinna cues unavailable). They found that listeners were unable to localise lowpass-filtered sounds monaurally in the vertical plane. However, subjects could localise the sounds binaurally, suggesting that only binaural temporal and intensity difference cues were used at these frequencies. They conclude that for sounds to be fully localised in the vertical plane, time and intensity difference cues must be available in addition to pinna cues.

However, it is not only the pinnae that provide spectral cues for the location of a sound source. Algazi *et al*. (2001) found that listeners were able determine the elevation of random noise bursts, lowpass-filtered at 3 kHz, away from the median plane. The intention of the research was to determine whether the removal of the pinna cues affects localisation in elevation, since spectral changes due to the pinna were considered to be the primary source of elevation information. Analysis of the HRTFs for the listeners revealed low frequency elevation-dependent features, which were attributed to head diffraction and torso reflections. Previous work reported by Avendano *et al*. (1999a) found that there are important low frequency binaural elevation cues due to multi-path head-diffraction effects and shoulder and torso reflections. They suggest that this may explain why many binaural recordings made with a dummy head but no torso produce elevated images. They go on to propose that including the proper torso cues should bring sound images down to their correct elevation.

Spectral cues also contribute to resolving the ambiguities described by the cone of confusion by supplying elevation information and whether a source is in front or behind the vertical plane. These elevation and front-back discrimination cues reduce the circle of possible locations from the ITD and IID cues to a single point (Shinn-Cunningham *et al.,* 2000a). Blauert (1997) suggests that it is possibly the orientation and angle of protrusion of the shell-like structure of the pinnae that allows us to resolve front-back ambiguities, as high frequencies will tend to be attenuated more for sources in the rear than for sources in the front.

### 2.4.3 Distance cues

The main acoustic cues used to determine the distance of a sound source from a listener are listed below and will be described in the following sections:

- Overall signal level (Section 2.4.3.1)
- Interaural intensity differences (Section 2.4.3.2)
- Spectral attenuation (Section 2.4.3.3)
- Direct to reverberant ratio (Section 2.4.3.4)

Shinn-Cunningham (2000b) investigated and summarised the dominant distance cues for sound sources within 1 m of the listener's head. She, along with others, (Duda and Martens, 1998; Brungart and Rabinowitz, 1999b; Shinn-Cunningham *et al*., 2000a), has shown that overall signal level and IID are particularly important for sound sources close to the listener. This is explained in more detail in the following subsections.

#### 2.4.3.1 Overall signal level

The overall signal level change for nearby sources does not simply follow the inverse square rule, but also depends on source direction (Shinn-Cunningham, 2000b). The level of sounds in the median plane changes relatively slowly with respect to distance, compared with distance on the interaural axis. Her results indicate that the overall sound level at the ears can produce distance information. However, unless the listener has prior knowledge of the source level, only relative distance information is available. This is because the level of sound reaching the ear varies with distance and with the level of the sound source itself and it is difficult to distinguish between the two. For sound sources that are further than 1 m away from the listener, the overall sound level tends to follow the inverse square law, and is independent of direction.

### 2.4.3.2 Changes in interaural intensity differences

The IIDs provide directional spatial information for high frequencies, as described in Section 2.4.2.3. Brungart (1998), and Chang and Tan (1998), have shown that for nearby sources in virtual auditory displays, it is the distance-dependent changes to the IID that provide absolute distance cues. In addition, for lateral sound sources located within 1 m from the head of a listener, a reduction in source distance causes a rise in IID for all frequencies, (Duda and Martens, 1998; Huopaniemi and Riederer, 1998). The increase in IID is substantial for lateral sound sources (Brungart and Rabinowitz, 1999b). This change in IID magnitude with horizontal position suggests our ability to determine source distance increases as source azimuth increases, which is confirmed by Brungart *et al*. (1999c). Shinn-Cunningham (2000b) splits the total IID into two components. The first is the traditional direction-dependent, frequency-dependent and distance-independent head-shadow component. The second is the distance-dependent, direction-dependent and frequency-independent component. It is shown that the IID distance cue is more robust than the overall level cue, as the IID only depends on source location and is independent of the level of the sound at the source. The dominance of the IID distance cue is confirmed by Brungart (1999d).

### 2.4.3.3 Spectral attenuation

Another, less influential, cue has been investigated by Coleman (1968), who has shown that a change in distance produces spectral changes in a free field sound source. Experiments were carried out that demonstrated how the perceived distance of a sound from a listener could be altered by manipulating the spectral content. More recent work has been published by Little *et al*. (1992). They found that high frequency sounds tend to be heard as closer to the listener than low frequency sounds. This is based on the fact that high frequency sound components lose more energy compared to low frequency ones when travelling through air. It can therefore be concluded that this form of spectral change provides a cue for the distance of the source from the listener. It is presumed that the spectral change will only provide a relative distance cue, and not an absolute

one, which would have to be based on prior knowledge of the source signal spectrum.

#### 2.4.3.4 Direct-to-reverberant ratio

The results from Shinn-Cunningham's work (Shinn-Cunningham, 2000b) clearly show that the direct-to-reverberant pressure ratio at the ipsilateral ear is a powerful distance cue for sources to the side of the head. For a sound source at an azimuth of 90º and a distance of 1 m compared to 0.15 m, the direct sound level in dBs increases linearly by more than 20 dB relative to the reverberant sound level. The levels at the contralateral ear, for the same source positions, change by only 8 dB. Landone and Sandler (1998a) suggest that the ratio between direct and reverberant sounds provide a cue for absolute source distance judgement.

### 2.4.4 Binaural advantage

If target and masker signals are presented *diotically*, i.e. with the same signals at each ear, we have the traditional experimental arrangement for masking tests, as discussed in Section 2.3. If the target is presented *dichotically*, i.e. different sounds are presented to each ear, different spatial locations can be simulated for the masker and maskee. The influence of the masker will now be different at each ear. When listening to spatially separated sounds, the human auditory system is able to exploit phase, level and spectral differences to its advantage when identifying the sound components. The increase in masker signal level necessary to keep a target signal at the masking threshold when it is presented dichotically instead of diotically is known as the *binaural masking level difference* (BMLD). More generally, the ability to distinguish between sound sources at different locations, using two ears as opposed to one, is known as the *binaural advantage*.

Moore provides a table of results for masking level differences (MLDs) using a variety of broadband maskers and low frequency target signals.

| Interaural Condition | MLD in dB |
|---|---|
| $N_uS_\pi$ | 3 |
| $N_uS_0$ | 4 |
| $N_\pi S_m$ | 6 |
| $N_0S_m$ | 9 |
| $N_\pi S_0$ | 13 |
| $N_0S_\pi$ | 15 |

**Table 2-1: Binaural masking level differences taken from Moore (1997)**

$N_0$ denotes that the phase of the noise is the same at the two ears, this is known as *homophasic*. $N_\pi$ denotes that the noise at each ear is π radians out of phase, or *antiphasic*. $N_u$ denotes that the noise is uncorrelated at each ear. $S_m$ denotes that the target signal is presented monaurally (i.e. to one ear only). The results shown are relative to the $N_0S_0$ or diotic condition. For the target and masking signals described there is an obvious advantage to having the target signal inverted in one ear, which corresponds loosely to it being located to one side of the listener, with the noise masker lying in the median plane.

## 2.5 Summary

This chapter has outlined the workings of the outer, middle and inner ear of the human hearing system, including how the peripheral hearing components enable a listener to distinguish between sounds from different locations.

Interaural time difference (ITD) is used to determine the spatial location of a sound source, in particular for sound sources on the horizontal plane. ITD is most effective at frequencies below approximately 1500 Hz. For higher frequencies, interaural envelope difference (IED) is used if the listening scenario does not also provide the stronger low frequency ITD cue. The detection and manipulation of ITD forms a core part of the technical work for the research described in this thesis in Chapter 7.

IID cues help to determine the spatial location of a sound source. IID is most effective for frequencies above approximately 1.9 kHz. It is frequency dependent and is considered to be a secondary cue to ITD when both cues are available to a listener. It has also been shown that the absolute loudness of sounds, when presented binaurally, does not follow a simple linear model.

Spectral changes that are imposed on the sound signals entering the ear canal due to the torso, head and pinna shape also provide localisation cues. The auditory system relies on their spectral signature to resolve the ambiguities created using ITD and IID cues alone.

The principle distance cues have been discussed. These include amplitude, spectral intensity difference and reverberation. Used together, a listener can roughly estimate the distance of a sound source.

Finally, a glimpse into the benefit of listening with two ears compared to one, in terms of masking, has been introduced in the form of the binaural advantage. This short overview provides the terminology used throughout the rest of this thesis and provides a physiological basis for what follows.

# Chapter 3 Sound spatialisation

Chapter 2 covered the cues and processes required for determining the location of the auditory signals presented to a listener for free field sounds - a process known as localisation. This section describes the complementary processing required for spatialisation - the artificial reproduction of spatialised sounds. The methods involved tend to fall into one of two categories, sound field reconstruction or binaural synthesis.

## *3.1 Sound field reconstruction methods*

The simplest method of producing spatialised sounds is to use multiple loudspeakers positioned around a listener. This technique ranges from using two loudspeakers in front of a listener to hundreds of loudspeakers. Both methods have the potential for completely surrounding a listener with virtual sound sources.

### 3.1.1 Stereo

The most familiar configuration of multiple loudspeakers uses just two. This is the two-channel stereo pair used in the majority of home hi-fi systems. The recommended positioning of the two speakers is shown in Figure 3-1.

A virtual sound point source can be positioned between the loudspeakers by adjusting the relative amplitude of the signal delivered to the left and right channels. This is known as panoramic positioning, or panning. The majority of modern audio mixing consoles have panoramic potentiometers for performing this operation.

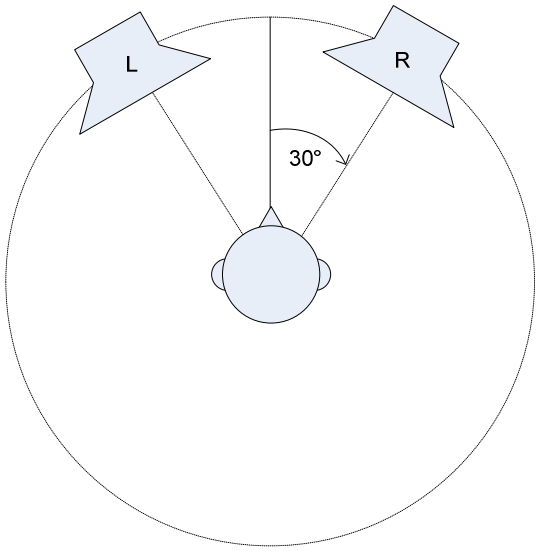



**Figure 3-1: The recommended loudspeaker placement for a stereo configuration.**

Another stereo positioning process allows the sound source to be perceived as being outside the space between the two loudspeakers. The algorithm works by increasing the difference between the left and right channels, $left_{stereo}$ and $right_{stereo}$, respectively. This is shown below in Eq. 3-1 to Eq. 3-4.

$$sum = left_{stereo} + right_{stereo} \quad \textbf{Eq. 3-1}$$

$$difference = left_{stereo} - right_{stereo} \quad \textbf{Eq. 3-2}$$

$$left_{wide} = \frac{sum}{2} + \left(\left(difference\right) * width\right) \quad \textbf{Eq. 3-3}$$

$$right_{wide} = \frac{sum}{2} - \left(\left(difference\right) * width\right) \quad \textbf{Eq. 3-4}$$

The parameter *width* takes a value from 0 to 1, with 0.5 being no width adjustment, as this would produce a left and right signal that are the same as the stereo input signals. A width of 0 produces the same signal for the left and right channels, which is the monophonic sum of the left and right stereo input channels.

### 3.1.2 Surround sound systems

The most common arrangement of surround sound speaker systems is the 5.1 configuration. This is based on the same left and right channels as the stereo arrangement, with an additional centre speaker at the front and two rear speakers; these constitute the "5". There is also a low frequency enhancement (LFE) speaker, which is the ".1". The recommended speaker positioning is shown in Figure 3-2.

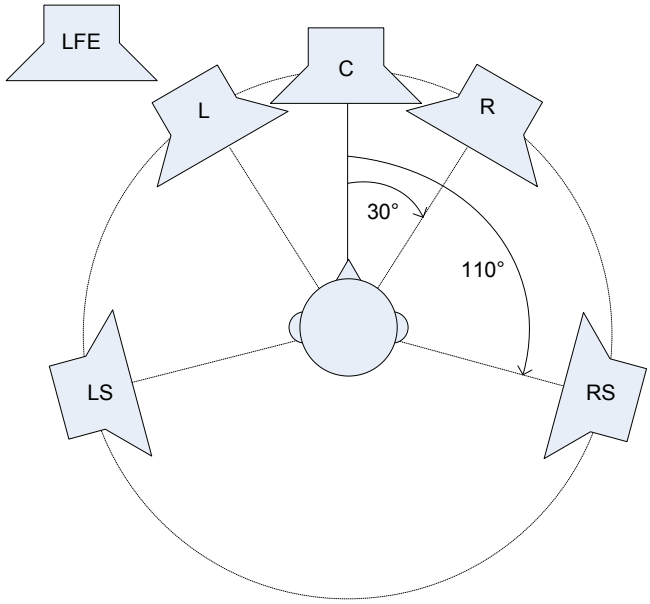



**Figure 3-2: The recommended loudspeaker positions for 5.1 surround sound.**

This configuration is mostly used for the audio soundtrack of films and it has become quite popular with the consumer format of digital versatile discs (DVD). Its specific use for audio, such as music, is starting to catch up with stereo since the introduction of DVD-Audio (DVD-A) format and super audio compact disc (SACD). The centre and rear channels are mostly used for special effects and speech, in the case of audio for films, and for reverberation and backing tracks for music. The additional channels are seldom used for accurately positioning a sound source, but simply immerse the listener in an auditory environment that cannot be recreated with stereo.

### 3.1.3 Ambisonics

Unlike conventional surround sound systems, ambisonics aims to reproduce a true three-dimensional sound image using an array of loudspeakers. The basic concept behind ambisonic systems is to record a sound source with multiple microphones that are located as closely as possible together, but with orthogonal directivity patterns. The original soundfield is then regenerated by processing the recorded signals and reproducing them over multiple loudspeakers.

### 3.1.4 Wave-field synthesis

The aim of wave-field synthesis is to reproduce sound sources by using an array of loudspeakers. This has been summarised and compared to ambisonics by Daniel *et al*. (2003). It is based on the concept that a wavefront can be considered as being emitted either by the original primary source or by secondary sources along the wavefront. The concept was first investigated by Berkhout (1988) and is shown in Figure 3-3.

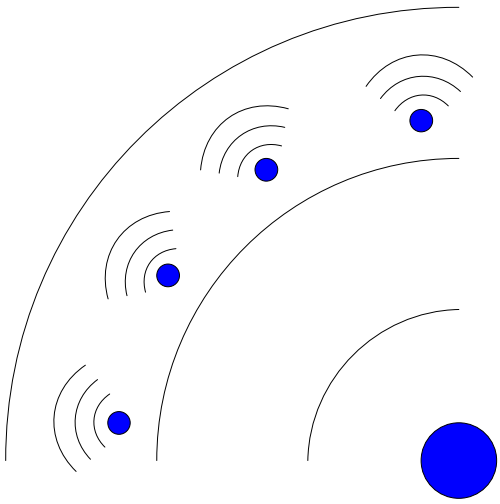


**Figure 3-3: Reconstruction of the wavefront from a primary sound source using secondary sound sources.**

The secondary sources are used to reconstruct the wave from the primary source. In practice this is achieved by recording the signals for a listening space which are then reproduced over loudspeakers, as shown in Figure 3-4.

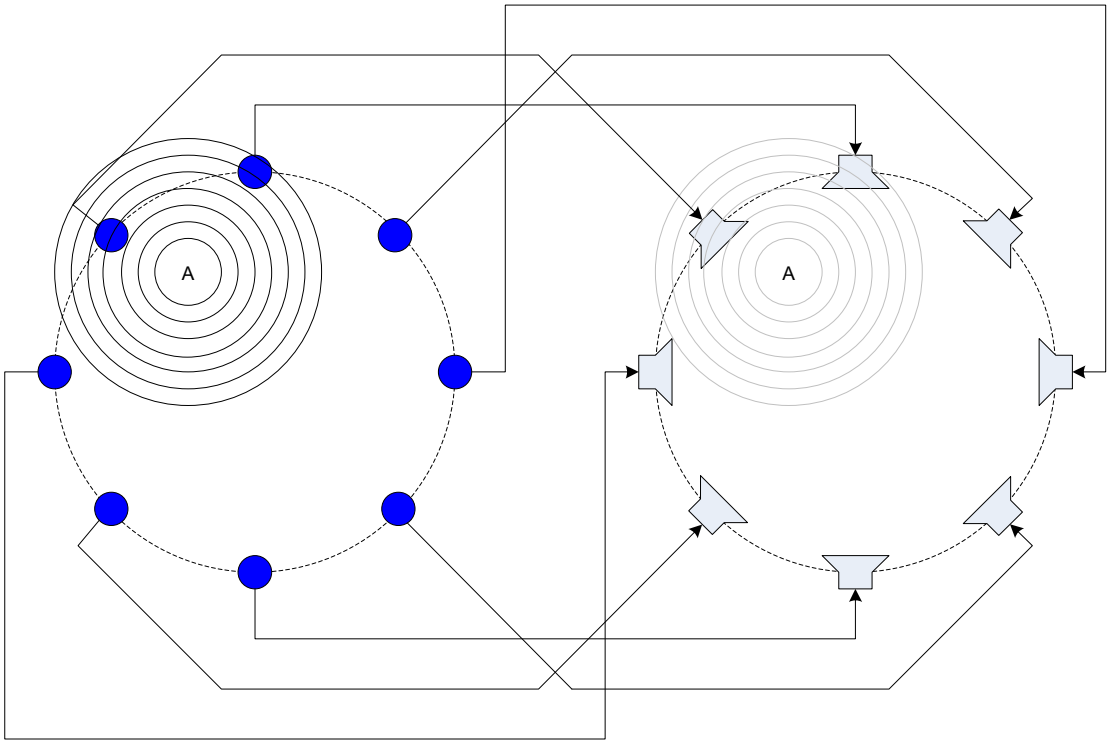



**Figure 3-4: The transfer of acoustical information from the recording space to the reproduction space in wavefield synthesis.**

The loudspeakers will then generate the equivalent sound that has been recorded, provided that an array of sufficiently close transducers is used and there are no system losses. In general, horizontal linear arrays are used rather than surfaces of transducers.

## *3.2 Binaural synthesis*

The aim of binaural synthesis is to replicate the movements of the eardrums that would have occurred with the equivalent signals in free field listening. Binaurally processed signals tend to be targeted at headphone presentation.

The most obvious way of recording a signal at the eardrum is to use a probe microphone in the ear canal, with the sound source positioned at the required location relative to the listener. Then, to replicate the free field signal, the recorded signal is presented through a tiny loudspeaker positioned at the same location as the probe microphone. This method is far from practical and, as a result, an extensive amount of work has been done to investigate more suitable techniques.

One approach for generating binaural signals uses interaural time and intensity differences as the basis for creating spatialised sound (Duda *et al*., 1999). These spatial cues were modelled using a spherical representation of the head by Duda and Martens (1998). Although the model is crude it is possible to adapt it to emulate, and so partially cater for, the anatomical differences between one listener and another. However, it has been found that sounds presented binaurally with the original time and intensity differences but lacking the spectral cues introduced by the external ears are typically perceived as originating inside the listener's head (Wightman and Kistler, 1989a).

### 3.2.1 Head-related transfer functions

A head-related transfer function (HRTF), is the description of the spectral filtering that occurs between a sound source for a specific location and the listener's eardrum (for example, see Cheng and Wakefield (1999) and Moore (1997)). HRTFs vary as functions of azimuth, elevation and distance. In the time domain they are referred to as head-related impulse responses or HRIRs (Møller *et al*., 1995a; Blauert, 1997).

It is widely accepted that if the spectral and temporal changes that occur in the free field can be replicated over headphones, the listener will experience the same free field listening conditions virtually (e.g. Møller *et al*., 1996).

We present here a brief overview of the description and use of HRTFs. For a more detailed review, the reader is referred to Cheng and Wakefield (2001).

### 3.2.2 HRTF measurement

We begin a discussion of the processes involved in recording an HRTF by considering the signal path for the free field condition. By examining the sound pressure at points along the signal path it is possible to produce what has been termed the free field transmission model (Møller *et al*., 1995b). The pressure positions are indicated in Figure 3-5.

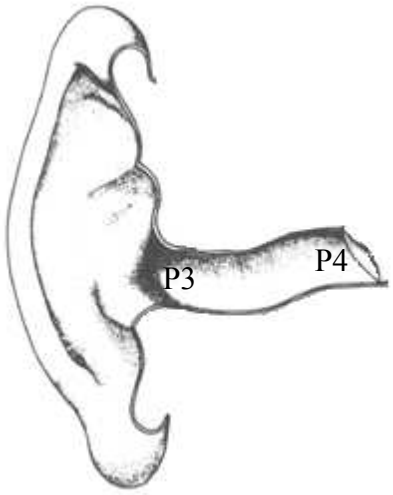



**Figure 3-5: Acoustic pressures at the ear canal (P3) and the eardrum (P4), adapted from Møller *et al.* (1995a).**

The pressure at the entrance to the ear canal due to an external sound source is denoted by $P_3$. The corresponding pressure at the eardrum is $P_4$. $P_2$ is the pressure at the entrance to the occluded ear canal. It is not shown in Figure 3-5 which shows the ear canal in the unblocked condition. Sound transmission along the ear canal is independent of the direction of sound incidence and is given by the division of the pressures at the entrance to the ear canal $P_3 / P_2$ and the transmission along the ear canal, $P_4 / P_3$. The transmission outside the ear canal is direction-dependent. $P_1$ denotes the sound pressure at the position corresponding to the centre of the head with the listener absent. The directional part is therefore given by $P_2 / P_1$ ($\theta$, $\phi$), where $\theta$ and $\phi$ are azimuth and elevation, respectively, as shown in Figure 2-1. The complete free field transmission to the eardrum is therefore given in Eq. 3-5 by:

$$\frac{P_4}{P_1}(\theta,\phi) = \frac{P_4}{P_3}.\frac{P_3}{P_2}.\frac{P_2}{P_1}(\theta,\phi) \qquad \textbf{Eq. 3-5}$$

However, the measurement of HRTFs is only concerned with the directional part of the transfer function. Therefore, measurements of $P_2$ and $P_1$ are all that is required. It follows that the directional characteristics of the sound transmission to the ear may be totally captured with HRTFs measured with occluded ear canals. What is more, these pressures are significantly easier to measure in the laboratory than $P_3$ and $P_4$. The effectiveness of HRTFs measured using occluded ear canals was confirmed by Møller *et al*. (1995a) by analysing the variation between HRTFs for 40 human subjects. They found that, across the range of subjects in their analysis, the variation in measurements with occluded ear canals was smaller than at the same points with open ear canals. The reduction in variance leads to a more accurate set of average HRTF data for representing a typical subject.

The method of placing a microphone at the entrance to the occluded ear canal and recording a (usually) wideband test signal from a sound source at a known location has also been used by Butler and Musicant (1993), Pralong and Carlile (1994), and Algazi *et al*. (1999). This simplifies the measurement procedure as it is unnecessary to insert miniature microphones into the ear canals of the listener, as used by Wightman and Kistler (1989a), which is far from practical and can potentially be hazardous to the subject. The signal recorded at the occluded entrance to the ear canal is deconvolved with the original test signal at the desired spatial location to produce the HRTF for that direction. This process is repeated to produce a three-dimensional matrix of HRTF data for all the required locations around a listener.

For the HRTF measurement techniques described above only the directional free field signal path is required. In practice, however, the signal will also have been affected by the transfer functions of the full signal path, to create what is known as the total system transfer function (TSTF). For example, the signal will have

been subjected to the transfer functions of the signal generator, amplifier, loudspeaker, microphone, room, cables, as well as the wanted acoustic effects of the hear, torso and pinnae. This additional coloration of the signal, called the system transfer function (STF), is independent of direction and can be measured in the room without the listener being present. Eq. 3-6 shows that the STF can be extracted from each of the measured HRTFs for the subject, to leave just the directionally-dependent information of interest. The origin of the transfer function for each of the components in the system is denoted by the subscript in Eq. 3-6.

$$HRTF = \frac{TSTF}{STF} = \frac{TF_{SigGen}TF_{D/A}TF_{Amplifier}TF_{Loudspeaker}TF_{HRTF}TF_{Microphone}TF_{Preamplifier}TF_{A/D}}{TF_{SigGen}TF_{D/A}TF_{Amplifier}TF_{Loudspeaker}TF_{Microphone}TF_{Preamplifier}TF_{A/D}}$$

**Eq. 3-6**

Once the undesirable response of the system has been removed the reproduction signal path can be considered. For example, it is usually necessary to remove the transfer function of the binaural sound reproduction system (typically headphones). As with the measurement system transfer function, the reproduction system transfer function can be obtained by playing a test signal through the headphones that the listener will use and recording the acoustic output at the listener's occluded ear canal using a probe microphone. Møller *et al*. (1995c) discuss the importance of headphone equalisation for accurate rendering of a binaural sound field. They describe measurements of the transfer functions from headphone to ear canal for 14 sets of headphones and concluded that none of the headphones they tested were adequate in a binaural reproduction system without equalisation. This was due to the large frequency response differences they observed between subjects even when using a single model of headphone. This led them to recommend individual equalisation of the headphone transfer function when subjects are presented with spatialised audio. The influence of headphone transfer functions is also confirmed by Pralong and Carlile (1996).

To perform the direct acoustic measurement of personalised HRTFs for each potential listener is impractical, (Wenzel *et al*., 1993), as it is difficult and time consuming. Therefore, a number of researchers are investigating the possibility of synthesising HRTFs from, for example, optically-captured models of the ears, head and torso of the listener. Different approaches to this method include boundary element modelling (Katz, 2001a, 2001b), spherical harmonics (Tao *et al*., 2003), elliptic Fourier transform methods (Hetherington and Tew, 2003), adaptive filtering (Norris, 1998), and the use of a database of measured HRTFs (Toyama *et al*. 1999). An alternative acoustic approach is measurement by reciprocity. This involves placing microspeakers in the subject's occluded ear canal and surrounding their head with miniature microphones in the directions for which HRTFs are required (Zotkin *et al*. 2006).

To reproduce an accurate 3D auditory environment for a particular listener, their HRTFs need to be measured for numerous positions around them. The number and positions of the measurements varies greatly depending on their intended use. Examples of some of the data sizes that have been captured are shown in Table 3-1.

| **First author** | **Number of HRTF locations measured** |
|---|---|
| Møller (1992) | 97 |
| Butler (1993) | 104 |
| Wightman (1991) | 265 |
| Kistler (1992) | 265 |
| Middlebrooks (1990) | 324 |
| Carlile (1994) | 343 |
| Pralong (1994) | 343 |
| Evans (1998) | 648 |
| CIPIC database | 1250 |
| Chen (1995) | 2188 |

**Table 3-1: The number of locations used by various researchers for measuring HRTFs**

The traditional reference point for measuring HRTFs is the centre of the head. However, for nearby sources there is a discrepancy between the azimuth angle subtended at the mid-point of the interaural axis and the angles subtended at the ears of the listener. This is explained diagrammatically in Figure 3-6. The dashed lines show the direction of arrival for a distant sound source, the solid lines show approximately how the angle changes for a nearby sound source.

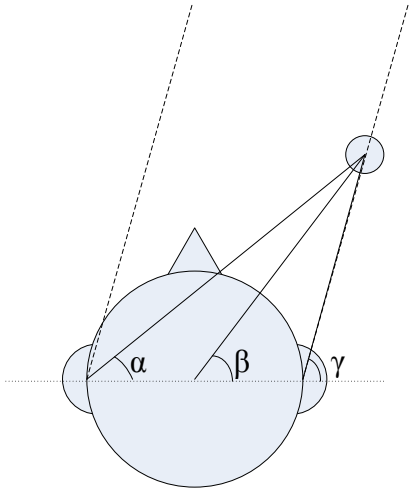



**Figure 3-6: The acoustic parallax effect for nearby sources, which results in the three angles *α*, *β* and *γ* being different.**

This is known as the acoustic parallax effect (Brungart, 1999a). Generally, for sound sources over a metre away from the listener, the effects of acoustic parallax tend to be ignored.

The duration of each HRIR in a set also varies across different experiments. Examples of the lengths typically used are shown in Table 3-2. Kistler and Wightman (1992) use a duration of 10.24 ms for their test signals. However, in earlier experiments Wightman and Kistler (1989a) and also Asano *et al*. (1990) use 20.48 ms signals for measuring HRIRs (at a sample rate of 50 kHz). The entries in the CIPIC database use 200 samples for each direction at a sampling rate of 44.1 kHz, equivalent to a length of approximately 4.5 ms (University of California, 1998).

| First author | Duration of HRIR (ms) |
|---|---|
| CIPIC database | 4.5 |
| Kistler (1992) | 10.24 |
| Wightman (1989a) | 20.48 |
| Asano (1990) | 20.48 |

**Table 3-2: Durations of HRIR measurements across different research groups.**

The amount of data required to represent the listening environment accurately used to be seen as a major problem, as HRIRs and HRTFs can require hundreds of kilobytes of memory. However, the memory requirement is no longer a concern as the cost of electronic memory has reduced significantly over the last few years, with the majority of modern computers having gigabytes of memory. In addition, personal computer processor speeds have increased at a rapid rate (Moore, 1965), and most even provide highly efficient single-instruction-multiple-data (SIMD) instructions that can be utilised for audio signal processing algorithms such as convolution. Modern machines allow researchers to generate the output from spatial audio processing algorithms in a few seconds rather than in hours. Although this improvement in technology facilitates the use of longer test signals and HRIRs, and higher sample rates, this may result in HRIRs that are contaminated by reflections from equipment in the room. Reflections from objects are generally removed from the impulse response simply by cropping the waveforms. Landone and Sandler (1998a) found that it is possible to truncate HRIRs to 1.5 ms (66 samples at 44.1 kHz) without degrading the most important spectral features, which include the cues for front-back disambiguation.

Although memory size and processor speed is no longer considered a limiting factor in the laboratory, there have been numerous studies investigating methods of reducing the amount of data required to portray a set of HRTFs accurately. It is envisaged that these forms of optimisation may prove useful when migrating spatial audio processing algorithms to portable and low power devices, such as mobile phones and hearing aids. Some examples of these studies include principal component analysis (Wightman and Kistler, 1991, and Kistler and Wightman, 1992), pole-zero approximation (Blommer and Wakefield, 1992,

1994, 1997; Kulkarni and Colburn, 1995; Jenison, 1995), modelling of the contralateral HRTF from the ipsilateral HRTF (Avendano *et al*., 1999b), other mathematical models (Chen *et al*., 1995; Kahana *et al*. 1999) and genetic algorithms (Cheung *et al*., 1998a, 1998b).

### 3.2.3 Personalised and generic HRTFs

Due to the anatomical differences between listeners their HRTFs will also be different (see Lopez and Gonzalez, 2001, for example). The most accurate spatialisation will occur when a listener is presented with binaural signals that have been generated using their own set of HRTF data. This has been verified by Wightman and Kistler (1989b), who reported that subjects were able to locate signals processed with personalised HRTFs almost as well as real free-field signals. However, they also found that there was an increase in front-back confusions and at the time of publishing they were unable to offer an explanation for why this occurred. It is possible that this may have been due to spectral coloration as a result of the signal processing applied prior to presenting the signal to the listener. For example, in experiments performed by Bronkhorst (1995), lowpass-filtered signals performed better than wideband signals. He suggests that this may be due to inaccuracies in the simulation of the pinna cues above the 7 kHz lowpass filter cut-off frequency used. This reinforces the conclusion that accurate localisation requires accurate HRTF measurement and precise reconstruction of the time domain signals. Another possibility is that the absence of a visual cue for frontal virtual sources tends to make them appear behind the listener.

When applying spatialisation cues to music, some believe that a generic set of HRTFs is adequate, on the basis that a listener is not concerned about the precise location of each performer. Others point to the widely-acknowledged artefacts caused by the use of generic HRTFs. For example, it is well documented that generic HRTFs are unable to reproduce spatial cues accurately for all listeners (Blauert, 1997; Moore, 1997). More troublesome in the context of high quality sound reproduction are spectral colouration, sound source diffuseness due to the

contradictory localisation cues, and increased front-back reversals (Møller *et al*., 1996; Blommer and Wakefield, 1997). Another phenomenon caused by the use of generic (i.e. non-individualised) HRTFs is the collapse of the frontal sound stage, which causes sounds in front of the listener to appear inside the listener's head. However, it has been reported by Begault *et al*. (2000) that perceived externalisation of speech sound sources can be significantly improved with the use of a head-tracker and reverberation.

Typically, a music sound stage is reproduced over loudspeakers or headphones. The listener is assisted with the segregation and localisation of the sounds by prior knowledge of real acoustic instruments. These cues will be lacking for synthesised sounds, which become detached from the listener's acoustic space, McIlwain (2001). The advantage of using a generic set of HRTFs is purely pragmatic in that they are easily obtained, for example, by using a dummy head. The failings of dummy heads, however, have been documented by Minnaar *et al*. (2001a) and Møller *et al*. (1999), who have shown that many such heads provide poorly spatialised signals compared to HRTFs recorded using other human heads. Based on this they have attempted to improve the design of a dummy head so that its HRTFs more closely match those for a typical human listener (Christensen *et al*., 2000; Hammershoi *et al*., 1992). Attempts have also been made to adapt generic HRTFs for use with any listener (Advanced Micro Devices, 1997). However it is to be achieved, it is likely that the easy availability of perceptually accurate individualised HRTFs will be a determining factor in the widespread adoption of binaural sound spatialisation in next-generation consumer technologies for high quality entertainment and mobile communications.

Other work that includes the use of personalised HRTFs includes Wenzel *et al*. (1993) and Wightman and Kistler (1989b), who have carried out experiments to investigate localisation accuracy. The ability of a listener to localise sounds correctly using a particular set of HRTFs is related to how close the set is to their own HRTFs (Middlebrooks, 1999). This suggests that a generic set of HRTFs may perform well for one listener but poorly for another. Non-individualised HRTFs may provide adequate spatial cues for the majority of listeners, (Wenzel

*et al*., 1993), but in many situations the other problems they introduce must also be taken into consideration.

### 3.2.4 Sound spatialisation using HRTFS

In spatial sound synthesis the aim is to produce a realistic simulation of the auditory space around a listener. The auditory cues that need to be emulated to produce a credible three-dimensional auditory event can be generated using an HRTF left-right pair (Landone and Sandler, 1998b). The result is known variously as a virtual acoustic scene, a virtual auditory scene or a virtual auditory display. A brief background on the theory of digital signal convolution is presented, followed by a description of methods for the reproduction of spatialised sounds.

#### 3.2.4.1 Convolution

Convolution describes how the input to a linear, time-invariant system interacts with the system to produce the output. In discrete time, the output of the system $y(n)$ is given by convolving the input sequence $x(n)$ with the sampled impulse response of the system $h(n)$. This is known as direct convolution, where $\otimes$ is the convolution operator and is written mathematically as shown in Eq. 3-7.

$$y(n) = \sum_{m=-\infty}^{\infty} h(m)x(n-m) = x(n) \otimes h(n) \qquad \textbf{Eq. 3-7}$$

The convolution sum can therefore be described as the cross-correlation of one sequence with a time-reversed second sequence; that is, the elements of the first sequence are multiplied point-by-point with the time-reversed elements of the second sequence.

The reverse process can be used, for example, to obtain the impulse response of a system if the input and output sequences are known. This is known as system identification.

$$h(n) = \frac{y(n) - \sum_{m=0}^{n-1} h(m)x(n-m)}{x(0)} \qquad n \geq 1,\ x(0) \neq 0$$ **Eq. 3-8**

Similarly, if the impulse response of the system and the output sequence are known, the input sequence can be calculated.

$$x(n) = \frac{y(n) - \sum_{m=1}^{n} h(m)x(n-m)}{h(0)}$$ **Eq. 3-9**

Eq. 3-10 describes the fast linear convolution of signal $x_1$ with $x_2$, having sample lengths of $n$ and $m$ respectively and $-\infty < k < \infty$.

$$x_1(n) \otimes x_2(m) = \mathrm{F}_{\mathrm{D}}^{-1}[X_1(k)X_2(k)]$$ **Eq. 3-10**

Where, $\mathrm{F}_{\mathrm{D}}^{-1}$ denotes the inverse discrete Fourier transform (DFT), $X_1(k)$ is the Fourier transform of $x_1(n)$, and $X_2(k)$ is the discrete Fourier transform of $x_2(m)$. The equation states that the convolution of two sequences is equivalent to the inverse DFT of the product of the DFTs of the sequences. This assumes that the two sequences are the same length and periodic, as required by the DFT. The DFT is one of the core mathematical operators used in the audio processing described in Chapter 7.

For sequences of length $N$, the direct convolution method requires $N^2$ real multiplications. The method of fast linear convolution requires $12N \log_2 2N + 8N$ real multiplications (Ifeachor and Jervis, 2002). Table 3-3 shows that by using sequences of more than 128 elements the fast convolution method requires fewer multiplications than the direct method.

| N | Direct method | Fast convolution | Ratio fast:direct |
|---|---|---|---|
| 8 | 64 | 448 | 7 |
| 16 | 256 | 1088 | 4.25 |
| 32 | 1024 | 2560 | 2.5 |
| 64 | 4096 | 5888 | 1.4375 |
| 128 | 16384 | 13312 | 0.8125 |
| 256 | 65536 | 29696 | 0.4531 |
| 512 | 262144 | 65536 | 0.25 |
| 1024 | 1048576 | 143360 | 0.1367 |
| 2048 | 4195304 | 311296 | 0.0742 |

**Table 3-3: Computational saving of fast convolution compared with the direct method in terms of multiplications required. Taken from Ifeachor and Jervis (2002).**

#### 3.2.4.2 Binaural synthesis of a single sound source

In the direct approach to binaural synthesis, the HRIR samples become the coefficients of a finite impulse response (FIR) filter. This filter convolves the HRIR coefficients with the mono anechoic input signal to produce a spatialised signal for each ear that is *theoretically* indistinguishable from a real signal at that location. As has been shown in Section 3.2.4.1, which describes the theory of convolution, the equivalent operation in the frequency domain is to multiply each HRTF by the DFT of the audio signal and to take the inverse DFT of the result. This allows for a more efficient implementation.

The topology of the FIR filter requires that, for each output sample, the input samples are shifted along a delay line and repeatedly multiplied by the coefficients. By continually streaming input samples into the system, the filtered output samples will be produced continually. However, in the frequency domain it is necessary to split the audio signal into temporal segments and process them individually (Rabiner and Gold, 1975), otherwise the output signal is not computable until the entire signal has been received and stored.

To prevent discontinuities occurring at the boundaries of each window, a shaping function is usually applied to the samples in each segment. An example of a raised cosine bell shaping function is the Hanning window, which is shown in Figure 3-7.

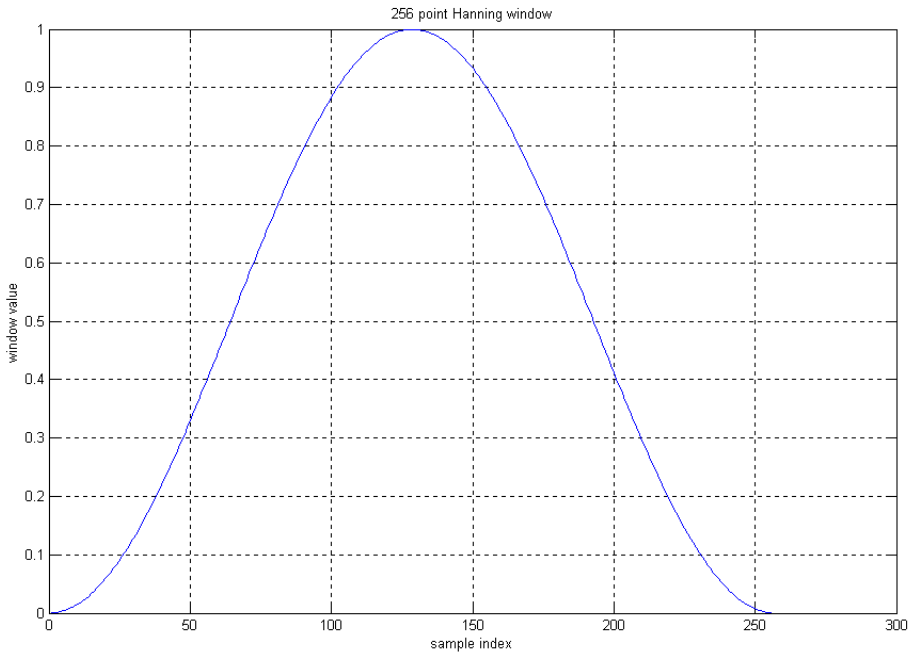



**Figure 3-7: A 256-point Hanning window**

Overlapping by 50% and adding windows with this shape produces the composite window signal shown in Figure 3-8. This produces a constant signal gain factor of 1, apart from the first and last half windows.

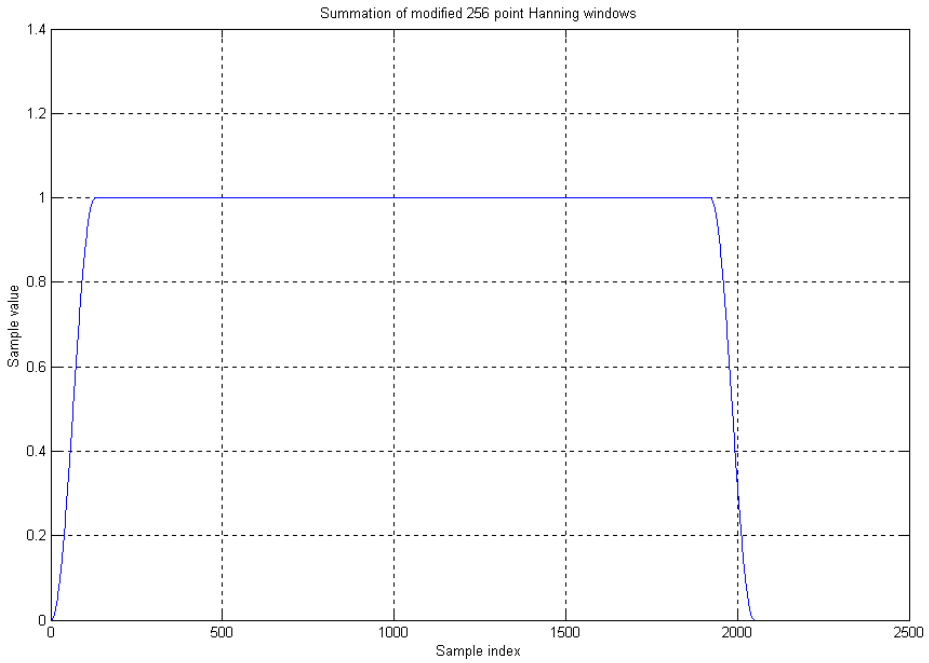



**Figure 3-8: The effect of applying the Hanning window using overlap-add creates a constant signal gain of unity.**

Therefore, an input signal can be windowed and summed to produce an identical output signal, so long as padding zeros are placed at the start and end of the input sequence of samples. The length of the signals used for the convolution is also important. In general the two input signals will be of different lengths. If a signal with $N$ data points is convolved with another signal having $M$ data points, the resulting output sequence will have $N + M - 1$ data points. The $M - 1$ points at the end of the sequence must be overlapped and summed with the first $M - 1$ points of the next block of convolved data.

### 3.2.4.3 Binaural synthesis of multiple sound sources

It was shown in Section 3.2.4.2 how the overlap-add method of signal processing allows an input signal to be segmented and processed in short blocks. After the spectral processing is complete the segments can be converted back to the time domain and reconstructed to form a continuous output signal. The diagram below shows how two spatialised signals, A and B, are combined to generate mixtures for the left and right ears.

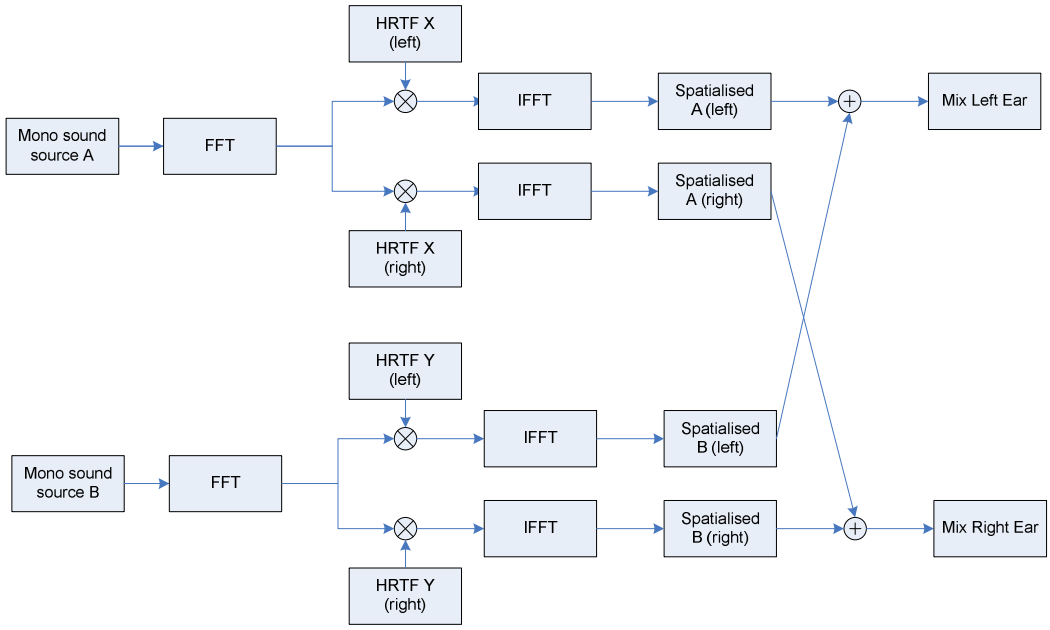



**Figure 3-9: How two spatialised signals are combined to produce a binaural mix output.**

It can be seen that the mono anechoic sound sources are individually spatialised with HRTF pairs to produce separate left and right signals for each source. These are summed to produce the left and right signals that are presented to each ear. Only two sources are shown in Figure 3-9, but the concept can be extended to allow any number of spatialised input sound sources.

## *3.3 Summary*

This chapter has presented a number of methods to artificially spatialise sound sources. The aim of sound spatialisation is to reproduce a three-dimensional auditory environment that is similar, if not identical, to the original free field listening environment. It has been shown that this can be accomplished either by using multiple loudspeakers to recreate the original sound field or binaurally using a pair of stereo headphones. Binaural synthesis requires the measurement, usage and application of head-related transfer functions (HRTFs). These have been discussed, along with techniques for optimising and improving the effectiveness of the digital filtering techniques required to reproduce realistic spatialised audio. An example of the overlap-add method for processing windowed blocks of digital audio has been provided. This processing model

forms the core of the system architecture that has been implemented for subsequent algorithm development. Further details about the system architecture and algorithms for sound source spatialisation are discussed in more detail in Chapter 7.

# Chapter 4 Auditory masking and the factors that influence it

It was explained in Section 2.3 how a masking sound can prevent a quieter sound from being heard, simply because the ear does not convey the displacement of the basilar membrane due to the quieter sound during the louder one. In addition, it was described how consecutive sounds can also be affected by temporal masking due to the propagation of the signals through the higher levels of the hearing system. The relationship between the masker and the target signal for simultaneous sounds is not straightforward, as the work described in this section illustrates.

Before considering the many factors that can affect the ability of one signal to mask another, it is useful to understand how auditory streams are grouped and formed within the hearing system. There are strong links between auditory grouping, auditory masking and auditory continuity. The background to this is discussed in Section 4.1.

## *4.1 Auditory grouping*

A listener presented with a single sound source, be it spatialised or diotic, is easily able to attend to it. This is because all the information is only attributable to the single stream of audio data that the brain has decoded. If one or more additional sound sources are introduced the task of apportioning attention to the individual streams becomes more complicated. This is attributed to the overlap of the physical properties of the audio streams, i.e. the spectral and temporal content, and also the listener's ability to predict the content of the target sound based on previous information (Bregman 1990). Figure 4-1 illustrates how this can be a significant problem when listening to multiple simultaneous sound sources.

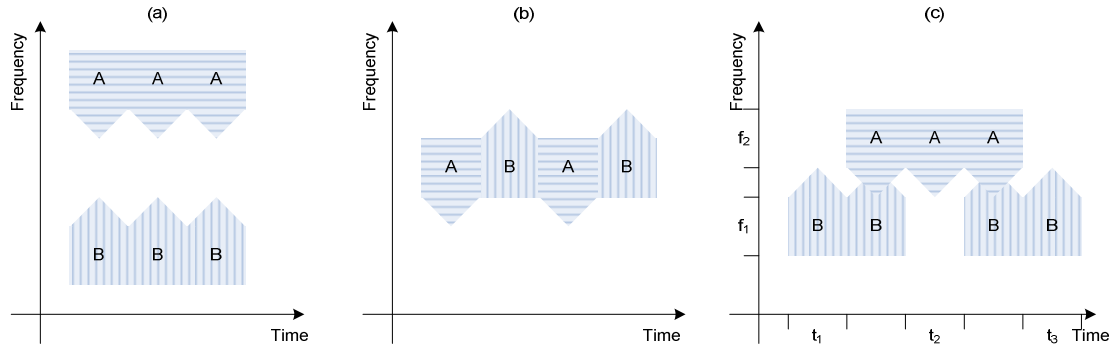



**Figure 4-1: (a) shows a configuration that has a temporal overlap of two signal components A and B, (b) shows a configuration that produces a spectral overlap and (c) shows a combination of spectral and temporal overlap with regions $f_n$ and $t_n$ denoting no overlap in the frequency and time domains, respectively.**

It is only in exceptional circumstances that a target sound source is heard without any interfering sound sources, for example in a well insulated anechoic chamber. This may help to explain why the human hearing system is so well adapted to concentrating on a single target sound source in a multi-source environment. The ability of human listeners to focus their auditory attention has been termed the *cocktail party* effect (Cherry, 1953). The name evokes the concept of listening to a single conversation in the midst of speech background noise, known as babble, at a cocktail party. It can be seen from Figure 4-1 that there are time slices $t_n$ and frequency bands $f_n$ that only contain sound information for a single sound source. One of the methods used to resolve the cocktail party problem involves detecting these regions and assigning them to an appropriate sound source.

### 4.1.1 Auditory Scene Analysis

The area of research that investigates the natural processes used to interpret the complex sounds presented to the human hearing system is referred to as *auditory scene analysis* (ASA). ASA seeks to explain how complex signals arriving at the eardrums are segregated and localised. Its goal is to recover separate descriptions of each sound object in the auditory environment. A considerable amount of research and experimental data has been collated by Bregman (1990).

This section presents only a brief overview of the primary processes of relevance to the technical work described in Chapter 7; namely, *proximity* and *spatial segregation*.

If two sound sources are sufficiently similar in one of more respects they will be grouped together. The property of auditory similarity is known as proximity. The dimensions used for measuring proximity in this context are time, frequency, amplitude and spatial location. An example of grouping by frequency proximity is shown in Figure 4-2, using tones A, B, and X.

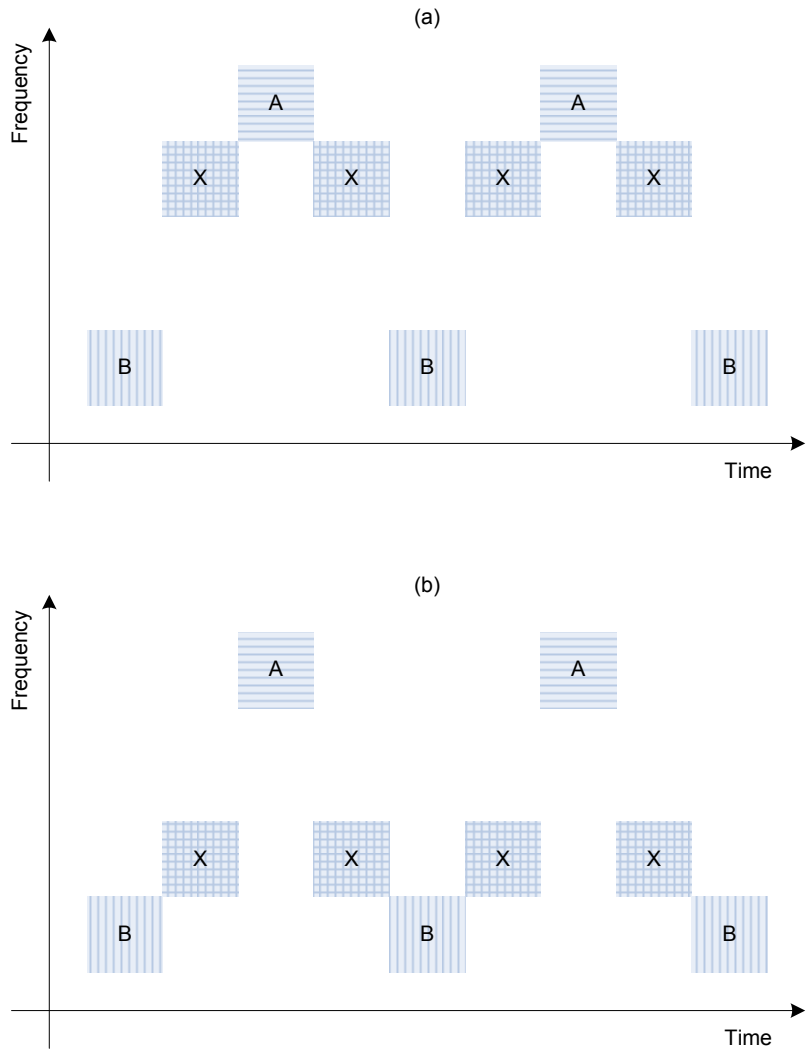



**Figure 4-2: Illustration of grouping by frequency, (a) shows that tone X is grouped with tone A, (b) shows that by changing the frequency of tone X it becomes grouped with tone B.**

Figure 4-2 shows, by means of a visual analogy, that if the frequency of sound source X is close to the frequency of sound source A, it will be grouped with A into a single stream and sound source B will heard as a separate stream. However, if the frequency of X is reduced such that it is closer to the frequency of B, it will be grouped with B into a single stream and A will be heard as a separate stream.

The temporal relationship of the tones can also influence the grouping. Figure 4-3 shows that spectral grouping can be disrupted and temporal grouping forced instead.

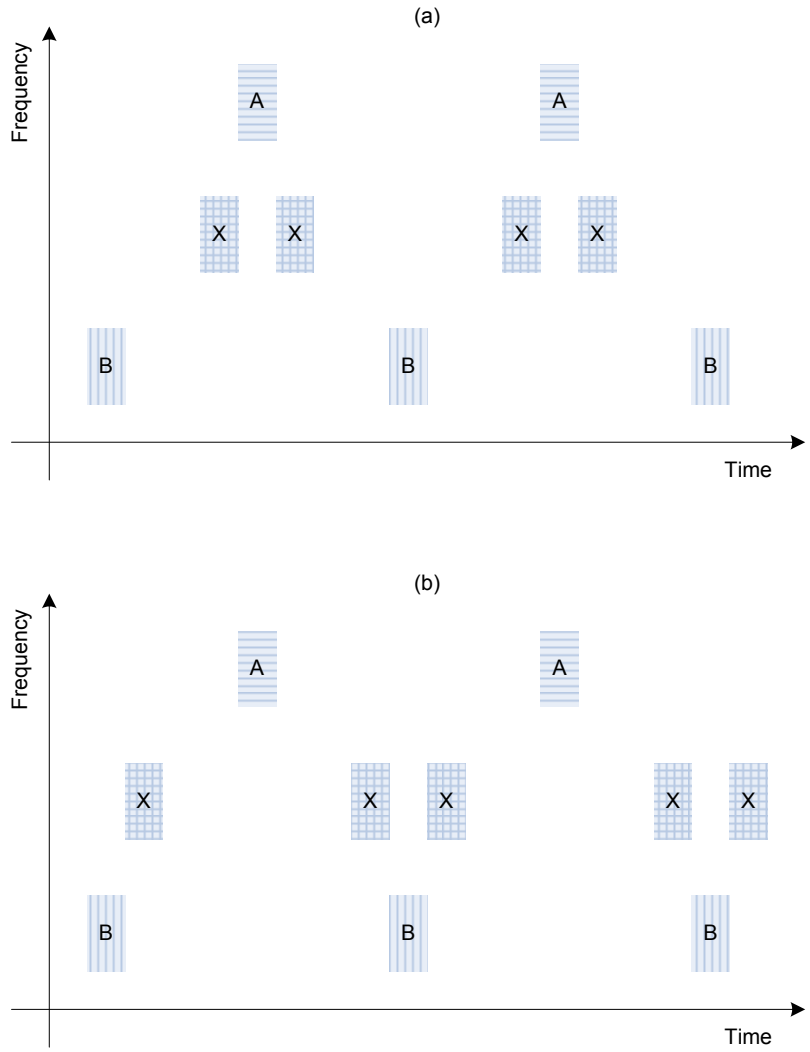



**Figure 4-3: Illustration of temporal grouping, (a) shows that tone X is grouped into a triplet with tone A and (b) shows that tone X is grouped into a triplet with tone B by altering its temporal pattern.**

It is possible to adjust the temporal and spectral relationships between the tones such that a listener can perceptually switch the stream that X is assigned to. However, it has been shown by Bregman (1990) that the tone X can only be allocated to a single stream at any time and not both. This is known as *exclusive allocation*.

With reference to Figure 4-1 for a typical 3-D acoustical scene containing multiple talkers, the frequency components from the different voices will tend to overlap spectrally and temporally. Therefore, early spectral analysis alone, performed by the auditory system, will be insufficient to segregate them. A discussion by Bregman of work by Kubovy suggests that if two sounds differ only by spatial location they will be fused into a single sound at an intermediate location. This is because in audition, the indispensable attributes are time and frequency, which means that two simultaneous sounds that differ only in frequency will be heard as separate. Similarly, two sounds of the same frequency but presented at different times will be heard as separate sounds. However, space is considered not to be an indispensable attribute, which is compatible with the observation that sounds at different spatial locations are perceived as a single source, if they have the same spectral content.

To investigate this further, Bregman (1990) describes an informal experiment he carried out. Two sounds were constructed, S1 consisting of frequency components at 200, 400, 600 and 800 Hz, S2 consisting of frequency components at 300, 600, 900 and 1200 Hz. All components had equal intensity. Both sounds were spatialised using dummy head HRTFs using an elevation of 0º, with one being located at an azimuth of +45º and the other at –45º. The sounds always had different start and end times, but overlapped for substantial lengths of time. As each sound has a 600 Hz component, Kubovy's argument suggests that it would be heard as a separate tone approximately in front of the listener. However, this was not the case; instead the 600 Hz component always remained spatialised in the directions of both sounds. Bregman suggests that this is because the tone is grouped with the other frequency components at the onset and offset times and is not extracted by presenting another sound containing the same tone at a different location.

This work was extended, to investigate whether the 600 Hz tone could be forced into the central location by suppressing the level of the other components in each sound. However, it was found that the tone still remained spatialised with S1 and S2. This provides further support for the notion that the onset and offset times are the critical factors in creating this effect. This is illustrated diagrammatically in Figure 4-4.

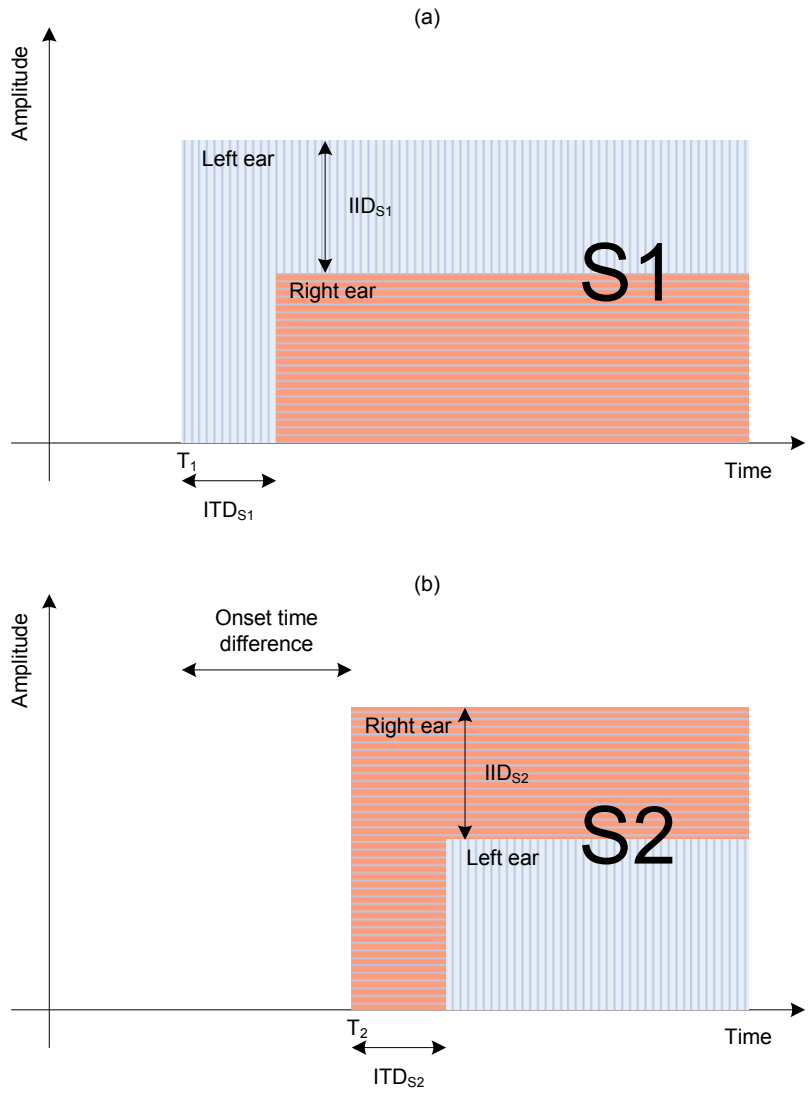



**Figure 4-4: Illustration of the two sound sources S1 and S2 used in experiments by Bregman (1990) to determine the location assignment of a single-tone component. (a) Sound S1 is spatialised to the left of the listener, (b) sound S2 is spatialised to the right of the listener.**

If the 600 Hz tone is first presented from the left side of the listener as a component of the first sound source S1, the ITD and IID cues at the onset are strong. When the second sound source S2 is introduced, from the right side of the listener, the ITD and IID cues indicate that the signal is located on the opposite side and not directly in front of the listener. Therefore, the 600 Hz tone within the second sound source is not fused with the existing tone within the first sound source on the left, but is heard after time T1 in the direction of S1 and then after time T2 also in the direction of S2. This is one of the few instances where the concept of exclusive allocation fails, as the tone component is allocated to two spatial locations simultaneously.

Research by Best *et al*. (2007) investigated the impact of binaural interference on auditory grouping. Binaural interference is used to describe the disruption of binaural cues due to the presence of simultaneous energy in other spectral regions. Best *et al*. (2007) suggest that this is most likely to occur when grouping cues bring the target and interferer into a single sound object. It follows that the effect of binaural interference depends on the way listeners use simultaneous and sequential grouping cues to organise a mixture into sound objects. Best *et al*. (2007) aimed to break the grouping of a simultaneous target and interferer, having common onsets and offsets, by capturing the interferer into a repeating auditory stream. The intention was to force the listener to switch from using simultaneous grouping to using sequential grouping. If the interferer were successfully extracted into a different stream the listener should be more able to determine the target's spatial location. Their results show that this is indeed the case. They propose that "…binaural interference is the result of obligatory grouping that occurs when there are no cues indicating the presence of two distinct sound sources (other than the spatial cues themselves)", Best *et al*. (2007). This again suggests that binaural cues alone are not always sufficient for segregating simultaneous sounds and ties in with the idea that they are not indispensable. They go on to add that the introduction of stronger auditory segregation cues, such as asynchronous onset times and sequential streaming, reduce the across-frequency grouping and enable listeners to access the binaural information in different frequency bands.

## 4.2 Masking and critical bands

It might be expected that the masking ability of a signal would be correlated with the range of frequencies it covers, i.e. a single frequency is more easily masked by a wideband noise than a narrowband noise. There is, however, a limit to the effectiveness of increasing the bandwidth of the noise masker, as discussed in Section 2.3. Bregman (1990) refers to this as the critical band of masking, based on earlier work by Fletcher (1940). This means that a tone will only be masked by frequencies that are adjacent to it and within the same critical band. Frequencies outside the critical band make no additional contribution to the masking of the target tone. Indeed, it has been shown that increasing the frequency range of the masker can actually reduce its masking ability. An experiment that illustrates this has been described by Hall *et al*. (1984). They found that when a tone is masked by a narrow band of noise, the amount of masking can be decreased by simultaneously presenting an additional flanking masker that has a different centre frequency to the main masker and target signal. Similar work by van de Par and Kohlrausch (1998) confirms this. If low frequency amplitude modulation is applied to the masking bands a release from masking is observed. This can even occur if the signal and on-frequency noise band are presented to one ear and the flanking noise band is presented to the other ear. In addition to this, Moore (1997) describes a similar experiment using a wideband noise masker. If low frequency amplitude modulations are imposed on the noise, the amplitude fluctuations across auditory filters in the ear will be correlated. As the bandwidth of this type of noise increases, it actually reduces its ability to mask the target signal. Hall *et al*. (1984) call this *comodulation masking release* (CMR).

Bregman (1990) provides a possible explanation for why this release from masking occurs, based on the 'peek' theory of scene analysis. Bregman argues that if an interfering sound source is amplitude modulated at a different rate to a target sound source it can be detected more easily as the listener gets a better 'peek' at the target signal when the interferer is at a lower intensity. This

phenomenon does not occur when frequency modulation is applied to the masking signal.

Moore (1997) observes that the reduction in masking ability due to an increase in masker's bandwidth also occurs in the case of forward masking. He suggests that as the bandwidth of masker noise is increased we can more easily detect when the noise stops. This might allow a listener to detect that there is one frequency band, containing the target tone, that did not stop at the same time and therefore something different was occurring in that spectral region.

The following sections discuss the factors that influence masking. In particular, the effects of different masking and target sound source types on speech intelligibility are considered in respect of their spectral, temporal and spatial relationships.

## 4.3 The influence of spatial separation between sound sources

Shinn-Cunningham *et al*. (2001) investigated the influence of source distance for speech reception thresholds (SRT). The SRT is defined as the target signal level that produces a speech intelligibility score that corresponds to 50% of keywords being identified correctly. The binaural advantage within the experiments is given as the difference (in dB) between the SRTs of two target/masker configurations. A diotic reference SRT was set based on the configuration where the target and masker are at the same location, in front of the listener at 1 m. This is shown in Figure 4-5 with a 0dB advantage at that location. A high context sentence was used for the target signal and speech modulated noise for the masker. A "high context" sentence is one where there is a strong contextual relationship between the keywords in the sentence, for example, "the **desk** and **both chairs** were **painted tan**", where the keywords are marked in bold. Speech modulated noise is generated by applying typical spectral and amplitude modulations found in speech to a white noise source. This allows the temporal

and spectral masking impact of speech to be recreated without the informational masking inherent in real speech. These were pre-processed using spherical head HRTFs. Shinn-Cunningham *et al*. claim that although these HRTFs would not contain precise localisation cues for each listener they would provide adequate cues for the scope of their listening tests. Each sound source could be spatialised to 0º, 45º or 90º to the right of the midline of the listener, and either at 1 m or 0.15 m from the centre of the listener's head. The results are shown diagrammatically in Figure 4-5 where the more positive number is the greater the binaural advantage. They found that when the target is fixed at 1 m in front of the listener (Figure 4-5 (a)), the largest binaural advantage of approximately 6 dB, occurs when the masker is also at 1 m but at an azimuth of 45º. That is, the SRT is 6 dB lower in this target-interferer configuration compared to the diotic reference condition. It is interesting that the greatest binaural advantage is achieved with the interferer at 45° and not 90° as might be expected. Yet, when the masker is at 0° (Figure 4-5 (b)), the greatest advantage of 6 dB occurs when the target is at 90°. By moving the masker to 0.15 m from the listener (Figure 4-5 (a)) an increase of approximately 13 dB is required for the target to maintain the same speech intelligibility levels. In Figure 4-5 (b), if the masker is fixed at 1 m in front of the listener, the advantage is approximately 6 dB when the target is at 1 m and 90º. However, by moving the target closer, to 0.15 m, the advantage increases to 30 dB. If the target and masker are both at 90º (Figure 4-5 (c)) there is an advantage of 21 dB by having the target closer than the masker to the listener.

These results show that in addition to increasing the angular separation between the masker and target signals, an increase in distance separation also produces substantial advantages.

As discussed in Section 2.4.3, one of the cues for determining sound source distance from a listener is the absolute sound level. Shinn-Cunningham's results also confirm the obvious statement that, to increase the audibility of the target sound source, the interferer sound source amplitude should be reduced as well as spatially separated.

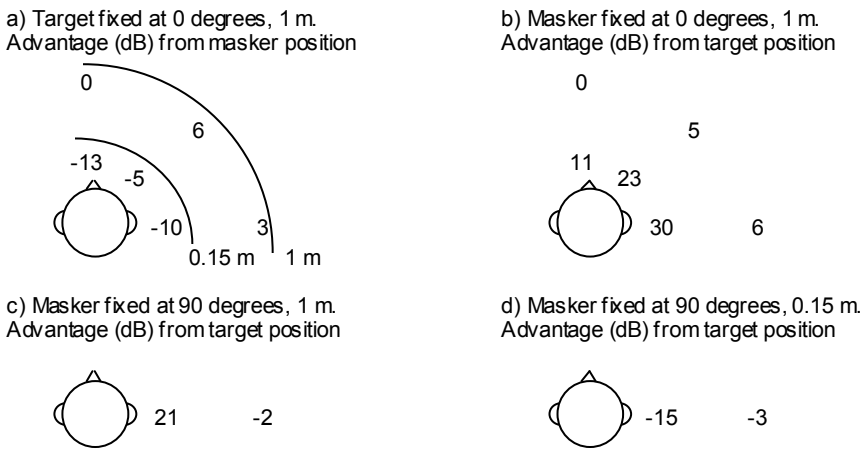



**Figure 4-5: Speech reception thresholds for target/masker configurations differing in angular separation and distance. From Shinn-Cunningham *et al.* (2001).**

Kidd *et al*. (1998) carried out experiments to investigate the release from masking for spatially separate sound sources. They were based on earlier findings when considering the binaural advantage for tone pattern identification (Kidd *et al*., 1995). Kidd *et al*. (1998) used signals that were presented from loudspeakers in a semicircle in front of the listener. The target signal was one of six tone patterns, which could occur within one of 16 frequency bands. Two types of masker signal were used, either a wideband noise (Figure 4-6 (a)) or a multi-band signal (Figure 4-6 (b)), which contained 8 randomly selected tone sequences, excluding the frequency bands that contained the target tone sequence.

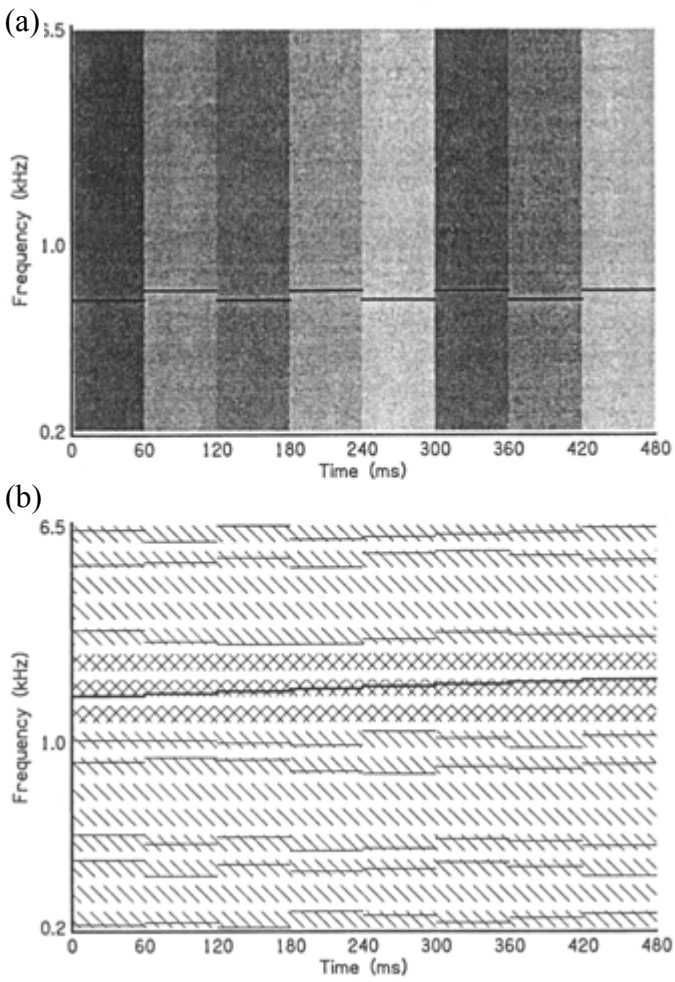



**Figure 4-6: (a) an alternating tone sequence target signal with a wideband masking noise. The shading highlights that the noise bursts are synchronised with the tones, each burst was at the same amplitude. Plot (b) shows an increasing ramp tone sequence target signal with 8 random tone sequences used as maskers. The maskers are assigned to frequency bands outside the band reserved for the target tones (cross-hatched area), as used in the experiments by Kidd *et al.* (1998)**

The first type of masker is classed as an energetic masker, the second is classed as an informational masker. Kidd *et al*. explain that an informational masker is able to disguise the target signal more than a noise masker when each is presented from the same location. This is because, in the case of the informational masker, the listener does not know which tone sequence is one of the patterns they are listening for, or which frequency band it will occur in. Therefore, the first tone in each frequency band could be assigned to the target or the masker and the listener must analyse the tonal movements to distinguish which frequency band to concentrate on. The results of the experiments indicate

that spatial separation provides a far greater improvement in detecting the target sequence with the informational masker, than with the noise masker. They suggest that this is because the listener is able to hear a specific narrow band that is uncontaminated by the masker and coming from a single direction. This makes it possible to focus on and extract the target sequence. The energetic masker, however, overlaps with the frequency band of the target sound. This may disrupt the localisation of the target signal, thereby reducing the effect of spatial separation. The results show that spatial separation between the target and the energetic masker produced advantages of 5 dB for targets below 3 kHz and up to 10 dB for higher frequencies. Spatial separation of the target from the informational masker produced advantages greater than 20 dB for all frequency bands tested.

All these points highlight the benefit of having spatial separation of the target and masker, in particular if the interferer is an informational masker of the target, e.g. if both are speech sound sources.

In more recent research by Gallun *et al*. (2007), the influence of a combination of interferer signals at both ears was investigated when attention was focussed on a target signal that was only presented to one ear. A target male speech signal was split into narrow frequency bands and presented simultaneously with an interferer that used either the same frequency bands or different bands. Five types of interferer sound were used: different-band speech, different-band time-reversed speech, broadband noise, same-band noise and different-band noise. One of their experiments showed that target speech recognition performance, with an ipsilateral different-band speech interferer, reduced as the level of a contralateral same-band noise interferer increased. The ipsilateral speech interferer was a different male talker to the target male speech. This result demonstrates that the listeners were unable to isolate the SNR advantage in one ear from the contralateral interference. Gallun *et al*. suggest that this is because listeners were unable to use two listening strategies simultaneously. That is, they had to select between an ear-based strategy (i.e. focus on the ear presented with the target) and a frequency-based strategy (i.e. focus on the frequency bands that contained the target). Therefore, there would have been no performance

degradation had the ipsilateral interference been removed. This was confirmed with an additional experiment in which target recognition scores were unaffected by a contralateral broadband interferer in these conditions. However, when a same-band interferer was used, the performance degraded as the number of frequency bands used for the target speech and noise interferer was reduced. Gallun *et al*. postulate that this was due to the reduced spectral content disrupting cues for determining similarity between the target and interferer and hence they were grouped into a single auditory stream. They recommend that further investigations be carried out to confirm this hypothesis.

## *4.4 The influence of sound duration*

Kohlrausch (1990) investigated the influence of target and masker signal durations on a listener's ability to detect the target sound source. The masker signal for his experiments was one of two types of dichotic flat-spectrum wideband noise. One had an interaural phase difference of zero below 500 Hz and of π radians above 500 Hz, denoted N0π. The other was the inverse, having an interaural phase difference of π radians below 500 Hz and zero above 500 Hz, denoted Nπ0. Therefore, the perceived lateral position of the masker depends on frequency. The noise was presented at 77 dB SPL for all experiments. The masker duration was either 500 ms or 25 ms, as shown in Figure 4-7. The target signals used in the experiments were sinusoids between 200 Hz and 800 Hz at intervals of 100 Hz. The target durations were either 250 ms with 20 ms linear ramps, or 20 ms with 5 ms linear ramps, also shown in Figure 4-7. Three bursts of the masker were presented, with gaps of 100 ms for the 500 ms masker and 300 ms for the 25 ms masker. The target was placed at the temporal centre of one of the gaps. The listener had to identify which gap contained the target signal.

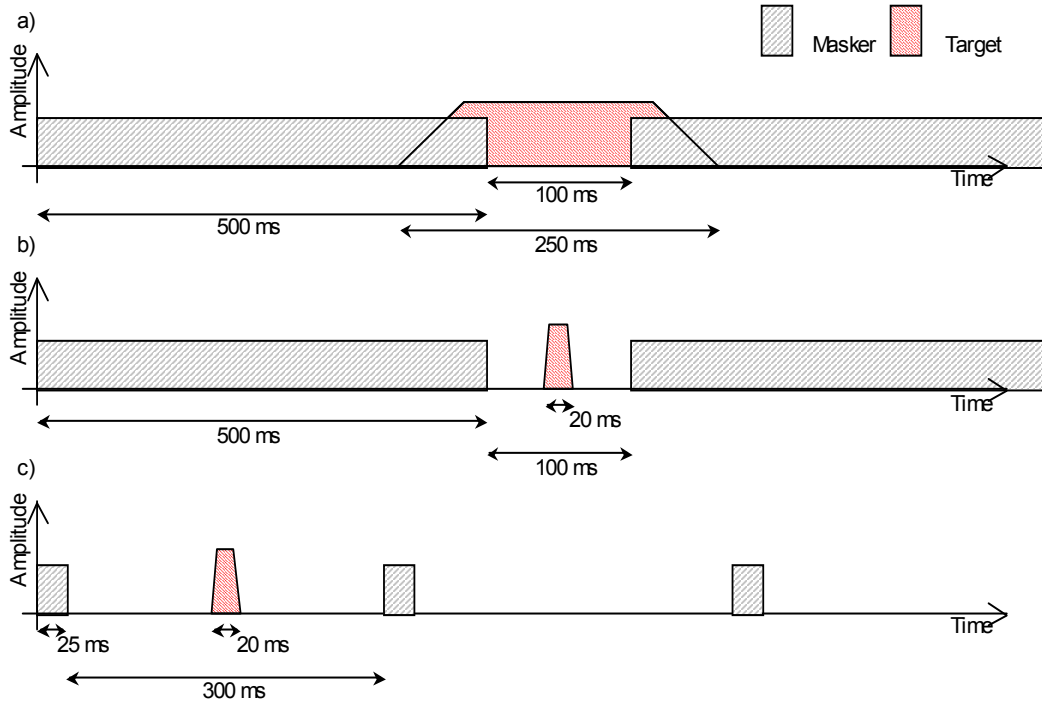



**Figure 4-7: Diagram showing the target and masker durations for the three different configurations used in experiments by Kohlrausch (1990). (a) shows the 250ms target sound overlapping with the 500ms masker sounds. (b) shows the 20ms target sound sitting in the middle of a 100ms gap between two 500ms masker sounds. (c) shows the 20ms target sound sitting in the middle of a 300ms gap between two 25ms masker sounds.**

Kohlrausch set out to show that the scenario depicted in Figure 4-7 (a) will produce the greatest binaural masking level difference (BMLD). That is, the difference in masking thresholds between a monaural or diotic reference condition and a dichotic condition. The reason given by Kohlrausch is based on work by McFadden (1966), who states that increasing the duration of the masker noise helps the binaural system to estimate the binaural parameters that lead to an accurate lateral image of the masker. The addition of an overlapping binaural target sound with a different interaural phase can then be detected more easily, as the lateral image of the corresponding frequency band of noise is moved away from its previous location and, after termination of the signal, is moved back to its original location. A signal gated simultaneously with the masker cannot produce these lateral movements at its onset and offset. Therefore, the cue of a moving image is not available to the listener. However, it was seen from the experimental results provided by Bregman (1990), discussed in Section 4.1.1,

that the lateral shift of a shared frequency component does not occur, which suggests other mechanisms must be involved in detecting the target in this type of masking scenario.

It was found that the binaural hearing system requires sufficient time to derive the spatial cues with which to localise a sound source. It appears that a masker duration of at least 300 ms is required for optimal detection of a simultaneously presented target signal. The results indicate that rapidly changing frequency components are unlikely to create a stable spatial image capable of affecting the binaural advantage of a spatially separate target and masker.

## *4.5 The influence of spectral content of sounds*

The frequency content of a target sound can influence BMLDs. For broadband noise maskers, BMLDs resulting from the comparison of a N0S0 condition with a N0Sπ condition are much larger at low frequencies than at high frequencies (Durlach, 1964). This is generally considered to be due to the reduced importance of ITD cues at high frequencies (see Zurek and Durlach (1987) and Moore (1997)). Experiments by van de Par and Kohlrausch (1997) also show that the failure of the hearing system to encode the temporal fine structure of high frequency signals contributes to the frequency-dependent difference in BMLDs. When narrowband noise maskers are used, the BMLDs are generally much larger, and depend on the centre frequency of the noise band. At low frequencies BMLDs can be as high as 25 dB, reducing to 15 dB at high frequencies (see Zurek and Durlach (1987)). McFadden (1972) measured BMLDs ranging from 11 dB to 2 dB for target tones ranging from 290 Hz to 600 Hz against a 400 Hz tone masker. As discussed in Section 2.3, maximum masking occurs when the target and masker sounds have similar frequencies, i.e. lie within the same critical band.

## 4.6 Sound source signal content

The influence of sound source location, signal duration and frequency content on masking has been discussed in Sections 4.3 to 4.5 respectively. This section considers how the masking affects speech intelligibility.

### 4.6.1 Phoneme recognition with a noise masker

Research by DeSimio *et al*. (1996) has confirmed the existence of the binaural advantage for phoneme recognition using sounds spatialised with generic HRTFs. They carried out experiments using a target speech signal presented simultaneously with a white noise masker in two spatial configurations. In the first, both sound sources were located in the same place, directly in front of the listener, and in the second, the target sound was directly in front and the noise masker 90º to the right. Nonindividualised HRTFs were used, based on data generated by Wightman and Kistler (1989a). Despite this, a binaural advantage was noted with the greatest phoneme recognition scores being achieved when the noise and masker were presented in the 90º separation configuration. Feuerstein (1992) also found that word recognition scores were significantly better for binaural compared to monaural listening for experiments based on simulated monaural hearing loss.

### 4.6.2 Speech recognition with a noise masker

Noble *et al*. (1997) describe work by Saberi *et al*. (1991), which suggests that improvements from spatial separation between a target sound and a masking sound may also be due to increases in signal-to-noise ratios in specific frequency bands, arising from directionally-dependent pinna filter effects. Additionally, this may explain the masking release for monaural listening where the noise and speech are located at different elevations in the median plane, where there are no interaural differences. The experimental work carried out by Noble *et al*. (1997) investigated whether listeners with hearing loss have a binaural advantage when

speech spectrum noise and a target speech source are spatially separated. The signals were presented via two arcs of loudspeakers, one horizontal and one vertical, each having a radius of 1.22 m. Subjects faced the central loudspeaker, for frontal tests and the leftmost loudspeaker, for lateral tests. The noise was presented from two loudspeakers, at equal angular separations on each side of the central loudspeaker. It was acknowledged that this might be perceived as a fused centralised noise source rather than distinct, spatially separated noise sources. Three groups of listeners were tested; a group with normal hearing, another with sensorineural hearing loss and one with conductive mixed hearing loss. The listeners with normal hearing displayed a clear benefit that increased as the spatial separation between the noise and speech increased laterally in the horizontal plane and vertically in the median plane, but remained constant for sound sources in front of the listener. It is possible that this may have been due to using a loudspeaker equidistant on each side of the target loudspeaker, which would generate almost identical noise signals arriving at each ear, irrespective of the spatial separation. The listeners with low sensorineural hearing loss followed the same pattern of benefit but at a much lower level. Listeners with more severe hearing loss found very little benefit even with large spatial separations. They suggest the reason for the difference in performance between normal and impaired hearing listeners is due to the pinna cues, which occur for high frequencies, and were therefore missing for the subjects with impaired high-frequency hearing. However, work reported in Section 2.4.2.4 (Butler and Humanski, 1992), suggests that pinna cues are less important for localisation and it is the ITD and IID that are dominant.

### 4.6.3 The articulation rate of speech

Culling and Colburn (2000) carried out experiments investigating the effect of the articulation rate of speech on intelligibility. They combined the two concepts, binaural intelligibility level difference (BILD) and binaural sluggishness. BILD is the level difference required to achieve the same level of intelligibility between a binaural condition and a reference diotic condition.

Binaural sluggishness is the inability of the binaural system to follow rapid changes in the temporal relationships between the sounds reaching the two ears.

Their first experiment investigated whether the performance of binaural perception mechanisms is impaired by rapidly varying temporal information. Listeners were presented with arpeggios consisting of three tones. The duration of each tone was consistent within a trial and ranged from 32 ms to 100 ms between trials, therefore producing different arpeggio rates. The arpeggios were repeated simultaneously with a wideband noise burst (0 – 10 kHz) for 1.6 s in either the N0S0 or N0Sπ conditions. N0 and S0 denote that the masker and the target signal, respectively, had the same phase at each ear across all frequencies. In the N0Sπ condition, the phase of the signal presented to each ear differed by π radians. The noise was presented at a spectral level of 38 dB SPL/Hz with start and end ramps that were much shorter than the steady-state duration of the tones. Two audio segments were presented to the listener in each trial, one with ascending arpeggios and the other with descending arpeggios. The listener was asked to indicate which segment contained the ascending arpeggios. The level of the arpeggios was adjusted using a two-down-one-up protocol according to the response of the listener. The average signal level of ten trials was used to calculate the discrimination threshold. To determine the detection thresholds the experiment was repeated but without the descending arpeggios. The listener had to indicate which noise burst contained the ascending arpeggios.

The general trend of the results showed that the arpeggios with a faster repetition rate had a higher discrimination and detection threshold than arpeggios at a slow repetition rate. In both the N0S0 and N0Sπ configurations the detection threshold was approximately 7 dB higher for the faster arpeggios than the slower ones. The detection thresholds showed a binaural advantage of approximately 12-15 dB between the two configurations. The discrimination thresholds showed a slightly different result: The binaural advantage at the slowest repetition rate was approximately 12-14 dB and at the highest rate the advantage reduced to approximately 3-5 dB.

In summary, Culling and Colburn (2000) observed that listeners found it more difficult to detect and discriminate faster arpeggios in noise, than slow arpeggios. Furthermore, the slower arpeggios exhibit a significant binaural advantage whereas the faster arpeggios only showed a small benefit.

Their second experiment investigated the speech reception thresholds (SRTs) at different speech articulation rates. Sentences containing 5 keywords were presented simultaneously with speech-shaped noise. Each sentence was processed with an acceleration factor of 1.5 and 2 times the original speed. The SRT was set based on a 50% intelligibility level. The speech signal level was altered by 2 dB between trials depending on the percentage correct word score, i.e. the speech signal level was increased if the listener was unable to correctly identify at least 50% of the key words in the sentence and decreased if they were able to. The average of the signal level in the final eight trials in each test was used to calculate the SRT. The results show that for an acceleration factor of two the SRT increased by approximately 6 dB and 8 dB for the N0S0 and N0Sπ conditions respectively. There was a binaural advantage of approximately 3-5 dB between these conditions. As the binaural advantage was similar across all three acceleration rates, this suggests that the binaural sluggishness that was measured for tone arpeggios is not apparent for speech.

The experiment revealed that frequencies between 4 and 8 Hz contribute to increasing the intelligibility of speech in noise. They also found that the average modulation spectrum for male voices peaks at about 4 Hz, approximately the natural syllable rate. They suggest therefore, that the lack of impact from binaural sluggishness is due to the useful modulation frequencies of speech being sufficiently low that they are not vulnerable to binaural sluggishness, even at the increased articulation rates tested.

### 4.6.4 Similarity in target and masker speech content

Brungart and Simpson (2007) investigated the interaction between target and masker speech signals when they are presented from different spatial locations.

Combinations of the same and different gender talkers were used, as listed in Table 4-1.

| Within-ear masker | Across-ear masker | Designation |
|---|---|---|
| Same talker | None | TT |
| Same talker | Same talker | TT-T |
| Same talker | Same gender | TT-S |
| Same talker | Different gender | TT-D |
| Same gender | None | TS |
| Same gender | Same talker | TS-T |
| Same gender | Same gender | TS-S |
| Same gender | Different gender | TS-D |
| Different gender | None | TD |
| Different gender | Same talker | TD-T |
| Different gender | Same gender | TD-S |
| Different gender | Different gender | TD-D |

**Table 4-1: The masker sound source type, relative to the target speech sound source type used by Brungart and Simpson (2007).**

Their results show some interesting relationships between the target and masker signals depending on whether the masker is the same talker, the same gender or the opposite gender to the target. This is summarised in Figure 4-8.

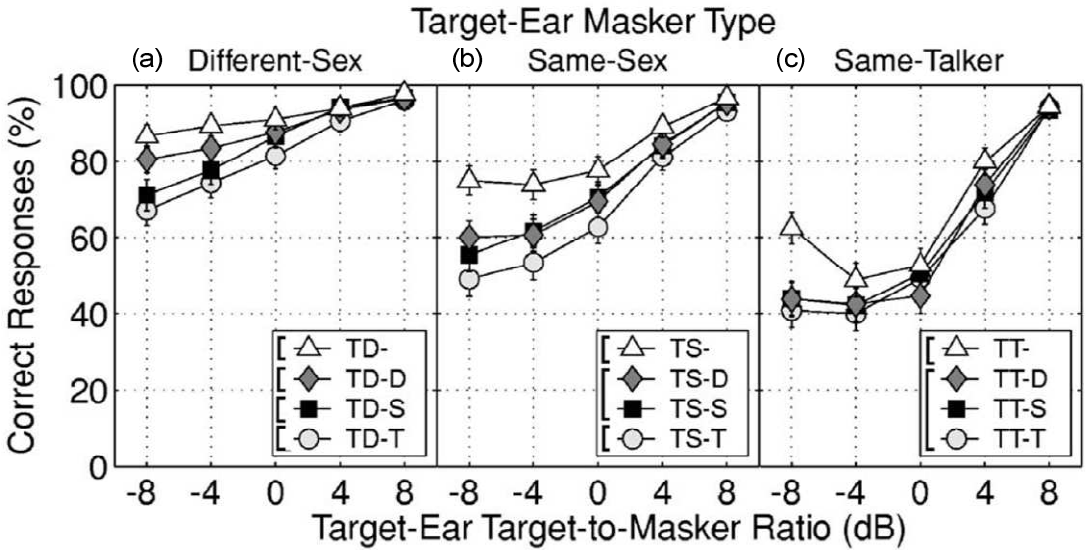



**Figure 4-8: The percentage correct word scores for three different masker sound types. The key refers to the masking designations given in Table 4-1. Taken from Brungart and Simpson (2007).**

Figure 4-8 (c) shows that when the target and within-ear masker are the same talker the type of interferer presented to the other ear has very little influence. However, when the interferer at the same ear as the target speech is of a different gender, the interferer at the other ear can significantly improve the listener's ability to understand the target speech (Figure 4-8 (a) and (b)). Evidently, there is a complicated relationship between the across-ear interference and its similarity to the target speech. Brungart and Simpson state that the relationship not only depends on the similarity between the across-ear masker and the target speech, but also on how similar the across-ear masker is to the target speech, compared to the similarity of the within-ear masker to the target speech.

They go on to explain this in relation to the *integrated strategy* model. That is, the listener assumes a single strategy for extracting the target speech from the multiple sound sources. Brungart and Simpson discuss this with the use of three scenarios:

1. The target speech is male and the within-ear masker is female. The optimum strategy is to segregate the male and female voices. If another female talker is presented to the other ear, the same strategy can be used. If however, the across-ear interference is a male talker the strategy breaks

down and the listener will confuse the interference with the target speech. Therefore a different, less effective, strategy may be employed to handle this change to the scenario. The result will be reduced target extraction performance.

2. The target speech is male and the within-ear masker is male. The strategy now is to focus on the smaller differences to identify the target talker. If a female talker is used as the across-ear interference, the performance is only slightly affected as it does not impact the strategy. However, if the across-ear interference matches the target talker it will greatly impact the strategy for extracting the target and performance will be degraded.
3. The target speech is male and the within-ear masker is the same talker. The listener has very little information to distinguish the target speech from the interference. If the across-ear interference is also the same talker this does not significantly impact performance as the strategy is focussed on the within-ear target extraction. Due to this, there will be very little benefit when the across-ear interference is a female talker as it will not influence the binaural strategy being used.

Iyer *et al*. (2007) compared the intelligibility performance for a speech target sound source in the presence of speech-shaped noise or same-talker speech. The interferer was presented in two configurations; simultaneously with the target, referred to as 'continuous', or temporally gated with the target such that only one sound source was presented at a time, referred to as 'interrupted'. The gating was applied using a square wave with a 50% duty cycle and a repetition rate of 4, 8, 16, 32, 64, or 128 Hz. In each time segment either the target was on and the interferer was off, or vice versa, so the resulting output signal was an alternating stream of the two sound sources. The target was temporally gated for all tests, including the 'continuous' configuration. The signals were presented diotically over headphones. The percentage correct word scores are shown in Figure 4-9.

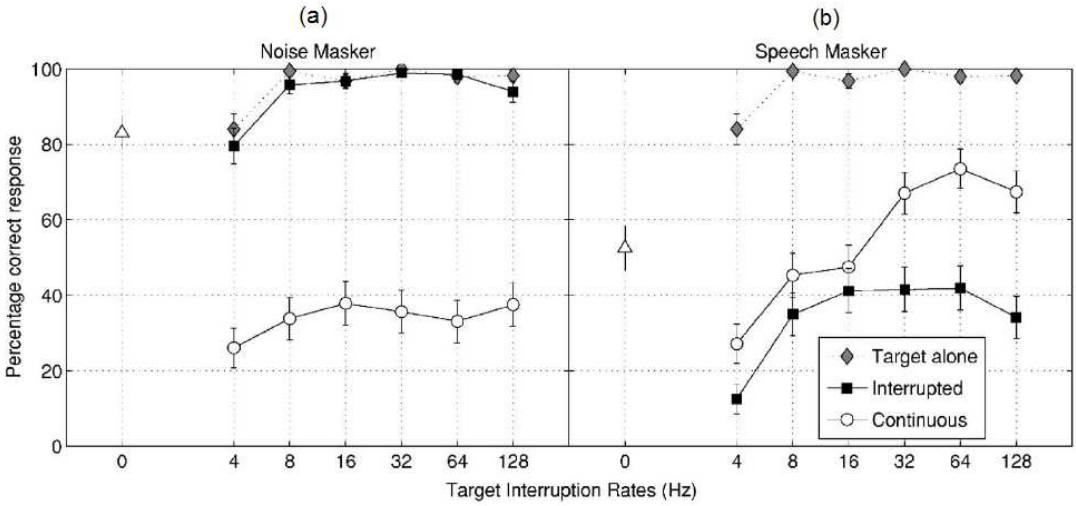



**Figure 4-9: The percentage correct word scores relative to speech target interruption rate for continuous and interrupted noise (a) and speech maskers (b). Taken from Iyer *et al.* (2007).**

Iyer *et al*. (2007) show that for a noise interferer sound source, performance is generally better when it is presented in the interrupted configuration as opposed to continuous (Figure 4-9 (a)). The opposite is true for a same-talker speech interferer (Figure 4-9 (b)). They attribute the increased performance with an interrupted masker to a release from energetic masking. This allows listeners to glimpse the target during the intervals when no, or very little, interference is present. They explain the reduced performance using the interrupted speech interferer as being because there is not only some loss of phonetic information in the target signal, but there is also nothing to make the target less 'similar' to the masker. They suggest that the continuity cues for the target and the interferer may also have been disrupted, preventing the two sound sources from being segregated. A continuous speech masker, however, could be segregated into a single stream, resulting in improved target speech recognition scores.

## *4.7 Summary*

This chapter has discussed a number of factors that influence the ability of an interfering sound source to mask a target sound source. It is helpful to know how the hearing system behaves, when presented with complex mixtures of sounds,

so that an effective processing system can be developed to enhance the intelligibility of a target speech source in challenging conditions. With this in mind, the key points in relation to masking can be summarised as follows:

The binaural advantage can be exploited by increasing the angular separation between the target sound source and masker sound source, giving nearly 20 dB advantage between speech reception thresholds for sounds close to a listener. Moving the target sound source closer to the listener increases the binaural advantage by approximately 20 dB if the target is at 0.15 m and the masker is at 1 m. Listeners with sensorineural hearing loss should also benefit from spatially separating target and interfering sound sources, but to a lesser extent than a listener with normal hearing.

To ensure sound sources are correctly segregated into separate streams the sound sources need to be *in situ* for at least 300-500ms. That is, the sound source's spatial location needs to be consistent and persistent for effective localisation to occur. This in turn allows the sound components to be grouped into streams.

Low frequencies provide more significant BMLDs than high frequencies. Therefore it is more important to emphasise the spatial segregation for lower frequencies.

The selection of target and interferer spatial locations needs to consider the effect of contralateral interference on a target sound source that is also affected by ipsilateral interference. A target sound source should be positioned such that it maximises the coherence between both ears, i.e. is spatialised at 0° azimuth, whereas the interferer sound sources should be located to minimise coherence between the ears, i.e. at approximately +/-90° azimuth.

It is not necessary to use personalised HRTFs to achieve a binaural advantage. Generic HRTFs are adequate for providing suitable spatialisation cues that allow exploitation of the binaural advantage due to spatially separated sound sources.

The natural syllable rate of speech makes it unsusceptible to binaural sluggishness. The binaural advantage is effective up to twice the normal articulation rates.

Listeners select the most appropriate listening strategy for the auditory environment they are presented with. It would be beneficial to guide the listener towards using a more effective listening strategy by emphasising the binaural difference for interfering sound sources and ensuring coherence for target sounds. This relates back to the earlier point of intelligently selecting the most effective spatial locations. Specifically, the listener will benefit if they are able to use the larger auditory stream differences rather than having to focus on the smaller spectral differences between target and interfering sound sources.

Temporal disruption of speech sound sources should be avoided as continuity is important for stream segregation. This is discussed further in the next chapter.

# Chapter 5 The cause and effects of auditory continuity

The factors affecting auditory masking were discussed in Chapter 4. In this chapter we consider auditory continuity, which is a topic closely related to masking and which appears to be the hearing system's way of reducing some of the problems caused by masking. In particular, we examine how inducing auditory continuity can increase the intelligibility of speech in the presence of interfering sounds.

According to Warren (1984), there are three types of continuity:

1. Homophonic auditory continuity – the same sound, different levels
2. Heterophonic auditory continuity – different sounds, different levels
3. Contextual catenation – phonemic restoration

Both homophonic continuity and heterophonic continuity involve the restoration of signal segments that have continuous spectral properties before and after the interruption, for example a constant amplitude sinusoid. The perceptual synthesis of obliterated fragments of a time varying signal is known as contextual catenation. Auditory continuity can be influenced in a number of different ways. These are reviewed in the following sections. Throughout these discussions, **A** is used to signify the inducee and **B** the inducer of auditory continuity.

## *5.1 Homophonic auditory continuity*

In homophonic auditory continuity if three levels of the same sound (e.g. noise at 60, 70 and 80 dB SPL) are presented in succession and repeated without pauses, the 60 dB level sound appears to be continuous. The other sounds are heard as intermittent pulsed additions. Homophonic continuity is a subtractive process, such that when a sound with a level of 80 dB SPL is alternated with the same sound at a level of 82 dB SPL, the 82 dB SPL sound is heard as a weaker pulsed addition to the 80 dB SPL sound. The difference between the two levels can be

calculated as shown in Eq. 5-1, which leaves a residue, $dB_{\text{residue}}$, of 77.7 dB. That is, the residue is the level of noise required to increase the 80 dB SPL signal to 82 dB SPL (Bashford *et al.*, 1992).

$$dB_{\text{residue}} = 10 * \log_{10}\left[10^{82/10} - 10^{80/10}\right] \approx 77.7\text{dB} \qquad \textbf{Eq. 5-1}$$

### 5.1.1 The influence of signal level differences

Homophonic continuity is induced when a louder segment of a sound interrupts a quieter segment of the same sound. McAdams *et al*. (1998) investigated the level of the residual sound that is left when part of a louder sound is allocated to a quieter sound in the continuity illusion.

Subjects were presented with a diotic target sound that reproduced auditory continuity. This was a noise burst or tone that alternated between a high ($L_H$) and low ($L_L$) amplitude. The low amplitude signal was heard as continuous ($L_C$), with the remaining high amplitude component being heard as a separate intermittent stream ($L_I$). This is shown diagrammatically in Figure 5-1.

The stream consisted of eight low amplitude components with seven interleaved high amplitude components. $L_L$ was presented at 60±{1, 3, 5} dB SPL. $L_H$ was either 2, 6 or 10 dB greater than $L_L$. The durations, in milliseconds, of the components were varied in four different duty cycles: 100/100, 200/200, 100/300, 100/700, where the first value of each pair is the duration of the high amplitude segment and the second figure is the duration of the low amplitude segment, both in milliseconds. This is shown schematically in Figure 5-2.

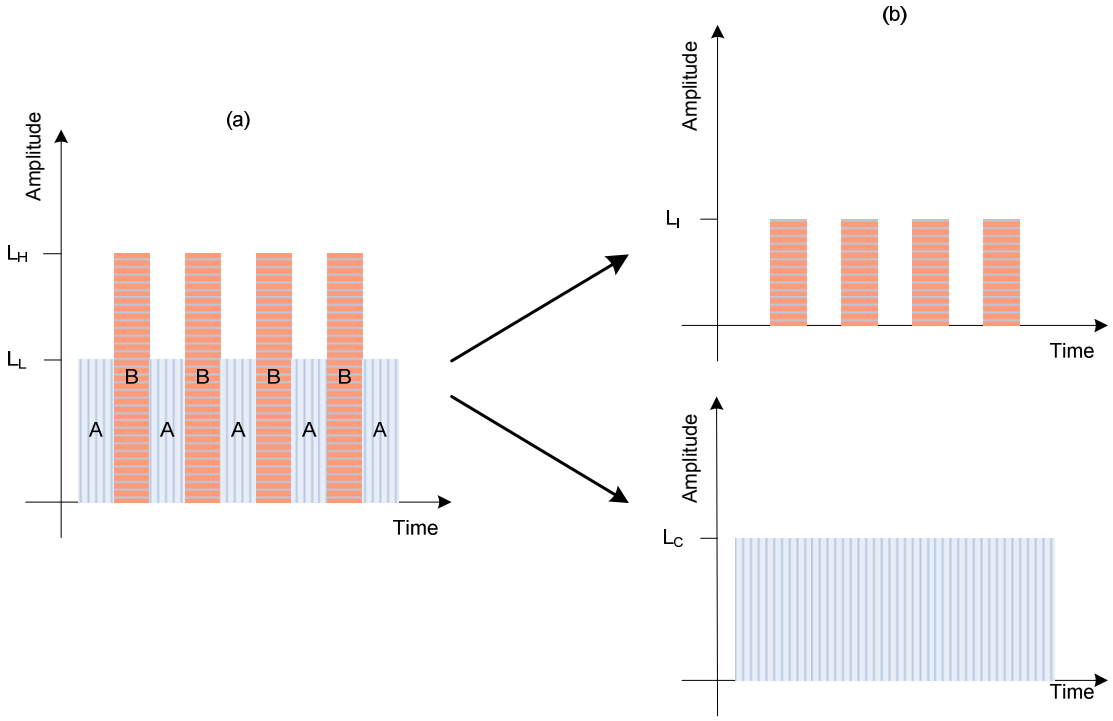



**Figure 5-1: Diagrammatic representation of (a) the actual stream presented to a listener and (b) the two perceived streams, as part of auditory continuity investigations by McAdams *et al.* (1998). Signal B induces continuity of signal A to produce a continuous signal with perceived amplitude $L_C$ and an additional intermittent signal having amplitude $L_I$.**

The reference stream, as depicted in Figure 5-2, was alternated with either an intermittent or continuous comparison signal. Subjects were asked to adjust the level of the comparison signal to match it with either the continuous ($L_C$) or intermittent residual ($L_I$) part of the reference stimulus. The experiment set out to investigate the perceived signal level that remained during auditory continuity. The results showed that the level did not match the expected levels calculated using loudness models or auditory segregation models, instead it lay between the theoretical predictions. Listeners tended to perceive a slightly quieter residual signal, compared to theoretical calculations, especially for small level differences between $L_H$ and $L_L$. In general the perceived level was within 2 dB of the calculated difference for continuous and intermittent comparison signals. When continuity was not induced, the level of quieter intermittent signal, $L_L$, was overestimated and the louder signal, $L_H$ was underestimated. The type of

reference signal, noise or tone, also influenced the perceived residue level for small level differences, although no clear pattern emerges from the results.

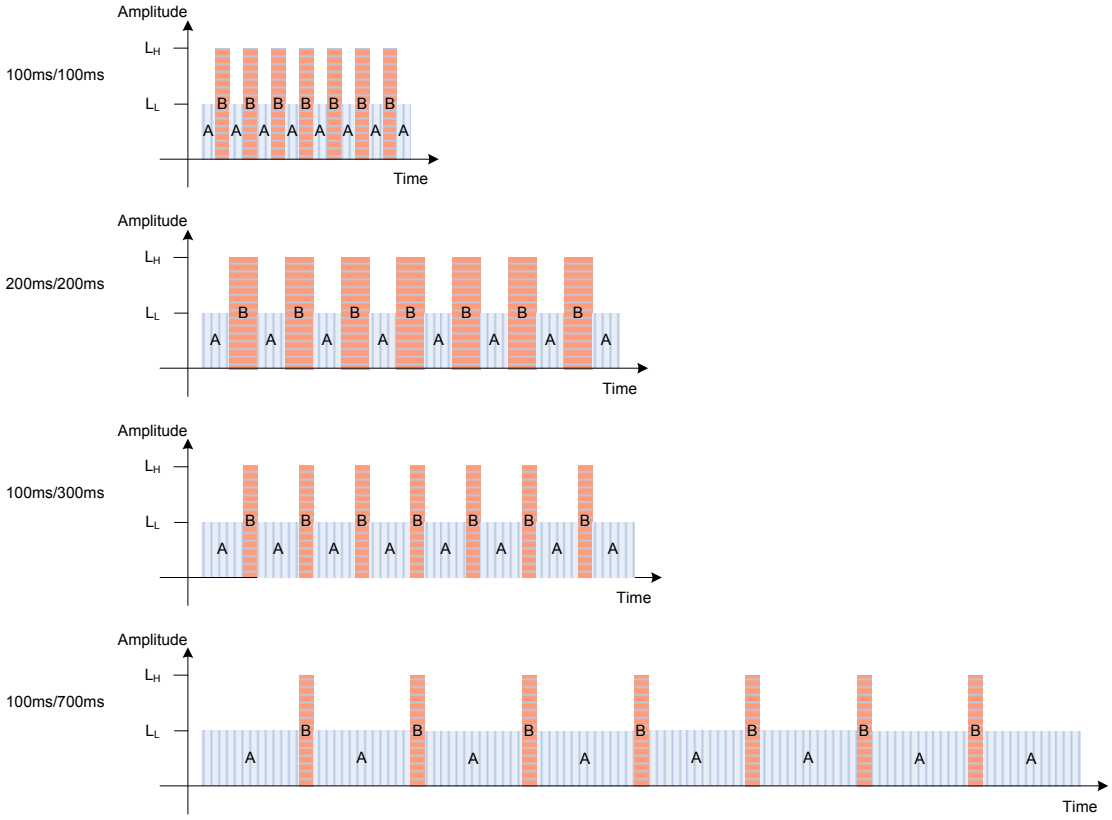



**Figure 5-2: The test signal duty cycles of high/low amplitude, used within the experiments by McAdams *et al.* (1998).**

## 5.1.2 The influence of signal duration

Drake and McAdams (1999) carried out experiments to investigate the influence of inducer and inducee signal duration when inducing the auditory continuity effect. Subjects were presented diotically with an alternating high and low amplitude 1 kHz tone in one of 12 different temporal configurations, shown in Figure 5-3.

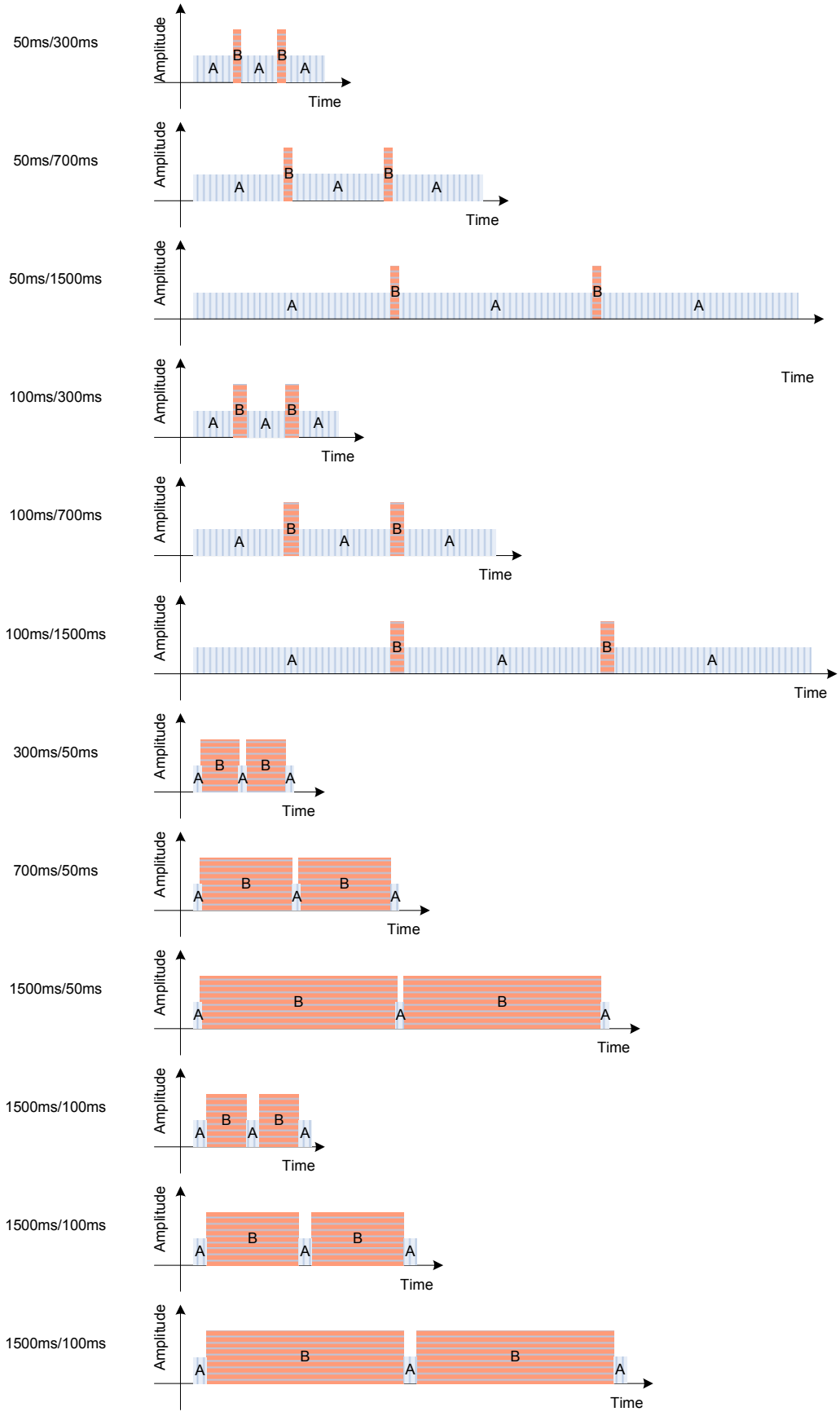



**Figure 5-3: The temporal configurations used in experiments by Drake and McAdams (1999).**

The first experiment investigated the temporal limits for which continuity was induced. The short tone had a duration of either 50 or 100 ms, whilst the long tone had a duration of 300, 700 or 1500 ms. In the normal configuration, in which auditory continuity would usually occur, the low amplitude inducee was the longer tone and the high amplitude inducer the shorter tone. The experiment also contained configurations in which the durations of the inducer and inducee were reversed. Their results confirm that the low amplitude tone must have the same or a longer duration than the high amplitude tone for continuity to occur. With longer low amplitude tones a stronger continuity phenomenon was perceived. In the second experiment they set out to investigate the influence of the ratio between the durations of the two tones and the total duty cycle duration. The duty cycles used for 1:2 and 2:1 were, 50/100, 100/200, 150/300, 300/600, 700/1400 ms and for 1:1 were, 100/100, 300/300, 700/700 ms. Their results show that the duty cycle duration did not influence the perception of continuity. That is, if continuity was not induced for a particular ratio, increasing the duration of both signals did not cause continuity to occur. They also found that, as expected, the ratios they tested of long, low amplitude tone to short high amplitude tone all produced continuity. In addition they found that duration ratios of 1:1 and even some of 2:1, where the high amplitude tone is twice the duration of the low amplitude tone, also produced the continuity phenomenon. They confirmed findings by Warren *et al*. (1994) that when the ratio is 2:1, continuity only occurs for high amplitude tone durations up to 300 ms.

## *5.2 Heterophonic auditory continuity*

Heterophonic auditory continuity involves two different alternating sounds of different levels. The fainter sound is heard as continuous if the louder sound is a potential masker of the fainter sound. The effect is strongly affected by the nature of preceding and subsequent sounds.

As an example, consider a sinusoidal signal **A** and a noise signal **B**, such that, when **A** and **B** are presented simultaneously under diotic conditions, **B** masks **A**.

Consider a new scenario in which both **A** and **B** are switched on and off intermittently. **A** is switched off during **B** and **B** is switched off during **A**. **A** resumes when **B** is finished with the phase it would have had if it had continued throughout **B**. This scenario is shown in Figure 5-4 (a).

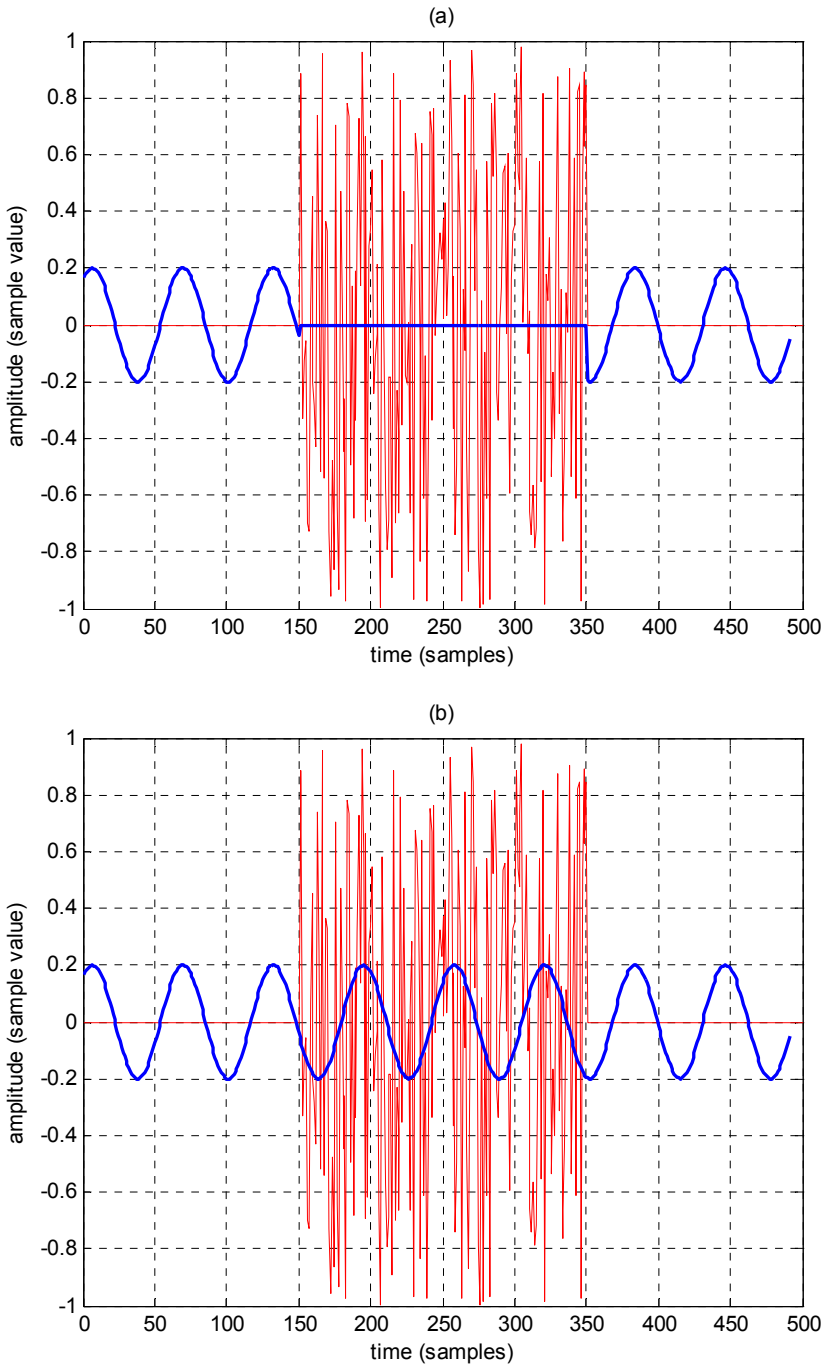



**Figure 5-4: Diagram of an example signal sequence where signal B masks signal A, and as a result induces continuity in A. In (a) the tone is stopped for the duration of the noise. (b) illustrates what is perceived when continuity is induced, the tone continuing through the noise.**

According to masking theory, when **B** is active, **A** should be inaudible. However, rather than only hearing signal **B**, **A** continues to be heard as though it had never stopped (shown conceptually in Figure 5-4 (b)). Due to masking, the hearing system cannot tell whether **A** is continuous or whether it ceased at the moment **B** started. Remarkably, the brain fills in the missing section based on its knowledge of **A** before and after the interruption. This is an example of the *continuity illusion* and it is a very powerful phenomenon which enables us to concentrate on a target sound in an auditory environment containing multiple distracting sounds (Bregman, 1990; Houtgast, 1972; Warren *et al*., 1972; Plack, 2000).

The relationship between masking potential and illusory continuity was expressed in the form of a rule by Houtgast (1972). This rule states that when a tone is alternated regularly with another sound, the tone is heard as continuous when a change from the tone to the louder sound involves no perceptible change in neural activity in any frequency region. Warren *et al*. (1970, 1972) have proposed a somewhat broader rule. This states that: "If there is contextual evidence that a sound may be present at a given time, and if the peripheral units stimulated by a louder sound include those which would be stimulated by a fainter sound then the fainter sound may be heard as present." They refer to this phenomenon as *auditory induction*.

The following sections discuss the factors that influence heterophonic auditory continuity, such as the spectral content of the inducee and inducer sounds and their spatial locations around a listener.

### 5.2.1 The influence of spectral content and frequency separation

Elfner and Homick (1967) carried out experiments to investigate how the continuity illusion is affected by the frequency separation between alternately

sounding tones. The tones were presented dichotically using the frequency pairings shown in Table 5-1.

| Tone **A** frequency (Hz) (inducee) | Tone **B** frequency (Hz) (inducer) | Tone **B** duration (ms) – dichotic |
|---|---|---|
| 200 | 300 | 27 |
| 250 | | 34 |
| 400 | | 25 |
| 500 | | 18 |
| 700 | | 17 |
| | | |
| 600 | 1000 | 10 |
| 800 | | 16 |
| 1500 | | 12 |
| 2000 | | 11 |
| 2500 | | 9 |
| | | |
| 2000 | 4000 | 9 |
| 3000 | | 10 |
| 4500 | | 16 |
| 5000 | | 13 |
| 6000 | | 13.5 |

**Table 5-1: Estimates of the continuity duration thresholds for experiments by Elfner and Homick (1967)**

Tone **A** was presented at 30 dB SPL and tone **B** at 45 dB SPL. Tone **A** had a duration of 250 ms for all tests. Estimates of the durations for tone **B**, read from graphical data, are shown in Table 5-1. The results show that as the frequency separation between tones **A** and **B** increases, the maximum duration of tone **B** that is able to induce continuity of tone **A** is reduced. Results were also obtained for gap detection within tone **A**, by replacing tone **B** with silence. The tone **A** frequencies tested were 400, 1500 and 4500 Hz, and the gap duration thresholds were found to be 2.5, 2.7 and 3.7 ms, respectively.

Elfner and Homick's results show that continuity is more likely to occur when the spectrum of a masking or inducing signal is identical or similar to the inducee. Further, continuity is more likely to occur if the signals are presented to the same ear. When each signal is presented separately to opposite ears, continuity is significantly affected. The gap duration thresholds also drop markedly and very short segments of missing audio can be detected.

Gap durations in noise were considered by Hall *et al*. (2007). Four 50 Hz wide bands of noise were presented individually or in a group, either comodulated or with random modulation applied to each. For the groups of noise bands, the silence was inserted into one of the bands with the remaining bands being continuous. For the scenarios marked as "all bands", all of the bands had a simultaneous segment of silence inserted. The four bands were centred around 1000 Hz, 1932 Hz, 3569 Hz and 6437 Hz. Gap detection thresholds were measured and are shown in Figure 5-5.

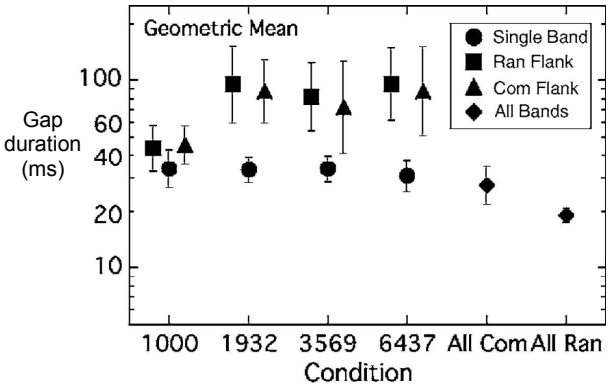



**Figure 5-5: Gap detection thresholds in ms for different noise configurations. Single narrow noise bands are denoted by the centre frequency. Groups of noise bands are marked as 'Com' for comodulated or 'Ran' for randomly modulated flanking bands. For these cases, the frequency denotes the band which contained the silence. 'All' denotes all bands had silence inserted. Taken from Hall *et al.* (2007)**

The results show that the addition of flanking noise bands produces an increase in gap detection thresholds, i.e. the subjects found it more difficult to detect gaps in the target band in these conditions. The effect is more pronounced for the three higher noise bands that were tested. For single and all comodulated bands

of noise the thresholds are similar at approximately 30 ms. Compared with the results of Elfner and Homick (1967), above, it suggests that perhaps as a signal becomes more complex, i.e. develops from a tone into narrowband noise, the more difficult it becomes for the auditory system to detect gaps in the signal. This trend continues as additional simultaneous noise bands are presented.

Kelly and Tew (2002) investigated the effect of spectral gating on the auditory continuity illusion. They used monaural, dichotic and binaural forms of signal presentation. The latter were spatialised using KEMAR measurements from the CIPIC HRTF database (CIPIC Interface Laboratory (1998)). Subjects were presented with a 1 kHz tone $a(n)$ interrupted by white noise $b(n)$, where $n$ is a time index. Each signal was separately spatialised to +/- 90º azimuth and 0º elevation. The spectral components, $A(n, k)$ and $B(n, k)$, identified by means of frequency index $k$, were calculated for $a(n)$ and $b(n)$, respectively using the discrete short-time Fourier Transform (STFT). Two sets of filter coefficients, $F_A(n, k)$ and $F_B(n, k)$, were generated according to the following conditions:

$$If\ |A(n,k)| > |B(n,k)|\ then\ \ F_A(n,k) = 1, F_B(n,k) = 0 \qquad \textbf{Eq. 5-2}$$

$$If\ |A(n,k)| \le |B(n,k)|\ then\ \ F_A(n,k) = 0, F_B(n,k) = 1 \qquad \textbf{Eq. 5-3}$$

A ‘1’ in these equations signifies that the filter has a unity gain at that frequency and a ‘0’ signifies that it has zero gain. These sets of spectral values were converted into two linear-phase time-domain filters, $F_A(n, k)$ and $F_B(n, k)$, by taking inverse discrete Fourier transforms and applying a windowing function. The filter coefficients were convolved with the corresponding input signals to produce spectrally gated output signals $a'(n)$ and $b'(n)$. The gating process is shown in Figure 5-6

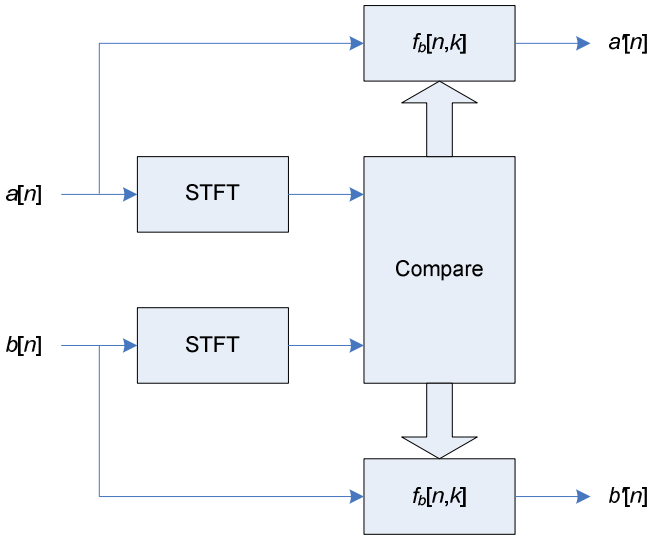



**Figure 5-6: The spectral gating process used in experiments by Kelly and Tew (2002)**

Subjects were presented with a test signal that had the temporal structure shown in Figure 5-7.

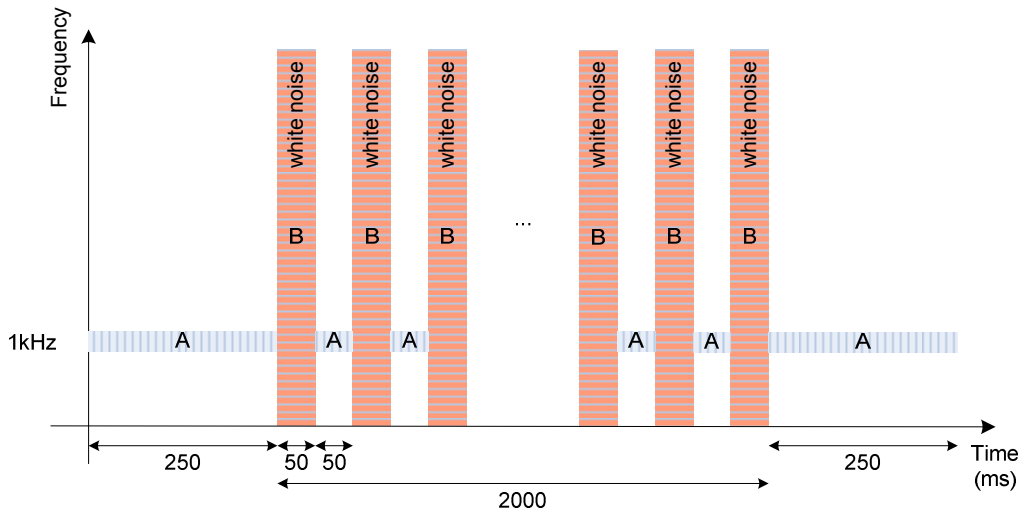



**Figure 5-7: Temporal structure of the test signals used in experiments by Kelly and Tew (2002).**

The tone and noise were initially set to 70 dBA SPL. The subjects were asked to adjust the level of the tone until continuity was perceived. A two-down, one-up procedure was repeated until the subject returned to the same tone level on three consecutive trials. Their results show that there is barely any difference between the continuity thresholds for simple pure-tone-in-noise signals and complex

spectrally gated signals. This suggests that the spectral gating approach satisfies the conditions for inducing continuity of the target sound.

### 5.2.2 The influence of spatial separation

Kashino and Warren (1996) investigated how interaural phase differences (IPDs) affect continuity. The inducee signal for these experiments was a 500 Hz tone, the inducer was 1/3-octave band-filtered noise centred at 500 Hz. The inducer was always presented at 70 dB SPL, the inducee was varied between 35 and 70 dB SPL. Both signals were 200 ms in duration. The signal configuration is shown in Figure 5-8.

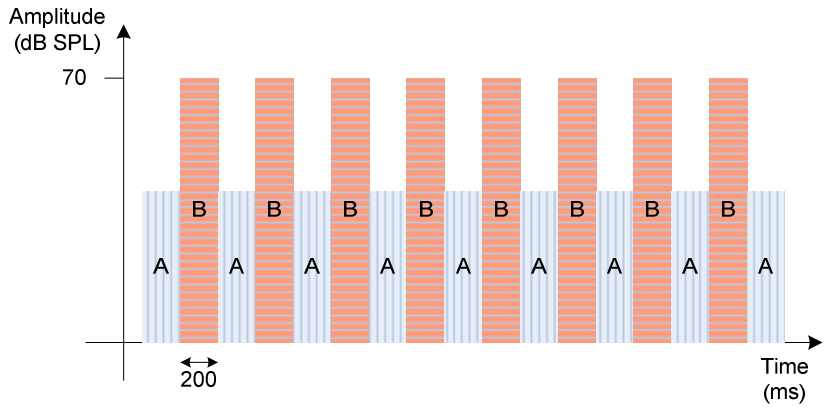



**Figure 5-8: The signals used in experiments by Kashino and Warren (1996).**

The phase of the target signal **A** and/or the masker signal **B** could be independently inverted at each ear, allowing different spatial locations to be simulated. The phase conditions tested were B0A0, B0Aπ, BπA0 and BπAπ. They found that continuity thresholds were 5-7 dB lower if one of the signals had a phase difference between each ear, compared with when both signals having the same phase at each ear. This is expected, as the inducer should mask the inducee more easily if the signals are perceived as being at the same spatial location. This was confirmed by Kashino and Warren in a second experiment in which they measured the binaural masking level differences for the same listeners used in the IPD experiment.

Elfner (1971) demonstrated a large increase in inducer signal durations before continuity fails when using free-field signals. The aim of this work was to investigate the influence of spatial separation on the continuity illusion for alternating tones. The tones had an angular separation of either 0º, 60º or 120º, centred around the 0º azimuth line in front of the listener. The trend for the frequency separation followed that found in previous work, described in Section 5.2.1. That is, the inducer is more effective if it has a similar frequency to the inducee. The configuration using a 0º separation produced the largest continuity thresholds. The approximate inducer tone duration thresholds for this configuration are shown in Table 5-2.

| Tone **A** frequency (Hz) (inducee) | Tone **B** frequency (Hz) (inducer) | Tone **B** duration (ms) 0º separation |
|---|---|---|
| 200 | 300 | 50 |
| 250 | | 92 |
| 400 | | 47 |
| 500 | | 31 |
| 700 | | 24 |
| | | |
| 600 | 1000 | 28 |
| 800 | | 45 |
| 1500 | | 33 |
| 2000 | | 23 |
| 2500 | | 18 |
| | | |
| 2000 | 4000 | 20 |
| 3000 | | 28 |
| 4500 | | 42 |
| 5000 | | 27 |
| 6000 | | 21 |

**Table 5-2: Maximum tone B (inducer) signal durations for inducing continuity of tone A (inducee), taken from Elfner (1971).**

A 60º separation produced thresholds that were approximately 5 ms less and 120º separation 10 ms less than those for a 0º separation.

Work by Darwin *et al*. (2002), has shown that binaural auditory continuity is influenced by ITD and not by spatial location. They presented subjects with a Huggins pitch target signal and a negative phase Schroeder complex signal. The Huggins pitch signal was generated using 300 Hz to 1 kHz bandpass-filtered white noise. This was used as the signal for one ear, the other ear received the same signal except that a linear phase change from 0 to 2π was applied across a 77 Hz band, centred around 500 Hz. Individually, these signals are heard as noise, but combined, they are heard as a weak pitch of 500 Hz on one side of the head and noise in the middle of the head. The Schroeder phase signal is a combination of frequencies with either a positive or negative phase relationship. Examples of positive and negative Schroeder phase signals are given in Figure 5-9, as described by Recio and Rhode (2000). It can be seen that one is the time reverse of the other.

The negative phase Schroeder complex was bandpass-filtered between 300 and 1400 Hz and presented diotically, so that it was perceived as a low frequency buzz in the middle of the head. Signals had durations of 120 ms and were presented in groups of 8 alternations with 10 ms crossfades. The Schroeder phase signal was used to induce continuity in the Huggins pitch target signal. Subjects were asked to listen for continuity either in the pitch signal, which could only be perceived using binaural hearing, or in the noise signal, which was essentially presented diotically. Darwin *et al*. found that the threshold of continuity, set as the level at which 50% of trials caused continuity, was approximately 2.4 dB lower when the target was the binaural pitch compared to the monaural noise. This implies that there is a binaural contribution to continuity.

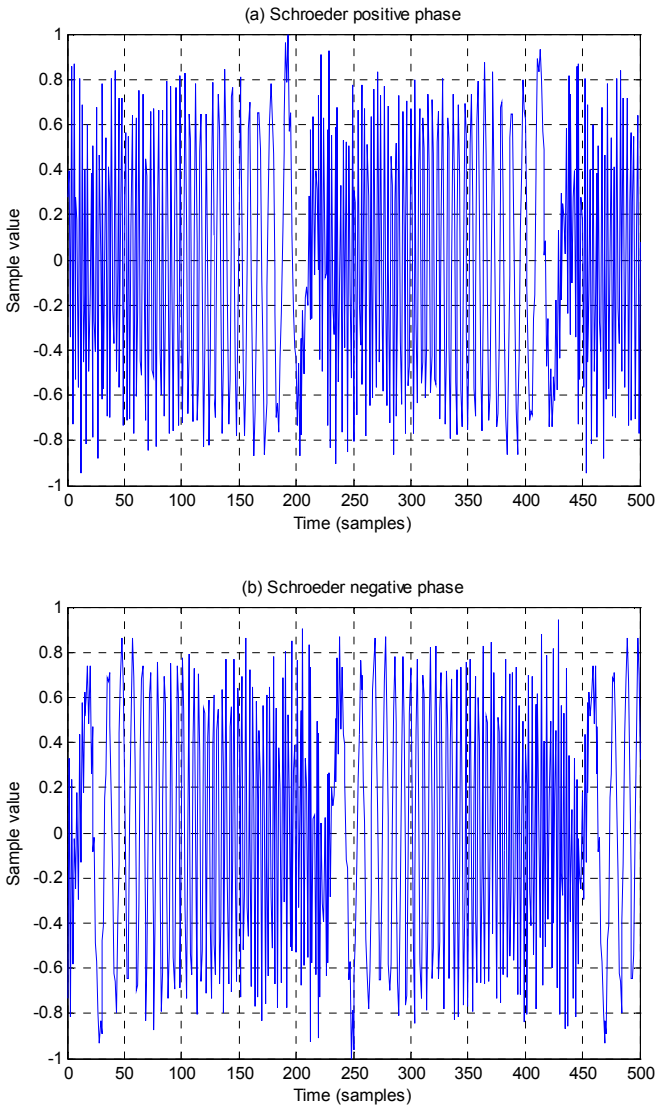



**Figure 5-9: (a) positive and (b) negative Schroeder phase signals. Both use harmonics 5 to 100 for a 200 Hz fundamental frequency. Generated from the description in Recio and Rhode (2000).**

In a second experiment described by Darwin *et al*. (2002), the Huggins pitch signal was replaced by a 500 Hz tone with an ITD of 300 µs. In this case the continuity threshold was increased from approximately -7 dB to -5.4 dB. Again, the negative phase Schroeder signal was used as the continuity inducer. Applying an appropriate ILD, which turned out to be 8 dB, the spatial image of the target could be brought back to the median plane. The results shown in Table 5-3 indicate that although the spatial location could be altered by adjusting either ITD or IID, the binaural benefit only occurred using ITD. For example, the trials using an ITD of 300 µs have a continuity threshold average of –5.4 dB, whereas

the trials that have a corresponding IID of 8 dB have an average continuity threshold of –1.3 dB.

| ITD (µs) | IID (dB) | Continuity Threshold (dB) |
|---|---|---|
| 300 | 0 | -5.4 |
| 0 | 0 | -3.8 |
| -300 | 0 | -5.2 |
| 300 | -8 | -3.6 |
| 0 | -8 | -2.0 |
| -300 | 8 | -3.6 |
| 0 | 8 | -1.3 |

**Table 5-3*:* Results from experimental work investigating the contribution of ITD to binaural continuity, taken from Darwin *et al.* (2002).**

The negative phase Schroeder inducer signal was perceived in the median plane, as in the previous experiment. Therefore, it was expected that it would more easily induce continuity in a target that is also in the median plane, rather than to one side (see Elfner and Homick (1967)). If the continuity of the target were affected by spatial location then the lines marked in bold should have the higher (less negative) continuity thresholds, as the target sound for those conditions would be perceived in the median plane. However, if ITD is the contributing factor then the lines marked in italics would have the higher continuity threshold. For the experimental conditions described it was found to be the ITD that affects the induction of continuity, rather than the perceived spatial location of the target. These results confirm the theory discussed in Section 2.4.2, which illustrates that lower frequencies are affected more by ITD than IID. The pitch of the target signal in the experiments described by Darwin *et al*. is 500 Hz and will therefore be influenced more by ITD than by IID.

### 5.2.3 Contralateral induction

Warren and Bashford (1976) investigated what they have termed *contralateral induction*. This can be interpreted as a special form of heterophonic continuity. A tone is presented to one ear and noise is presented to the other ear. They are then swapped around, with a duty cycle of 1 second, as shown in Figure 5-10.

The noise signal causes induction in the tone such that the noise is heard on one side of the head, but the tone is heard in the middle, and not on the other side of the head. When the locations of the sounds are swapped the noise is heard to move to the opposite side, but the tone remains approximately in the middle.

Bregman (1990) offers an explanation. The stream at each ear is an alternating sequence of the tone **A** and noise **B**. The sequence **A-B-A-B-A** is presented to one ear and its complement **B-A-B-A-B** is presented to the other ear. Heterophonic auditory continuity is induced in each ear as the tone is grouped with the relevant components in the noise. Therefore the tone will be heard as continuous in each ear. The tone is not heard precisely in the middle, however, as it is grouped with frequency components in the noise that have random phase and amplitudes. Therefore, the perceived location will change from moment to moment because of the changing interaural differences and an average location is detected.

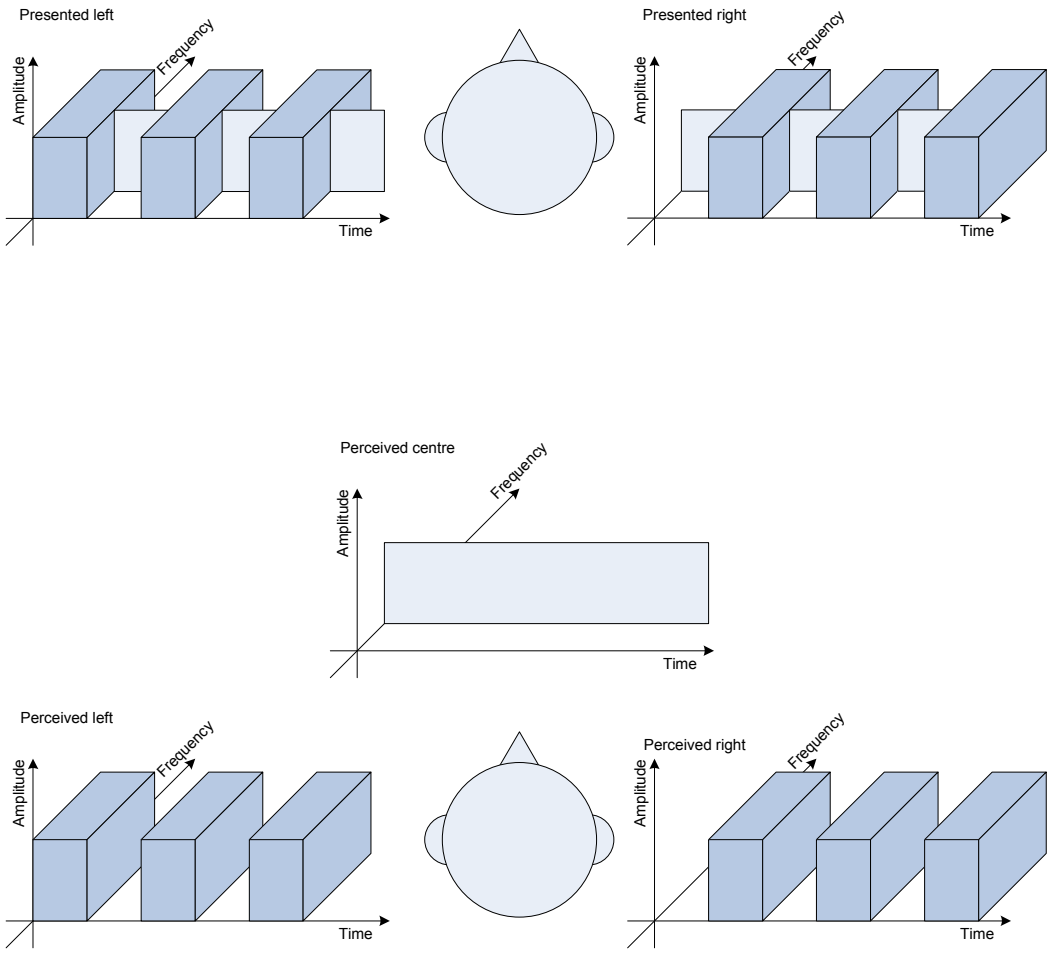



**Figure 5-10: The concept of contralateral induction. The left and right signals presented to a listener are heard as three components, a broadband noise that alternates between left and right ears and a steady tone that is heard approximately in the centre of the head.**

## *5.3 Phonemic restoration*

The intelligibility of a target speech signal broken by interruptions of silence improves when the silences are filled with noise (Bashford *et al*. (1992); Cherry and Wiley, (1967)). A speech target whose intelligibility has been impaired by the removal of a phoneme or even an entire syllable and then restored by the insertion of noise, displays *phonemic restoration* (Bashford *et al*., (1988); Bashford and Warren, (1979); Warren, (1970); Warren and Obusek, (1971)). However, investigations by Kewley-Port *et al*. (2007) have revealed that

intelligibility may actually be reduced by this practice for some listening conditions.

This section discusses the factors that influence the restoration of speech intelligibility when a target speech source is corrupted by missing temporal and/or spectral data. Different forms of target speech are considered, such as highly predictive sentences or random word lists using male and female talkers, along with a number of methods for perceptually recovering the missing audio data.

### 5.3.1 The spectral content of inducer signals

Bashford and Warren (1987) found that the characteristics of the noise that is used to interrupt speech affects the auditory continuity of the speech. Listeners were presented diotically with speech that was 1/8-octave bandpass-filtered at 1500 Hz. This was then interrupted with bursts of 1/3-octave bandpass-filtered noise or silence. The noise and speech were presented at the same level of 80 dBA SPL. The noise had a centre frequency of 375, 750, 1500, 3000 or 6000 Hz. The target and interrupter had the same duration, i.e. a 50/50 duty cycle. Listeners adjusted the duration of the interruption until they could not detect it, within a range of 20 ms to 1 s. The average interruption durations are given in Table 5-4.

| 1/3-octave noise band centre frequency (Hz) | Interruption duration threshold (ms) |
|---|---|
| silence | 79.2 |
| 375 | 135.7 |
| 750 | 220.7 |
| 1500 | 304.1 |
| 3000 | 129.2 |
| 6000 | 127.9 |

**Table 5-4: The minimum detectable gap durations for different centre frequencies of noise interferers, from an experiment by Bashford and Warren (1987).**

The results show that an interruption by silence requires the shortest duration before it is detected. The interruption that is most difficult to detect is the narrowband noise centred on the spectrum of the target speech signal.

Experiments by Samuel (1981) have shown that the type of signal that is inserted in the gaps can influence the success of the restoration. He found that white noise is better for restoring fricatives and tones are better for restoring vowels. In addition, short silent gaps increased the restoration of stop consonants. However, Miller and Licklider (1950) found that, for interruption rates up to 10 times a second, intelligibility was not improved when noise was inserted in regularly spaced gaps in sentences. This was a contradiction which Bashford *et al*. (1996) set out to resolve. Their experiments involved filling gaps in bandpass-filtered speech centred at 1500 Hz. They compared the effects upon intelligibility of replacing regularly spaced portions of a diotically presented male speech target signal with either white noise or speech-modulated noise. In both experiments the target was presented at 70 dBA SPL, with the noise at 70, 78 or 85 dBA SPL. In their first experiment the gaps in sentences were filled with noise modulated by the amplitude envelope of the deleted speech fragments. The duty cycle for the two signals was 400 ms, for which the speech target was on for 150 ms and then off for 250 ms. The use of speech-modulated noise produced twice the intelligibility increase obtained with interpolated white noise. For the second experiment the target was a list of monosyllables. The duty cycle was 400 ms again, but this time the target and noise were each on for 200 ms. The results show that, compared to silent gaps, insertion of speech-modulated noise increased intelligibility whereas white noise reduced intelligibility. They suggest that the overall intelligibility of an interruption condition is based on a trade-off between the extent of the temporal masking of the intact speech by the noise and the extent of the perceptual synthesis required for continuity. That is, although a white noise signal may induce continuity when inserted into a speech signal, it will also inflict backward and forward masking of useful regions of the speech. Speech-modulated noise improves contextual restoration whilst still inducing continuity, unlike silent gaps. Therefore, it is concluded that the white

noise did not provide sufficient information for perceptual restoration and generated excessive masking of the target.

To summarise, their results show that the best of their signals for filling gaps in band-limited speech was bandpass-filtered noise with a centre frequency matching that of the bandpass-filtered speech. They have therefore confirmed that the spectral similarity between inducer and inducee needed for inducing auditory continuity is also a requirement for the restoration of intelligibility. They go on to suggest that there are two levels of continuity. The first level induces apparent continuity of a signal. The second level groups the interrupted segments of speech into a sensible sentence. This is based on results of another experiment they performed, which indicate that the intelligibility of word lists did not improve when noise was inserted in gaps in the speech. The greatest improvement was found for highly predictive sentences, i.e. once the listener heard the sentence as being continuous, the insertion of noise in the gaps meant they were able to reconstruct the sentence based on their lexical knowledge. They go on to suggest that the results of Miller and Licklider are flawed as they used trained listeners who were used to hearing interrupted sounds and were able to achieve reasonable intelligibility scores irrespective of the signal used to fill the gaps.

Cherry and Wiley (1967) investigated the intelligibility restoration of a heavily band-limited 350 Hz-to-800 Hz speech signal. The speech target was amplitude gated so that only segments of high amplitude were reproduced. This frequency-limited, staccato sequence had white noise inserted into the gaps, which would otherwise have contained low amplitude speech. Intelligibility was totally restored using this method. They found that the level of the noise required for restoration varied over a 40 dB range between listeners.

Bashford *et al*. (1992) investigated the influence of the frequency content of an interrupting signal for restoring intelligibility. They describe two experiments, the first of which used male spoken sentences for the target signal and the second used male spoken word lists. In both experiments the target signal was presented at 70 dBA SPL and the interrupting noise at 80 dBA SPL, both diotically. The

speech on/off time was 175 ms/175 ms for the first experiment and 200 ms/200 ms for the second. The speech signal in the first case was bandpass-filtered centred at 1500 Hz with slopes of 48 dB per octave. The noise signal had energy in a 1/3-octave band centred at one of 375, 750, 1500, 3000 or 6000 Hz. Compared to using silence as the interrupter, the results from the first experiment show that intelligibility scores, based on percentage of words correctly identified, are greatest when the noise has a matching centre frequency to the speech and decrease as the centre frequency is moved further away. They therefore suggest that the spectral requirement for inducing continuity, discussed in Section 5.2.1, is also applicable to the restoration of intelligibility. The speech target signals were not band-limited for the second experiment and the noise had a broader bandwidth of 100 to 8000 Hz. Their results show that the percentage of correct words for word lists hardly improve when using noise interruptions compared to silence interruptions. However, a significant improvement was seen when high and low predictive sentences were used, indicating a contextual influence.

Warren *et al*. (1997) replaced spectral bands of a speech signal with noise and investigated how well listeners were able to restore the missing information using the unaffected bands. They band-limited a target speech signal into two narrow bands of 1/20-octave, centred at 370 Hz and 6000 Hz. A band-limited noise, spanning 600 Hz to 4 kHz, was inserted into the spectral gap and was played continuously with the band-limited speech signal. The signals were presented diotically over headphones. They varied the level of the noise and found that with the speech level set at 70 dBA SPL, a noise level of 60 dBA SPL gave the greatest intelligibility scores. They found that intelligibility significantly increased compared to the case with no band-limited noise inserted in the spectral gap. Their results are shown in Figure 5-11 which also indicate that the benefit of the noise insertion increases depending on the predictability of the sentences.

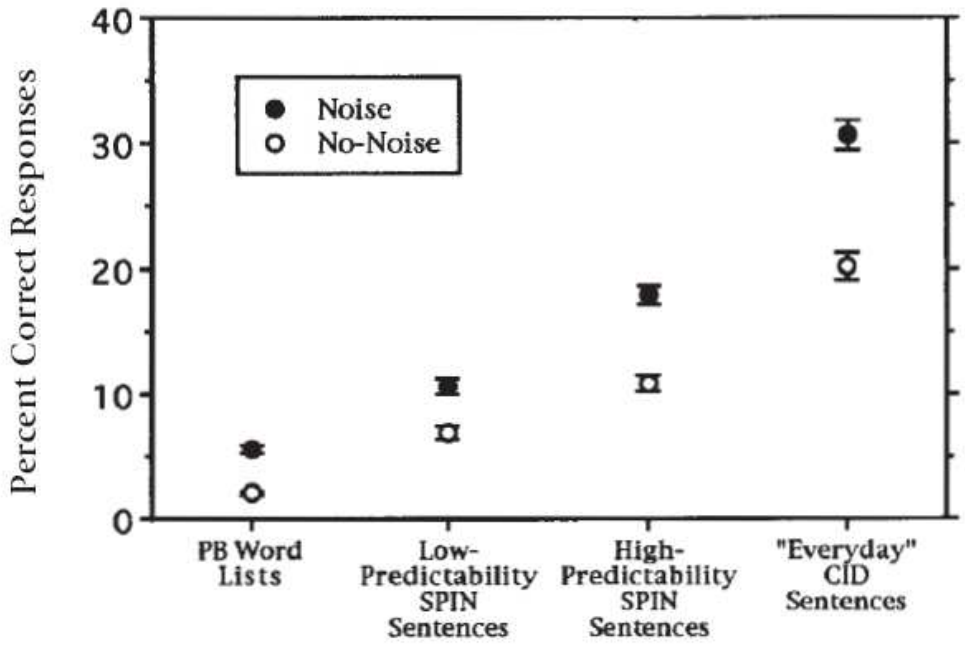



**Figure 5-11: The intelligibility scores for four types of sentence with and without narrowband noise inserted into spectral gaps. Taken from Warren *et al.* (1997)**

### 5.3.2 The relevance of target speech content

Bashford and Warren (1987) found that the context of the target speech signal is important. They used wideband target speech and a wideband noise interrupter. The target speech signal was either normal discourse, backward discourse or a stream of monosyllabic words. The average durations of interruption that could be detected are given in Table 5-5.

| Target signal | Noise interrupter (ms) | Silence interrupter (ms) |
|---|---|---|
| Normal discourse | 304.4 | 52.0 |
| Backward discourse | 147.7 | 50.0 |
| Monosyllabic words | 161.5 | 61.1 |

**Table 5-5: The detectable gap durations for different contexts of target speech signals, used in experiments by Bashford and Warren (1987).**

These results show that the context of the target signal has an influence on the interruption duration detection threshold. For the normal discourse interrupted

by the wideband noise it can be seen that there is no significant difference compared to using bandlimited noise in their previous experiment described in Section 5.3.1.

### 5.3.3 The influence of spatial separation of auditory objects

Shinn-Cunningham and Wang (2008) considered the influence of auditory object formation on phonemic restoration. They specifically looked at the restoration of speech when periodic gaps are filled with modulated noise compared to unmodulated noise and silence. Target speech was presented with interruption rates of 1.5, 2.2 or 3 Hz, with a 50:50 duty cycle. The noise and speech were either presented diotically, so that both sounds were perceived in the middle of the head, or the noise had an ITD of 300 μs to give the perception that it was located to one side of the listener. Figure 5-12 shows the percentage of correctly identified keywords for the condition that the speech and noise are collocated in the middle of the head.

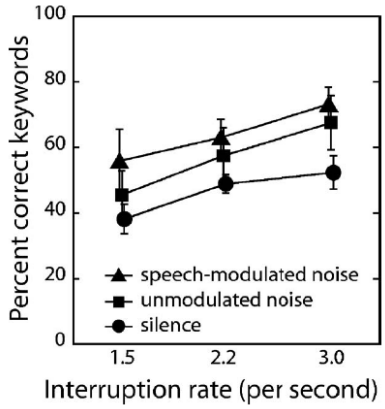



**Figure 5-12: The percentage of keywords correctly identified in relation to the interruption rate for 3 interruption types. Taken from Shinn-Cunningham and Wang (2008).**

The speech-modulated noise provides the best phonemic restoration across the interruption rates tested. Shinn-Cunningham and Wang (2008) raise the following key points based on these results. Firstly, the speech modulation is based on the missing segment of speech, but it only provides simple amplitude

modulation cues and no indication of the spectral content that is missing. It is therefore surprising that such a minor cue is used so effectively by listeners to restore the missing segments. Secondly, listeners reported that the modulations were perceived as part of the noise and not attributed to the speech. However, the modulations still provided an improvement over unmodulated white noise. The results for the spatially separated configuration are shown in Figure 5-13.

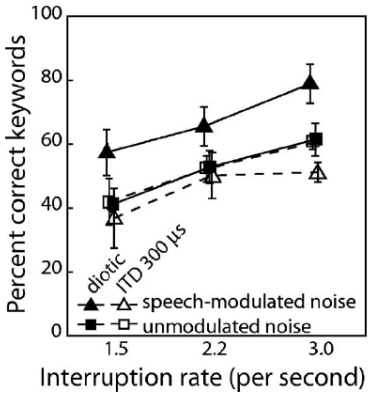



**Figure 5-13: The percentage of keywords correctly identified for spatially separated noise interruption and target speech for 3 interruption rates. Taken from Shinn-Cunningham and Wang (2008).**

It may be seen from Figure 5-13 that the speech-modulated noise loses its effectiveness at restoring the target speech when it is spatially separated. However, the unmodulated noise maintains its effectiveness. Shinn-Cunningham and Wang suggest that this is because the noise signals are more easily grouped into separate auditory objects when they are spatially separated from the target speech. The restoration of the target is based on the masking ability of the interferer sound. Therefore, the unmodulated noise achieves better correct word scores as it is the more effective masker. In fact, the spatially separated modulated noise appears to be only marginally better than the silent interruptions of the target signal shown in Figure 5-12.

### 5.3.4 Methods for improving phonemic restoration

Smith and Faulkner (2006) investigated the impact of missing spectral regions on hearing-impaired listeners. Their research reveals that missing spectral regions significantly affect intelligibility. Experiments were carried out on normal-hearing listeners with a simulated band of frequency loss corresponding to a damaged region of the cochlea. The speech signals were passed through a bank of bandpass filters and the amplitude envelopes were extracted for each band. Each envelope was used to modulate a corresponding bandpass-filtered white noise signal. The noise bands were then summed to produce a signal where the temporal and amplitude cues were maintained, but the spectral detail had been removed. Correct word scores were measured for four processing conditions, shown in Figure 5-14. The 'matched' condition had no frequency information missing, the 'dropped' condition had a frequency band from 424 Hz to 2182 Hz removed. The 'A-warp' and 'S-warp' conditions had the information from the missing band mixed in two different ways into higher and lower bands, as shown in Figure 5-14. The benefit of this frequency warping was measured along with the benefits of training the subjects to cope with the spectral loss.

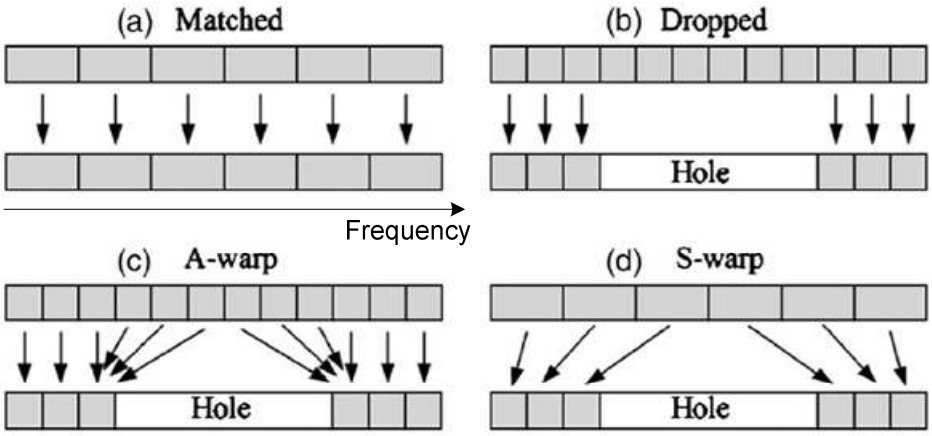



**Figure 5-14: The four spectral processing conditions implemented and evaluated by Smith and Faulkner (2006).**

The frequency warping significantly increased correct word scores for sentences and vowels, but less so for consonants. Table 5-6 summarises the results.

| | Matched | Dropped | A-warp | S-warp |
|---|---|---|---|---|
| Sentences | 83% | 34% | 59% | 70% |
| Vowels | 77% | 41% | 61% | 70% |
| Consonants | 75% | 67% | 69% | 70% |

**Table 5-6: The percentage correct word scores for different spectral processing methods taken from Smith and Faulkner (2006).**

These results show the benefit of the S-warping method used for moving the information from a damaged area of the cochlea into adjacent frequency bands. Smith and Faulkner (2006) also found that by training their subjects there was a performance improvement for all three spectral processing methods. The methods based on warping still achieved better performance compared to the method where the missing spectral components were simply dropped.

Srinivasan and Wang (2005) considered a top-down schema-based approach to phonemic restoration. They discuss the performance issues of solutions based on auditory continuity for restoring missing segments in speech. Restoration models based on inducing continuity are unable to restore phonemes that have no continuity with the remaining adjacent phonemes. Furthermore, spectral filtering or interpolation schemes tend to restore missing segments of unvoiced speech incorrectly. The model described by Srinivasan and Wang (2005) uses a missing data automatic speech recognition (ASR) processing stage. This provides a spectral template of the corrupted word. The time-frequency components of the spectral template are used to restore the corresponding corrupted components in the original speech. The spectral template provides a generic representation of the word that may not match the characteristics of the uncorrupted segments of the speech. Therefore, post processing is applied to smooth the pitch and the transitions between segments. Time warping is used to match the synthesised speech to the articulation rate of the original speech.

## 5.4 Summary

This chapter has discussed the causes and effects of auditory continuity. The close link between masking, discussed in Chapter 4, and auditory continuity imposes the requirement that a sound which induces continuity in a target sound must also be capable of masking the target sound.

In homophonic auditory continuity, described in Section 5.1, three levels of the same sound are presented in succession and repeated without pauses, the quietest of the sounds appears to be continuous. The other sounds are heard as intermittent pulsed additions. The perceived level of the residual sound is related to, but does not match exactly the level difference between the inducer sound and the target sound. Typically, for auditory continuity to occur, a low-level target sound is interrupted by a shorter, louder sound. However, in some cases, continuity is still perceived when the duration of the interrupter before and after the target is longer than the duration of the target sound.

Although the sound sources and configurations described in this chapter do not represent everyday listening situations, these simple measures of continuity performance certainly demonstrate that the human auditory system has a sophisticated mechanism for dealing with interruptions of a target sound source. However, the selection of appropriate inducer sounds is critical. The duration and level of the inducer must be chosen to ensure continuity of the target sound is induced, but without introducing excessive or unnatural residual sounds.

The human hearing system is able to detect silent gaps in tones for interruptions as short as 2 or 3 ms. As the bandwidth of the signal increases it becomes more difficult to hear the interruptions. For example, Section 5.2.1 shows that a noise with a bandwidth of 50 Hz requires an interruption of approximately 30 ms before it can be heard. The interruption becomes harder to hear when the original noise band is presented with additional comodulated bands of noise centred on other frequencies. This creates temporarily missing frequency components (spectral holes) rather than a complete absence of the signal. It

follows, therefore, that provided the spectral holes in a signal are small, the perception of that signal will not be significantly affected.

The spatial locations of the target and inducer sounds also influence whether auditory continuity is perceived. A direct link with auditory masking is observed. That is, a louder sound will more easily mask a quieter sound if they are collocated than if they are spatially separated. Therefore, the same limitations apply for auditory continuity. Section 5.2.2 illustrates that as the spatial separation between the sounds is increased, by the manipulation of ITD, interruptions are more easily detected and the inducer sound becomes less effective at invoking continuity.

Section 5.2.3 discusses the auditory continuity illusion when the target and inducer sounds are presented dichotically, one to each ear. This highlights that the continuity of a sound in one ear can be induced by a sound in the opposite ear. Alternatively, this can be viewed as dual streams of independent continuity within each ear which then become fused into a single sound object.

It is noteworthy that interruptions are more difficult to detect in speech compared to tones and bands of noise, as discussed in Section 5.2. If the interruptions are filled with noise that matches the bandwidth of the speech signal they become even more difficult to detect. That is, there is a strong link between the spectral content of the interrupting sound and the target speech sound. It is more difficult to detect the interruption if it has a similar spectral content to the target sound. It follows, therefore, that interruptions of silence are the easiest to detect.

The successful temporal restoration of intelligibility, by filling interruptions with noise, is limited to highly predictive sentences. The additional temporal masking caused by an interruption of white noise can reduce the intelligibility of random word lists compared to silent interruptions. This emphasises the importance of contextual information within the sentence when restoring intelligibility.

The spectral restoration of intelligibility, by filling missing frequency bands with noise, is very effective. When only narrow bands of speech are available to a

listener, the intelligibility is low, in particular if the missing bands contain a substantial proportion of the speech energy.

The brief discussion on methods for improving speech intelligibility, when spectral information is missing, has shown that spectral holes in the signal with energy should not be filled blindly in the hope that the hearing system will smooth across the interruptions. For example, speech has spectro-temporal regions of low energy which are important for the correct understanding of the words. Introducing excessive energy in these regions could alter the meaning of the words or worse still mask neighbouring regions of intact signal. Instead, when spectral holes are filled, consideration should be given to adjacent spectral activity, perhaps by using the temporal envelope and energy fluctuations in frequency bands that have not been disrupted. Thus, it is likely that careful consideration of the target speech sound and the interfering sounds will lead to a more effective and robust solution for improving speech intelligibility. This is discussed further in the next chapter, which considers how speech intelligibility is affected by multiple interfering sound sources.

# Chapter 6 Speech intelligibility with multiple sound sources

Up to this point, the literature review has focussed on auditory masking and its relationship with auditory continuity. Chapter 5 discussed how auditory continuity can improve intelligibility by invoking phonemic restoration. This section probes more deeply into the areas of research related explicitly to speech intelligibility in the presence of multiple sound sources. Intelligible is defined as "able to be understood" (Oxford Dictionary, 1995). Although they are linked, audio quality and intelligibility are not the same. That is, it is possible to have low quality audio where every word is understood. This distinction will become important for the technical development discussed in Chapter 7.

This thesis is concerned with the intelligibility of a target sound source in a complex auditory environment of multiple sound sources. If it were possible to extract the signal components for the target sound source entirely and exclusively from the mixture, the result would be the best that can possibly be achieved, i.e. the extracted sound would be identical to the target alone. However, despite a tremendous effort from a number of research groups, target extraction using techniques such as computational auditory scene analysis (CASA) has only been achieved under special conditions and is not easily attainable in real-time. Consequently, implementing pure CASA, for example, in a hearing aid application would be impractical due to the limited processing power available. Therefore, alternative simpler methods must be considered for resolving this issue.

The evidence so far has shown that listeners can more easily concentrate on a target signal that is at a different location from a masking or interfering signal, than on a target that lies in the same direction as the interferer. It has also been shown that our ability to understand a signal interrupted by gaps improves when the gaps are filled with signals that induce continuity. The next section discusses the effect of factors such as these on the intelligibility of a target speech sound

source in the presence of multiple simultaneous interfering sound sources and how they may be separated.

## 6.1 Blind Source Separation

This section begins by defining the blind source separation (BSS) problem. It goes on to describe practical means by which a mixture of sources may be separated into its constituent parts. The aim of BSS is to achieve source separation using only information contained in the mixture itself. 'Blind' means that very little is known about the mixing matrix of the sources. BSS can be considered as an instantaneous unmixing problem, which does not consider the convolutive effects of the sound source and its acoustic environment. These factors are discussed later in Section 6.1.1. Furthermore it does not consider sound sources that are moving relative to the sensors. A simple example of BSS is represented diagrammatically in Figure 6-1 for the case of two sources and two sensors.

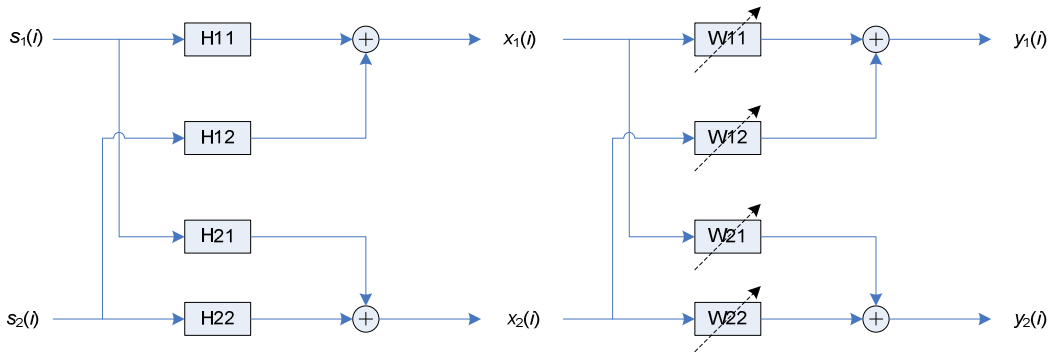



**Figure 6-1: Block diagram showing the mixing matrix for two sound sources with corresponding estimate of unmixing matrix to extract original sound sources.**

The sources $s_1(i)$ and $s_2(i)$ (where $0 \leq i \leq M\text{-}1$ and $M$ is the total number of samples in each signal) are mixed through the mixing matrix **H**. The outputs $x_1(i)$ and $x_2(i)$ simulate the input signals to a pair of sensors, which in this application are microphones. The naming convention used is summarised in Appendix E. It is assumed in this model that there is an insignificant amount of noise or distortion added to the signals by the sensors. In the unmixing process the sensor

signals $x_1(i)$ and $x_2(i)$ are passed through a matrix **W** to produce output signals $y_1(i)$ and $y_2(i)$. The goal of BSS is to estimate the coefficients of **W** such that $y_1(i)$ and $y_2(i)$ are an accurate estimate of the original sound sources $s_1(i)$ and $s_2(i)$, respectively. However, in a real acoustic environment the problem is convolutive and must take into account the delays and reflections of the sounds. In general, for $Q$ microphones, this can be written as (Davies *et al*., 2003):

$$x_q(i) = \sum_{p=1}^{P} \sum_{n=0}^{N-1} a_{qp}(n) s_p(i-n), \qquad q = 1,...,Q \qquad \textbf{Eq. 6-1}$$

Where $x_q(i)$ is the $q$th microphone signal at time sample $i$, $s_p(i)$ is the $p$th source signal, $a_{qp}(n)$ represents the impulse response from source $p$ to microphone $q$ and $N$ denotes the maximum length of the impulse responses.

Recovery of the original source signals given only the mixtures $x_q(i)$ is usually achieved by estimating a matrix of unmixing filters $w_{pq}(i)$, (Davies *et al*., 2003).

$$y_p(i) = \sum_{q=1}^{Q} \sum_{u=0}^{U-1} w_{pq}(u) x_q(i-u), \qquad p = 1,...,P \qquad \textbf{Eq. 6-2}$$

Where $y_p(i)$ is an estimate of a sound source $s$ and $U$ is the length of the unmixing filters. This theoretical modelling of the BSS problem does not include a representation of background noise or system noise and assumes the sources are stationary.

The time-domain representation is computationally expensive to implement due to the potentially large FIR filters. It is therefore more efficient to convert to a frequency domain system. The time-frequency components can be represented in a matrix form to give:

$$\mathbf{Y} = \mathbf{W}\,\mathbf{X} \qquad \textbf{Eq. 6-3}$$

Where matrix **X** represents the time-frequency components for the microphone signals and matrix Y contains the recovered source signals using the estimated unmixing matrix **W**.

There are several common methods for performing BSS. Hyvärinen *et al.* (2001) discuss the difference between BSS, an application that can be solved in many ways, and the theoretical methods used for the solutions. These methods include the use of, but are not limited to, independent component analysis, spectral and temporal correlation, frequency information, neural networks and beamforming. There are also many variants and combinations of these techniques.

Three methods are discussed next, along with some typical applications. These have been selected based on their popularity as baseline techniques for solving BSS and their relevance towards the technical work discussed later in this thesis.

### 6.1.1 Independent Component Analysis

The application of Independent Component Analysis (ICA) to the BSS problem described above requires some assumptions to be made (Douglas and Gupta, 2007). Firstly, the sound sources, denoted independent components, must be statistically independent of each other. For audio applications this is not such an unreasonable assumption to make. It is unlikely, for example, that two highly correlated sounds will be active simultaneously. If that situation did occur it would most likely be acceptable to group them together into a single sound source component. Secondly, the independent components must have non-Gaussian distributions. ICA fails if the observed variables have Gaussian distributions. Finally, it is assumed that the number of independent components is equal to the number of observed mixtures.

One of the practical implementations of ICA is in the Fourier domain. This generally utilises a short-time window to resolve temporal variations in the spectrum (Barry *et al.* 2005). The ICA model is applied to each frequency bin. This approach is not straightforward for several reasons. For example, the order and polarities of the separated sources can change from one analysis frame to the

next. To reconstruct accurately a source signal in the time domain it is necessary to know all of its frequency components. It is therefore essential to have a method of selecting which spectral components belong to each source, this is known as the permutation problem. As a result, many techniques have been investigated to assign continuity criteria to the extracted components.

The use of ICA for BSS has also been applied using a single channel of audio by Barry *et al*. (2005). They have successfully separated a flute and bass sound that were active simultaneously, but with minimal spectral overlap. This highlights how the underlying ICA concept can be varied and successfully applied to new use cases.

The performance of ICA when applied to BSS has been discussed by Shirley and Kendrick (2008). They have found that different microphone and algorithm configurations produce a variation in performance. For example, microphones that are spatially separated require processing of the time lag between the signals received at each microphone. Furthermore, performance is impaired when a reverberant environment or moving sound sources are considered, Talantzis *et al*. (2006). Shirley and Kendrick (2008) suggest this can be overcome by using more microphones (though this is not appropriate for the proposed hearing aid use case). They go on to suggest that a coincident pair of microphones is similar to beamforming for two sound sources. This will now be considered further.

### 6.1.2 Beamforming

The concept of beamforming refers to multichannel signal processing techniques that preserve the acoustic signals coming from a particular known position, and reduce the amplitude of signals coming from other directions. Beamforming is commonly used in an automotive environment, in particular to facilitate a hands free phone call in a vehicle.

One of the techniques used in beamforming is linearly constrained minimum variance (LCMV), Bourgeois and Minker (2009). It is a multiple-input-single-output (MISO) problem. The coefficients of the system filters are set such that the amplitude of the interference signals is minimised with respect to the target

signal. This contains an element of spatial filtering and spectral filtering. The coefficient delays can be synchronised for a particular direction of arrival, allowing the multiple signals from this direction to reinforce each other. Hence, the "beam" of the system is said to be "steered" in the direction of the target. The physical constraints of the microphone positions result in sidelobes, which can be further reduced using generalised sidelobe cancellation techniques (Allred, 2006). This can be restrictive when multiple occupants would like to participate in the phone call. In this case, further microphones are placed close to each of the additional occupants within the vehicle cabin. Although the approach is being used for premium in-car solutions, it is unsuitable for the more popular aftermarket hands-free systems.

The specific scenarios described and discussed by Bourgeois and Minker (2009) are directed towards a system that enhances an acoustic environment with the aim of extracting a target speech source. The resulting signal is transmitted over the air for a listener at the receiving end of a telephone connection. The listener is separate from the talker and is purely interested in the intelligibility of the speech to maintain a conversation. This is quite different to two or more people maintaining a conversation within the same listening conditions. A listener that shares the same auditory environment as the talker can make use of additional cues, such as head movement and visual information. Furthermore, they are advantaged by having the use of both ears and an unrestricted or "full" audio bandwidth. This is unlike a typical telephone conversation where the audio bandwidth is limited to approximately 3.5kHz for narrowband and 7kHz for wideband connections, thus reducing the available spectral information for the listener.

A physical limitation of the use of beamforming for a hearing aid based application is the restriction in the number and location of microphones in the system. Allred (2006) has shown that the performance of a beamforming system is good until the number of sound sources exceeds the number of microphones. In everyday listening environments this will be a regular occurrence.

### 6.1.3 Spectral processing

Alternative methods for processing the signals from only two microphones include spectral analysis of the signals. One of these spectral approaches used for solving BSS, and related to the work described later in this thesis, is generalised cross-correlation with phase transform (GCC-PHAT) (Knapp and Carter, 1976). This approach considers the correlation between the signals to determine the dominant signals for each direction of interest. This is used to provide an estimate of the direction of arrival of a sound. It allows sounds to be loosely distinguished based on their spatial location. If two digital signals $x_1(n)$ and $x_2(n)$ are acquired from two microphones, then GCC-PHAT is defined as:

$$l = \mathrm{F}_\mathrm{D}^{-1}\left\{\frac{\mathrm{F}_\mathrm{D}[x_1(n)]\mathrm{F}_\mathrm{D}[x_2(n)]^*}{\left|\mathrm{F}_\mathrm{D}[x_1(n)]\mathrm{F}_\mathrm{D}[x_2(n)]^*\right|}\right\} \qquad \textbf{Eq. 6-4}$$

Where $\mathrm{F_D}$ is the forward DFT, $\mathrm{F_D^{-1}}$ is the inverse DFT and $l$ is the time lag. The distance between the microphones determines the maximum delay that can be measured. Room reflections will disrupt the accuracy of this simple coherence method (Brutti, 2008).

The use of spectral information to determine relative attenuation and delay is the basis for the DUET (Degenerate Unmixing Estimation Technique) algorithm, Rickard (2007). It is also part of the ADRess (Azimuth Discrimination and Resynthesis) algorithm developed by Cooney *et al.* (2006).

### 6.1.4 Summary

This section has given a brief overview of BSS and some of the methods used to help solve the problem of separating a mixture of sounds into single components. This is useful for this research as it would provide an ideal starting point for the manipulation of the auditory scene. Having perfectly separated sound sources, or even a very good estimate, would allow the sounds to be respatialised very effectively. However, the aim of the research considered in this thesis is to investigate a solution for a low power wearable device such as a hearing aid. It is therefore important also to consider the practicalities of implementing such a

system. This most often results in a tradeoff between a theoretical or ideal solution and a compromised practical solution. Not only must the memory footprint and processing cycles of the algorithm be considered but also its responsiveness to real time stimulus. A large, highly accurate algorithm may be extremely good at processing pre-captured files of audio recordings, such as a neural network approach. However in a real portable system this may not be acceptable. Furthermore, the target processing core for a low power device may have a shorter data width, such as 8 or 16 bits compared to 24 or 32 bits for a high power device.

The techniques described in this section are generally too complex to implement or too processing intense for current low power DSP devices. Das *et al.* (2007) have estimated the complexity order of ICA based approaches as a cubic polynomial. However, in the future it is expected that these methods will become more practical due to algorithmic optimisations and improved digital core design.

## *6.2 The influence of spectral content*

A number of researchers have investigated the influence of different types of interference signals on the intelligibility of target speech signals. The research discussed here covers single and multiple speech signals and the spectral content of the target and interfering sounds. This section also examines the influence of gaps in the interferer that allow the target signal to be heard more easily, known as the 'peek' theory.

Warren *et al*. (1995) investigated the role of spectral redundancy in relation to speech intelligibility. Subjects were presented with male spoken sentences in English. The sentences were bandpass-filtered using 1/3-octave filters with a slope of 98 dB per octave, at centre frequencies of 370, 530, 750, 1100, 1500, 2100, 3000, 4200 and 6000 Hz. A single band of the speech was presented diotically over headphones at 70 dBA SPL. A second test group was presented with the two extreme bands, 370 and 6000 Hz, either dichotically with one band

sent to each headphone or summed and diotically. The results for the group presented with a single frequency band, are shown in Figure 6-2.

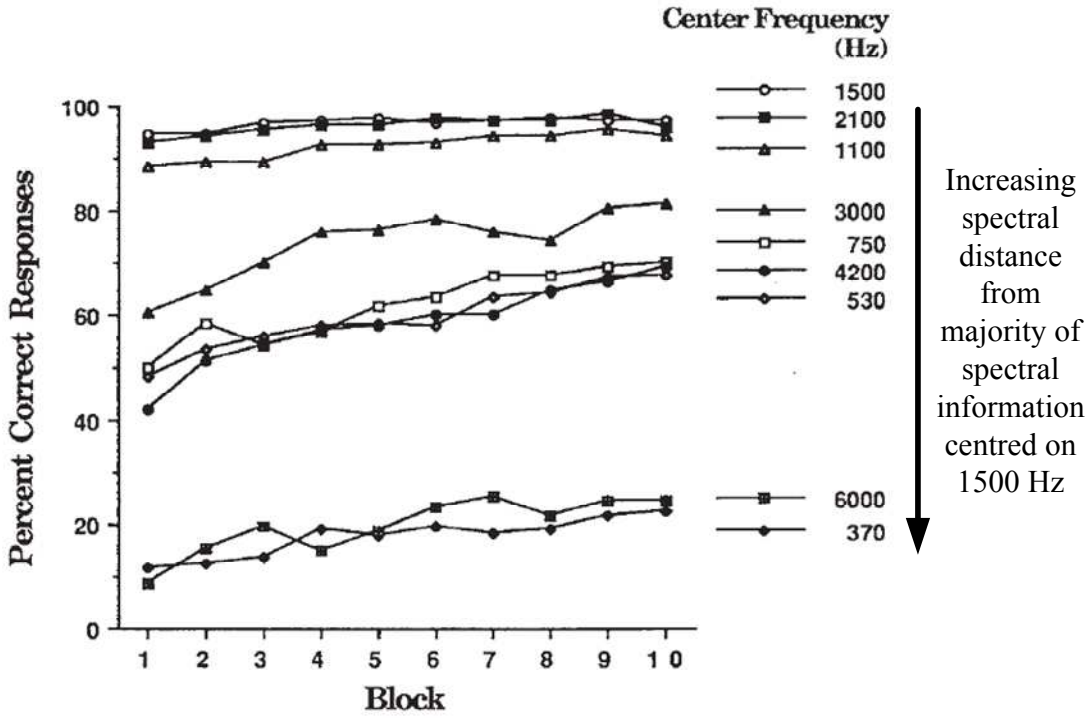



**Figure 6-2: The intelligibility results from the spectral reduction experiment by Warren *et al.* (1995). The plots show the mean percentage correct key word scores for 10 blocks of 10 narrowband sentences (1/3-octave, 96 dB per octave slopes) at nine different centre frequencies. Taken from Warren *et al.* (1995).**

The figure shows that as the centre frequency of the filter moves away from the frequency band containing the majority of the spectral information, in this case between 1100 and 2100 Hz, the intelligibility scores decrease. There is also evidence of a learning effect as the percent correct scores increase across the 10 blocks of sentences. The results for listeners who were presented with the combined signal of the two extreme bands are shown in Figure 6-3. It shows that when the two bands are presented together, either as their diotic sum or dichotically, the intelligibility is improved compared with the diotic presentation of each band individually. This suggests listeners are able to combine information from two disparate frequency bands to improve perception.

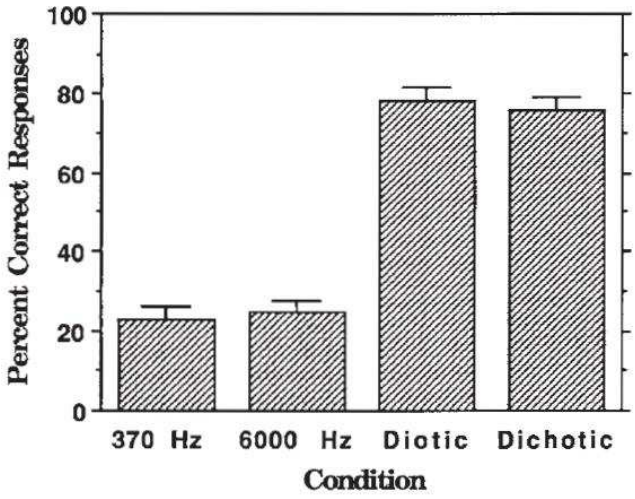



**Figure 6-3: The intelligibility results of the combined spectral bands (1/3-octave with centre frequencies of 370 and 6000 Hz), presented either individually or under diotic sum or dichotic listening conditions. Taken from Warren *et al.* (1995).**

The second experiment performed by Warren *et al*. (1995) used much narrower filters with a width of 1/20-octave and slopes of 115 dB per octave. Figure 6-4 shows that there is an obvious and expected degradation in the signal, as reflected in the intelligibility scores. However, the scores for the frequency band with a centre frequency of 1500 Hz are still relatively high.

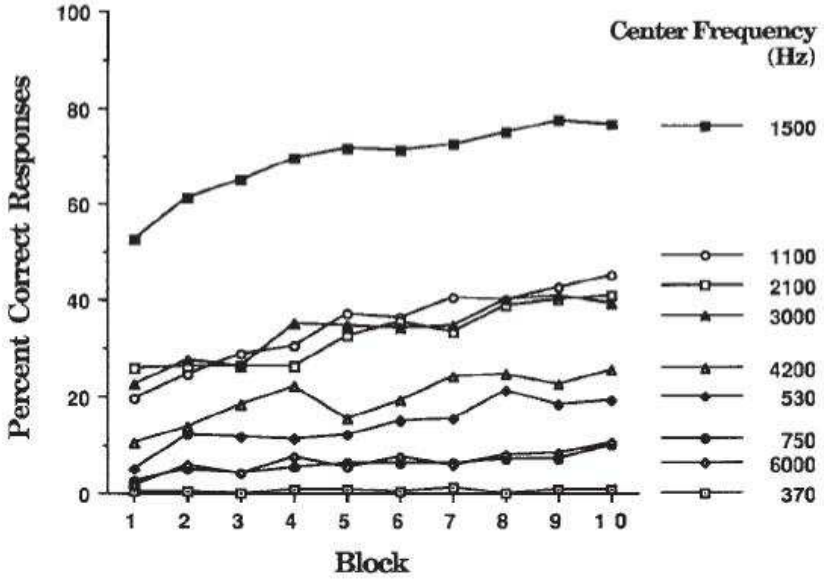



**Figure 6-4: The mean percentage correct word scores for 10 blocks of 10 sentences using filter widths of 1/20-octave and slopes of 115 dB/octave. Taken from Warren *et al.* (1995).**

The results from Warren *et al*. (1995) illustrate that we are very capable of understanding speech that has a large amount of spectral content removed.

Edmonds and Culling (2006) investigated the impact on intelligibility of three different listening scenarios. Firstly, the target and interferer speech signals were both presented to one ear (monaural). Secondly, the target was presented to the left ear and the interference to the right ear (dichotic). Finally, the target and interference were spectrally split between the ears (swapped). The swapped condition had the lower band of the target sound at the right ear and the upper band at the left ear. The interferer sound had the inverse spectral split, that is, the lower band at the left ear and the upper band at the right ear. Three split points were used, 750 Hz, 1500 Hz and 3000 Hz, each with an intentional spectral notch around the split frequency. Speech reception thresholds were determined for each of these configurations and for each of the frequency splits. The aim was to determine whether the separated bands of the target and interferer were fused together into single sound sources. This would produce two sound sources approximately in the middle of the listener's head, which should give similar intelligibility performance to the monaural condition. Alternatively, the split bands might be analysed in isolation, giving equivalent dichotic performance in each ear, which should give intelligibility performance similar to the true dichotic condition. The experiment would reveal whether listeners use a "better ear" rule (i.e. the ear with the best target-to-interferer ratio across all frequencies) or a "better band" rule (i.e. the frequency band for both ears containing the best target-to-interferer ratio). The results are given below in Figure 6-5.

The SRT shows a clear benefit of approximately 20 dB for dichotic listening compared with monaural listening. For the swapped condition there are two areas worth noting. Firstly, the SRTs are significantly higher than in the dichotic condition, which suggests listeners are using the better-ear rule rather than the better-bands rule. Secondly, the improved performance for the 750 Hz and 3000 Hz split frequency compared to the 1500 Hz split frequency confirms that the ear with the majority of the target signal, and therefore least interference signal, provides a better intelligibility advantage. They go on to suggest that the

reduced performance with a split frequency of 1500 Hz is due to the spectral notch, which removes some of the most useful information from the speech signals. It is also proposed that the listener focuses on the ear that produces the greatest advantage for intelligibility.

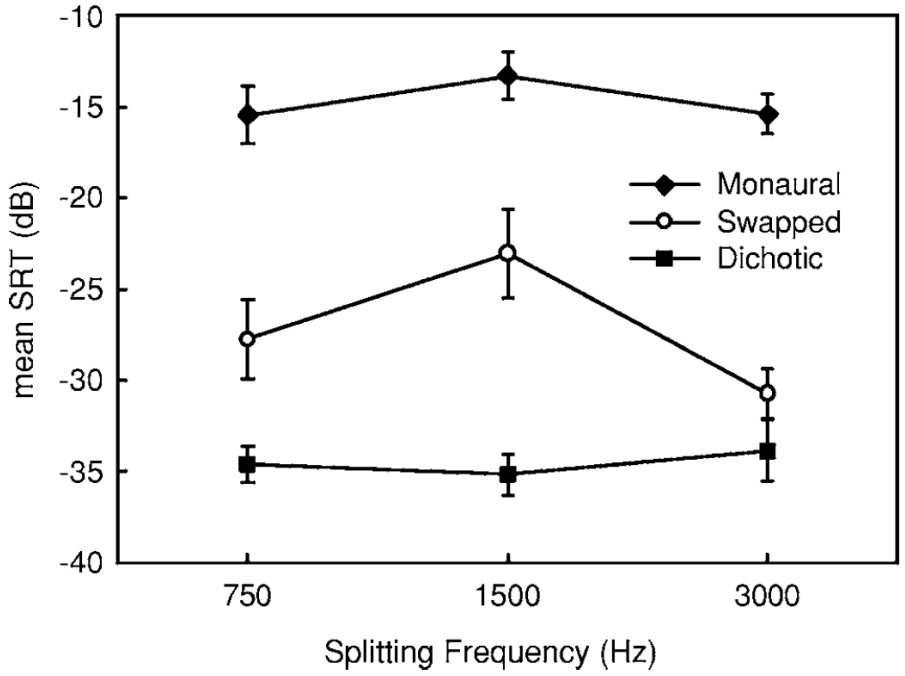



**Figure 6-5: The mean speech reception thresholds for different splitting frequencies, for monaural, dichotic and spectrally swapped speech signals. Taken from Edmonds and Culling (2006).**

Brungart *et al*. (2001b) carried out an experiment to investigate the influence of speaker gender on speech intelligibility, which was based on earlier work (Brungart, 2001a). A target speech signal was marked by a keyword within a phrase and was then masked by another speech phrase or noise. The masker was presented between 60 and 70 dB SPL, depending on the comfort level requested by the subject. The target was then adjusted to be between –12 dB and +15 dB of the masker in 3 dB steps. The target and masker could be the same talker, or another talker with the same gender or with opposite gender. Four male and four female talkers were used. Two types of noise masker were used, wideband noise with the same frequency range as the speech signal, and noise amplitude modulated by speech. The signals were presented simultaneously and diotically over headphones. The subjects were required to identify the key words in a

target phrase. The results show that the greatest percentage of correct scores is achieved when the target speech is presented simultaneously with modulated noise. The authors suggest that this may be due to the listener being able to catch a glimpse of the target signal during low amplitude sections of the noise signal. For the speech maskers, use of the same talker produced the worst intelligibility performance, with the talker of opposite gender leading to the highest performance. This is to be expected due to the increased spectral separation between the voices. They also found that one of the male speech sources was a particularly good target and poor masker, as it led to significantly better percentage correct scores than the other talkers. This implies that individual characteristics of the talker can also influence intelligibility. Apart from this particular case there were no significant differences between talkers, whether male or female.

Brungart *et al*. (2001b) also investigated the significance of the level difference between the target and interference signals. Within the experimental scenario already described, intelligibility continued to improve as the target was altered from being quieter than the masker signal, to being louder than it. This result is entirely predictable, as we would expect the masker to have less of an impact on the target signal if it has a lower amplitude. The results indicate an intelligibility improvement from 60% to 80% correct word scores for same gender talkers when the target was increased from 0 dB to 9 dB louder than the masker (Brungart, 2001a). It can therefore be concluded that suppression of the masker and/or increasing the level of the target signal will produce a significant improvement in intelligibility in an environment containing multiple sound sources.

In earlier related work, Edmonds and Culling (2005) investigated the influence of ITD within frequency bands for improving the intelligibility of a target speech sound source using simulated spatial separation. The target speech is presented to listeners simultaneously with either speech or noise masker signals. The signals were split into two frequency bands using split frequencies of 750 Hz, 1500 Hz or 3000 Hz. The lower and upper bands could have different ITDs of +/- 500 μs applied to simulate a spatial location to one side of the midline. The

target speech and speech masker were different male talkers. The noise masker was brown noise, generated using white noise and applying a 6dB/octave spectral roll off. The signals were presented to subjects in one of three conditions as shown in Figure 6-6.

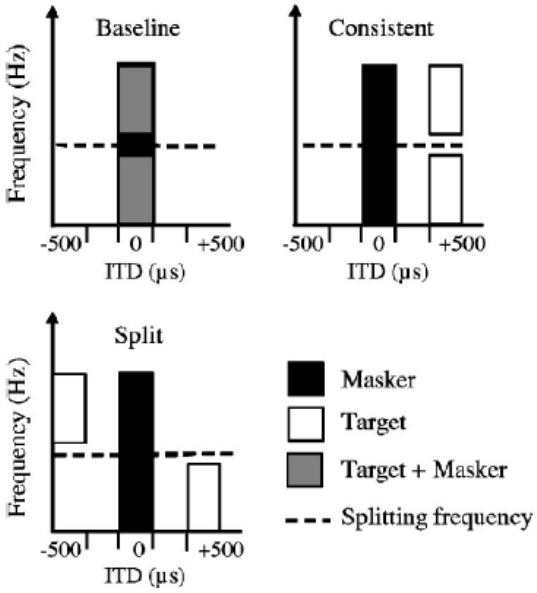



**Figure 6-6: The listening conditions for investigating the influence of ITD within spectral bands for improving intelligibility of a speech target signal. The dashed line represents the split frequency between the two bands. Taken from Edmonds and Culling (2005).**

Each condition has the masker presented with zero ITD, i.e. always at a simulated 0° azimuth. The target speech had either zero ITD, in the "baseline" condition, or the upper and lower band had an ITD of 500 μs for the "consistent" condition, or the upper and lower bands had different ITDs of +/- 500 μs for the "split" condition.

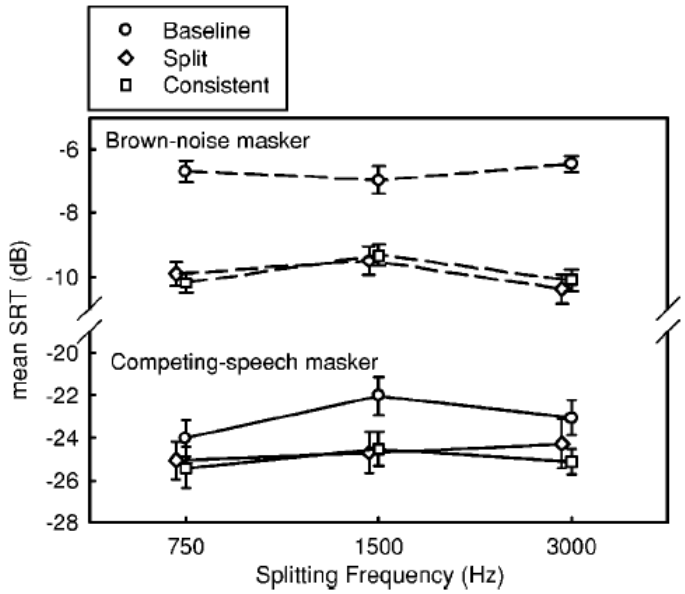



**Figure 6-7: The results for three forms of ITD processing of a target speech sound with a speech or noise masker. Taken from Edmonds and Culling (2005).**

The results in Figure 6-7 show that there is very little difference between the split and consistent configurations, but both perform significantly better than the baseline. Also, the target speech has a greater SRT when presented with a noise masker compared to a speech masker. This is to be expected according to the peek theory discussed in Section 4.2. However, it is surprising that the split and consistent listening conditions have such similar SRT values. Work by Healy and Bacon (2007) clearly indicates that the auditory system does not cope well with different time lags across frequency bands of the same sound source. However, their experiment used much larger time lags, of at least 12.5 ms, and the signals were presented diotically. The much shorter lag of 500 μs used by Edmonds and Culling is not only different across frequency bands but also different between the left and right ears. It would seem therefore, that the hearing system is more tolerant of small time differences across frequency. Furthermore, if the target and masker differ in ITD within each frequency band a full binaural advantage can be achieved.

### 6.2.1 Auditory glimpsing – the peek theory in practice

This section considers the ability of a listener to specifically use glimpses of dominant portions of a signal for improving intelligibility based on the peek theory. Bregman (1990) describes how in an acoustic environment with multiple sound sources, listeners use spectro-temporal segments of reduced interference activity to catch a glimpse of, or a "peek" at, the target sound source. This section reviews work that investigates the intelligibility improvement attributed to the peek theory.

Binns and Culling (2007) investigated the influence of the fundamental frequency (F0) contour on the intelligibility of a target speech source in the presence of a speech-shaped noise masker. Participants in a listening test were asked to identify key words in low predictability sentences. The F0 frequency contour was scaled to become 0.5, 0.25, 0 (monotone) or –1 (the inverse) of the original speech sentence. The processed conditions were compared to the unprocessed speech, i.e. with no processing of the fundamental frequency contour. The mixed signal of speech and noise was presented diotically over headphones. The noise level remained constant and the speech level was adjusted in +/-2 dB increments depending on the number of target words correctly identified. The speech reception thresholds were calculated from the average level of the target speech for the last seven from ten test runs in each condition. An additional set of experimental data was collected using another speech signal as the interference signal. The speech reception thresholds are compared in Figure 6-8 below.

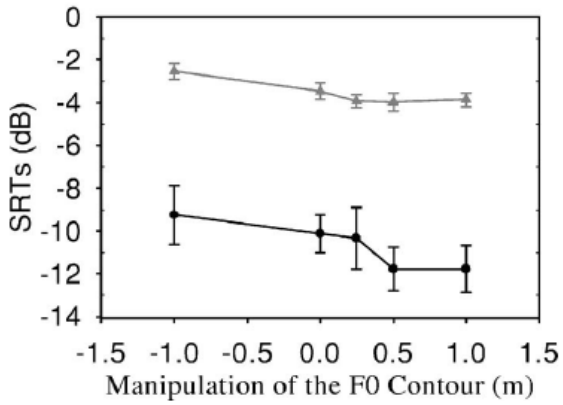



**Figure 6-8: Speech reception thresholds (dB) for a male speech target sound source with different F0 manipulations. Speech-shaped noise interference (grey line) and a different male speech interferer (black line) are shown. Plot taken from Binns and Culling (2007).**

The results show that across all F0 manipulations that were considered, there is an advantage of approximately 7dB in using a speech interferer compared to a noise interferer. It is suggested that this is due to the listener exploiting spectral and temporal gaps or lulls in the speech interferer. These beneficial lulls are not present in the constant-energy speech-shaped noise. This provides further evidence that, when listening in an environment containing multiple sound sources, a listener will make use of regions where the target sound source signal is dominant.

Li and Loizou (2007) directly investigated the influence of the peek theory on the intelligibility of a target speech sound source. Spectral holes, referred to as "glimpse windows", were created in the interference signals. The window size and position were adjusted according to the extent of the spectro-temporal region that was removed. The interferer was analysed and processed in 20 ms time frames. The durations of the windows used were: 20 ms, 200 ms, 400 ms, 800 ms and 'infinity', i.e. the total duration of the interferer. The total duration of the windows for each interferer was 800 ms, except for the 'infinity' configuration. The frequency bands used were low (LF), mid (MF), high (HF), low plus mid (LF+MF), full (FF) and random (RF). The time locations of the windows within the interferer signal were randomly selected. In each test the target was male speech and the interferer was 20-talker speech babble. These

were mixed and presented diotically using headphones. The results, showing the percentage of words correctly identified for each listening configuration, are displayed in Figure 6-9.

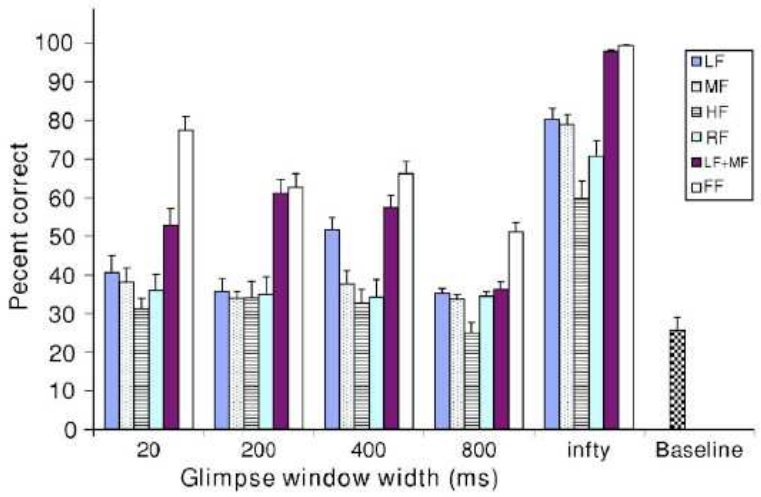



**Figure 6-9: The speech recognition performance, in terms of percentage correct words identified, for different glimpse window sizes. (LF: 0 – 1 kHz; MF: 1 – 3 kHz; HF: >3 kHz; RF: low, mid and high were randomly selected per frame; LF+MF: 0 – 3 kHz; FF: all frequencies). Taken from Li and Loizou (2007).**

The graph demonstrates that there is an advantage in introducing glimpse windows, compared to the baseline configuration of no glimpse windows. According to the peek theory the intelligibility performance should increase as the spectral range of the window is increased. This is confirmed by the results for the LF+MF and FF frequency bands. These show that intelligibility is improved compared with when only a single frequency band of the interferer is removed. The infinity configuration also provides an expected improvement as this removes spectral regions for the entire duration of the interferer. It also confirms that the majority of the information used for intelligibility is contained in the low- and mid-frequency bands. The slightly surprising result is the weak dependence of intelligibility performance on glimpse window duration, as wider windows would be expected to perform better than narrow windows. However, the total duration of the windows in the listening tests, as mentioned above, is limited to 800 ms, which is one third of the average target sentence duration of 2.4 s. Therefore, a single window of 800 ms leaves two other segments of

approximately 800 ms with no glimpsing benefit at all. On the other hand, the 40 shorter 20 ms windows are distributed throughout the entire duration of the target speech. Li and Loizou suggest that it is more beneficial to have multiple short windows than a few longer ones.

## 6.3 The influence of spatial separation

It is shown in Section 2.4.4 that there is a binaural advantage to be gained when listening with two ears compared to listening with one. The scale of this advantage is influenced by the spatial separation between sound sources. In general, the greater the angular separation between the sound sources, the more easily they can be distinguished. This section investigates the specific improvements in intelligibility which arise when spatially separated sound sources are presented to a listener.

Peissig and Kollmeier (1997) compared the speech reception thresholds of normal and impaired listeners in a multi-source environment. A male target speech source was presented with up to three interfering sources. The interference signals were either speech-shaped noise or continuous male or female speech. Each signal was spatialised using a set of HRTFs belonging to a listener who did not participate in the listening tests. Therefore, the signals were effectively nonindividualised. The spatial configurations used are shown in Figure 6-10.

The results for each of the configurations are shown in Figure 6-11. The varying interferer location in the first configuration shows that there is an advantage in increasing the spatial separation between target and interferer. Peissig and Kollmeier found that with the target fixed at 0°, the greatest advantage occurred with interferers at 105° and 255°. These were 9.8 dB for continuous speech-shaped noise and 6.2 dB for another speech signal. As can be seen from the spatial configurations in Figure 6-10(b) and (c), these directions were then chosen for the locations of stationary interference sources in subsequent tests.

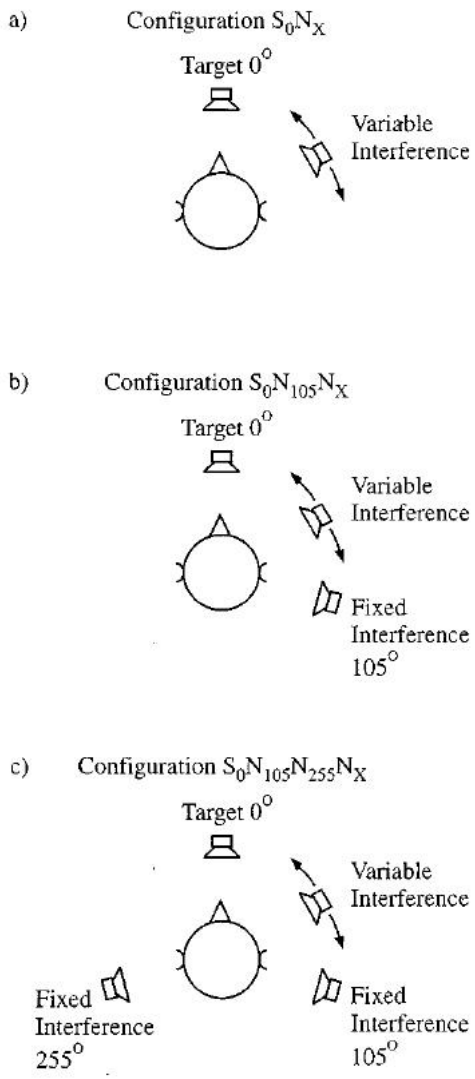



**Figure 6-10: The spatial configurations used by Peissig and Kollmeier (1997). Each configuration has the target sound source and one interferer whose location can vary, as shown in (a). (b) has an additional interferer at 105°. (c) has two additional interferers at 105° and 255°. Taken from Peissig and Kollmeier (1997)**

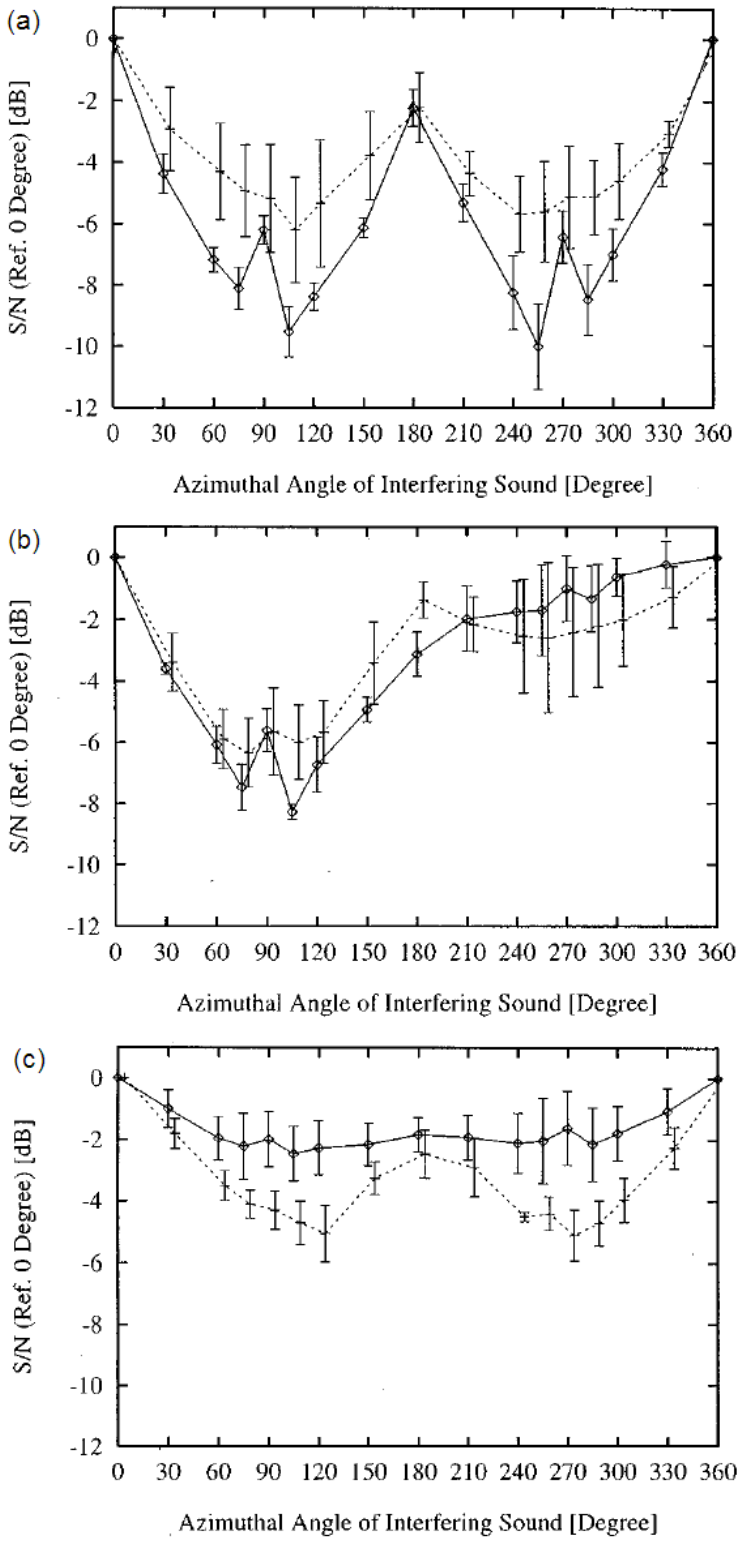



**Figure 6-11: The speech reception threshold results for the three interferer configurations shown in Figure 6-10. The solid lines with diamonds denote a continuous noise, the dashed lines with crosses denote the interferer talker configurations, (a) one varying interferer location, (b) one fixed and one varying interferer, (c) two fixed and one varying interferer. Adapted from Peissig and Kollmeier (1997).**

The results from the second experiment show that there is a maximum advantage at 105° of approximately 6 dB for the speech-shaped noise. There is a significant reduction in masking release when the mobile interferer is located in the opposite semicircle to the stationary interferer. They suggest that the binaural system is not able suppress two simultaneous interfering noise sources if they emanate from different sides of the head. Instead, binaural masking release is more easily achieved when the interfering sounds are close enough together to be heard as a single sound source. For the experimental condition using two speech sources as the interfering signals, the results are slightly different. There is a similar, distinct advantage when the mobile interferer is at the same location as the stationary interferer. However, there is also a small advantage of approximately 3 dB when the mobile interferer is on the opposite side of the head to the stationary interferer. The condition using two stationary interferer signals and one mobile interferer produces expected results: The masking release is reduced compared to a single stationary interferer, as there is now a constant interferer on the opposite side of the head. However, there is still an advantage to having the mobile interferer at the same location as one of the other stationary interferers. The advantage is only approximately 2 dB for the noise interferer, with the speech interferer reaching a maximum advantage of approximately 5 dB when the mobile interferer is near to one of the stationary interferers. In general, the masking release is greatly reduced when all interference signals are spatialised to separate locations. They suggest that the speech interferer performs better than noise in a multi-source and multi-direction scenario because the binaural system is provided with gaps in the interferer from one location during which the binaural advantage can be exploited for the interferer at a different location.

The results for hearing-impaired listeners are somewhat more varied than those for normal hearing listeners. With a single interferer there appears to be a loose relationship between interferer location and the hearing loss for the ear on that side of the head, inasmuch as the hearing loss reduces the impact of the interfering noise. In general, the hearing-impaired subjects showed an almost non-existent binaural advantage when presented with multiple noise or speech interference signals.

Freyman *et al*. (1999) investigated the role of perceived spatial separation between a target and interfering sound sources. The target signal consisted of meaningless sentences containing three keywords, for example "the tree ate the book". This was spoken by a female talker and for the baseline condition was presented from a loudspeaker directly in front of the listener, i.e. 0º azimuth and elevation. The subjects were asked to repeat back the target sentence they heard. The first configuration used continuous noise shaped with a speech spectral envelope and presented from the same loudspeaker as the target sound. As expected, they found that speech recognition increased as the signal-to-noise ratio increased. A further experimental configuration involved the speech spectrum noise being presented from a loudspeaker at 60º to the right of the listener as well as from the front loudspeaker delayed by 4 ms. This simulation gave the impression that the interfering noise was on the right of the listener and the speech target remained directly in front. They found that speech recognition did not improve relative to the baseline configuration. Freyman *et al*. suggest that the perceived spatial separation of speech and noise is not important when listening to speech in stationary background noise. The results were different when the interferer signal was a different female talker. They found a distinct improvement when the interferer was presented such that it was perceived on the right of the listener, compared to being presented from the same loudspeaker as the target, in front of the listener. They suggest that the reason for the advantage due to spatial separation using a speech masker is due to informational masking, whereas speech spectrum noise only provides energetic masking. This is consistent with findings by Kidd *et al* (1998), as discussed in Section 4.3.

Based on these findings, Freyman *et al*. (2001) investigated the spatial release from informational masking in speech recognition, that is, the advantage in speech recognition due to increasing the spatial separation between sounds. Their intention was to determine whether it is the temporal and spectral fluctuations in the speech signal that produce advantages. Firstly they extended the scope of the earlier experiment by using, in one case, a different female interferer talker and, in another, two different female interfering talkers. The spatial configurations were the same as for their previous experiments (Freyman

*et al.*, 1999). For the two-interferer-talkers configuration, both had the same spatial location. The target was always presented from the speaker in front of the listener at a level of 46 dBA. Four target-to-interference ratios were used: -12, -8, -4 and 0 dB. For example, for a ratio of -4 dB the interferer would have a level of 50 dBA. The mean percentages of words recognised correctly are shown in Figure 6-12. 'F-F' is the configuration with both the target and interferer in front of the listener. 'F-RF' indicates that the target sound was in front and the two interferers were in front and to the right, with the right leading by 4 ms. This is shown in Figure 6-13.

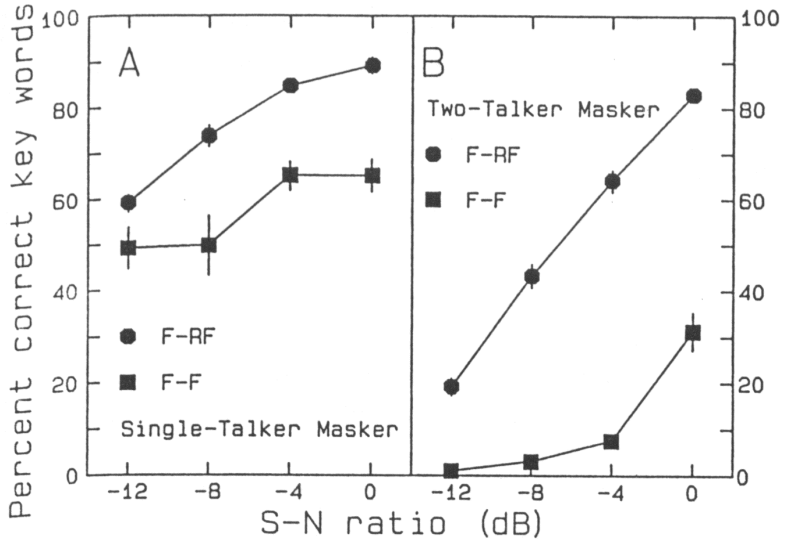



**Figure 6-12: Percentage correct keyword results from experiment by Freyman *et al.* (2001). (A) using one and (B) using two free-field spatially separated female speech interferer signals with different SNRs. Taken from Freyman *et al.* (2001).**

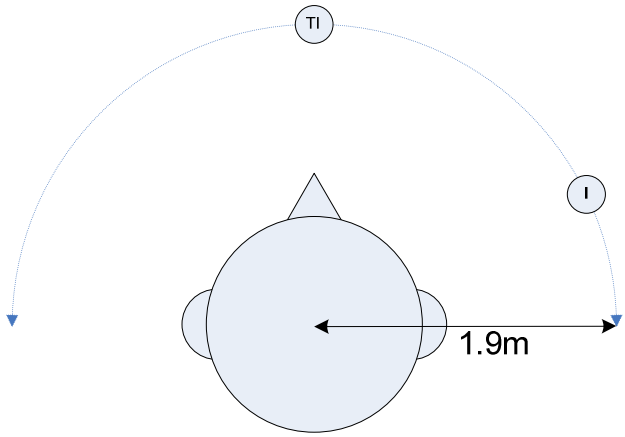



**Figure 6-13: The listening configuration for the listening experiment by Freyman *et al.* (2001). The target 'T' is presented from a loudspeaker directly in front of the listener. The interferers 'I' are presented from the front loudspeaker and another loudspeaker at 60° to the right.**

Spatial separation of the target and masker provided a clear improvement in recognition scores, which improved further as the signal-to-noise ratio was increased. For two interfering talkers the advantage is even more pronounced. When all the sound sources were presented from 0º the listeners found it very difficult to identify the target speech. Only 30 percent of target key words were correctly identified when all signals had the same level. However, with a spatial separation between the target and both maskers, over 80 percent of the target key words were correctly identified.

A follow up experiment was carried out to investigate the influence of presenting the signals binaurally using generic HRTFs. This was based on a reduced set of the configurations used in the main multiple-talker interferer experiment described above. Only a –4 dB target-to-interferer ratio was used for the case of two masker talkers. A subset of the sentences was recorded using a KEMAR. The manikin was positioned where the listener's head was normally located during the tests. The recorded signals were presented to subjects over headphones in one of four configurations: monaural left; monaural right; binaural with spatial separation; and binaural without separation. The results are shown in Figure 6-14. For the monaural conditions only the signal for either the left or right headphone was presented to the listener. The monaural configurations

show very poor correct word scores even though the target and interferers were spatially separated. The binaural results show an advantage due to a spatial separation, but not of the same scale as the free field experiment, shown in panel B of Figure 6-12. Freyman *et al*. acknowledge that the KEMAR recordings were not accurately externalised for the listeners and may have produced an inferior spatial distinction between the target and interferers compared to free-field listening. However, the advantage due to spatial separation of target and interferers using generic binaural cues is still significant and is shown in Figure 6-14. For the test signals used, there is an improvement from 5% for the F-F configuration to approximately 36% correct key words for the F-RF configuration. This is with an angular separation of 60° between the target and interferer sound sources. The equivalent free-field results, shown in Figure 6-12, have correct word scores of 8% for the F-F configuration and 63% for the F-RF configuration.

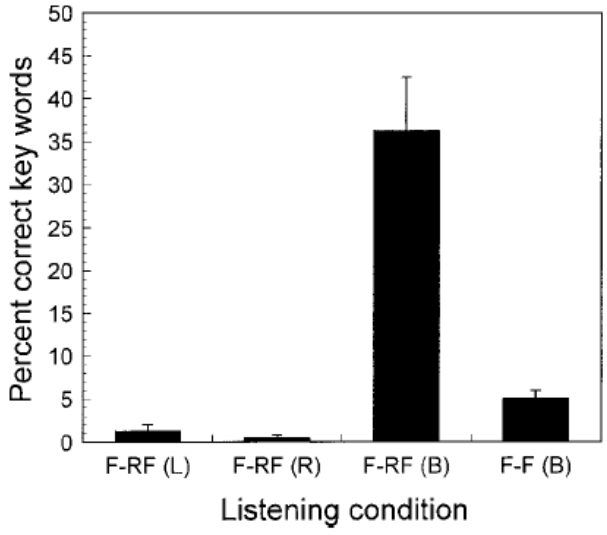



**Figure 6-14: Results from an experiment by Freyman *et al.* (2001), showing the significance of spatially separating the target and interferer, F-RF, compared to collocating them, F-F. The masker and target are either presented monaurally, to the left or right ear, or binaurally using recordings from a KEMAR manikin. Taken from Freyman *et al.* (2001).**

The next stage in their investigation involved the use of noise maskers instead of speech. This is effectively based on the peek theory discussed in Section 4.2. Two speech-envelope modulated (SEM) noise signals were created as follows.

The first was a single channel noise, generated by modulating white noise with the temporal amplitude envelopes of the summed two-talker interference from the previous experiment. The second was a multi-channel noise generated by splitting the two-talker interferer into eight adjacent frequency bands between 0 and 8 kHz. The temporal envelope of each band was then used to modulate eight similarly band-limited noise signals and these were then summed. The interferer signal for this experiment was either the single- or multi-channel SEM noise, or the sum of the single- or multi-channel SEM noise and its corresponding two-talker interferer signal. Their general finding was that there is almost no significant advantage in spatial separation between the target and the SEM noise masker, whether single- or multi-channel. Therefore, Freyman *et al*. (2001) conclude that the amplitude modulations of a masker are not directly responsible for creating the binaural advantage. The advantage returns, however, when the interferer signal is the sum of the SEM noise and the two-talker interferer signals it is based on.

Finally, the scope of the experiment was extended by employing yet more types of interferer signals. Either reversed speech or spoken Dutch was used as the interferer signal, to determine whether the intelligibility of the masker affected the intelligibility of the target. None of the participants in the experiment could speak or understand Dutch. Their results show that an advantage is still attainable even if the listener cannot understand the speech masker signals. They conclude that other differences, such as pitch variation or fine detail, are perhaps used in identifying a target speech signal against interfering speech maskers. It is possible that the eight frequency bands did not provide sufficient frequency resolution for the peek theory to apply. Therefore, it may be concluded that speech-envelope modulated noise is a more effective masker than normal speech.

Johnstone and Litovsky (2006) showed that some maskers provide only energetic masking, such as reversed speech, whereas others introduce informational masking as well, such as forward speech.

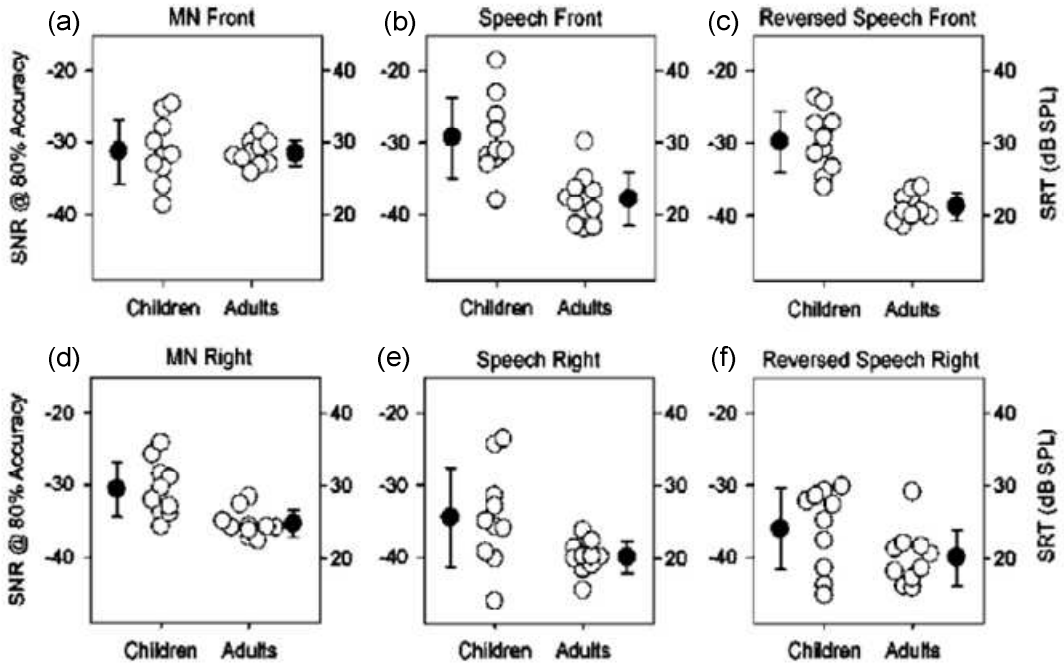



**Figure 6-15: The SNR and speech reception thresholds (SRT) for children and adults for different interferer signals when, (a) to (c) the target and interferer are collocated and (d) to (f) the target is in front and the interferer to the right of the listener. In each case, three types of interferer signal are used, modulated noise (MN), forward speech and reversed speech. Taken from Johnstone and Litovsky (2006).**

Their results are presented in Figure 6-15 and a number of key points are worthy of note. In general, adults tend to perform better than children and exhibit less variance in their SRT values. For the test procedure and sound files used, there is only a small amount of spatial release indicated by the difference between the top row of plots (collocated target and interferer) and the bottom row of plots (target in front, interferer on right). There is, however, a distinct difference between using modulated noise (MN) and speech as the interferer for adult listeners. The spatial release from masking for two different target talkers used in the experiments by Johnston and Litovsky is shown in Figure 6-16.

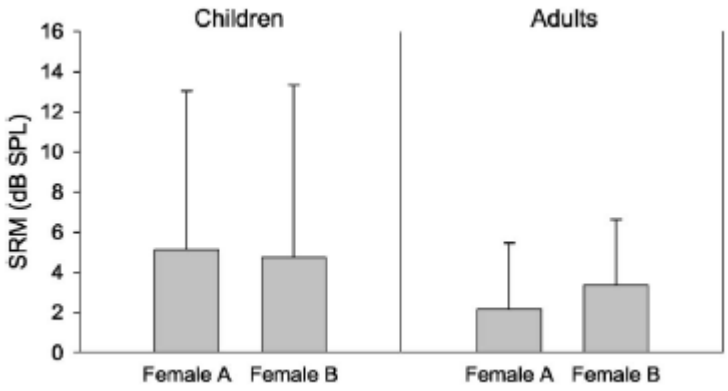



**Figure 6-16: The spatial release from masking (SRM) for children and adults tested by Johnstone and Litovsky for 2 female target talkers and 3 different masker types. Taken from Johnstone and Litovsky (2006)**

The spatial release from masking (SRM) is calculated as the difference between SRT for the different masker types in front and to the right of the listener. This is significantly lower than expected based on the large improvement in correct-word scores in other research covered in this section. This suggests that there is a narrow range of SNR across which intelligibility can change dramatically. Hence, under appropriate conditions, the intelligibility of a target with a collocated interferer can be significantly improved by increasing the SNR by only a few dBs, or by introducing an angular separation between the target and interferer sound sources.

The influence of different interferer signal types and the spatial configurations relative to a target speech sound source can have a very significant effect on its intelligibility. This is the focus of the next section.

### 6.3.1 The number, type and location of interferers

Hawley *et al*. (1999) carried out a number of listening tests to investigate the influence of localisation on speech intelligibility in a multi-source environment. The target and masker signals were speech sentences recorded by two male talkers and these were all scaled to the same average level of approximately 62 dBA. Within each trial the same talker spoke the target and competing

sentences. The sentences were either spatialised using KEMAR HRTFs and presented over headphones, or presented from frontal loudspeakers arranged in a semicircle five feet from the subject. Seven possible directions were used: -90º, -60º, -30º, 0º, +30º, +60º and +90º, all at 0º elevation.

The spatial configurations were split into three groups; close, intermediate and far, based on the angular proximity of the target to the competing sound sources. When the target is to one side of the listener and the nearest interferer was intermediate or far, the percentage of key words in error was between 0 and 20%, for one, two or three competing sound sources. However, when the competing sounds were located close to the target, the percentage keywords in error increased to 35% for one, 65% for two and 80% for three competing sound sources.

In a second experiment, Hawley *et al*. (1999) tested subjects' localisation ability in both the real and virtual sound-field and found that performance was significantly better in the real sound-field. However, the choice of sound field had no effect on their performance in a speech intelligibility experiment. The main conclusion that can be drawn from their results is that the proximity of the competing talker to the target talker impacts intelligibility more than the number of simultaneous competing sentences.

This result is similar to that found in work by Drullman and Bronkhorst (2000), who have investigated the influence of multiple talkers and spatial separation on speech intelligibility. Figure 6-17 shows that intelligibility scores measured by them are significantly worse when multiple competing sound sources are present. One female and four male talkers were used, with one of the male talkers being selected as the target. The signals were either sentences or meaningful syllables. All the signals were lowpass-filtered at 4 kHz and presented simultaneously for each test. For the monaural condition, the target and interferer sounds were all presented to the right ear. For the "binaural" condition, the target is presented to the right ear and each competing sound is presented to either the left or right ear. The 3D spatialised signals only varied in azimuth and were processed using either KEMAR HRTFs or individualised HRTFs. The azimuths used were –90º,

-45º, 0º, +45º and +90º, all at 0º elevation. The results for the average percentage correct scores in Figure 6-17 show that there is very little difference between intelligibility scores when using individualised or generic KEMAR HRTFs. The authors suggest that this is due to the lack of high frequency cues in the speech signals and that the sets of HRTFs were very similar below 4 kHz. On the other hand, the results using spatialised signals clearly show an improvement compared with monaural or dichotic signals, especially as the number of competing signals is increased.

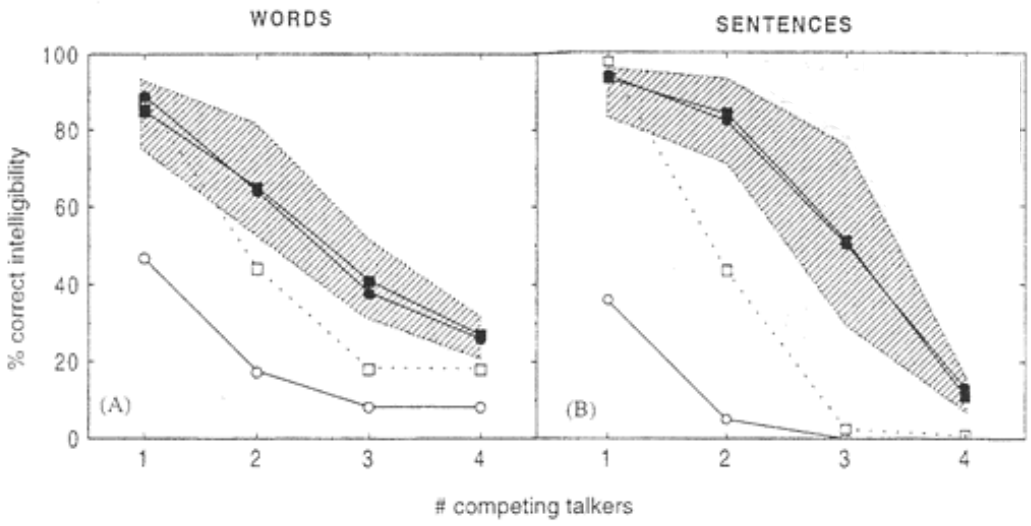



**Figure 6-17: Average percentage correct scores for intelligibility experiment by Drullman and Bronkhorst (2000). Panel A shows scores for words and panel B shows scores for sentences, as a function of the number of competing talkers. The symbols represent the results for, ○ – monaural, □ – binaural, ● – individualised 3D HRTFs and ■ - generic 3D HRTFs. The hatched area indicates the range of scores for the 3D configurations. Taken from Drullman and Bronkhorst (2000).**

Drullman and Bronkhorst (2000) also investigated the influence of spatial location for the conditions using HRTFs. They considered the angular separation between the target and the nearest interfering signal, and termed this the target-competing talker angle (TCA). The results for 45° and 90° TCAs are shown in Figure 6-18.

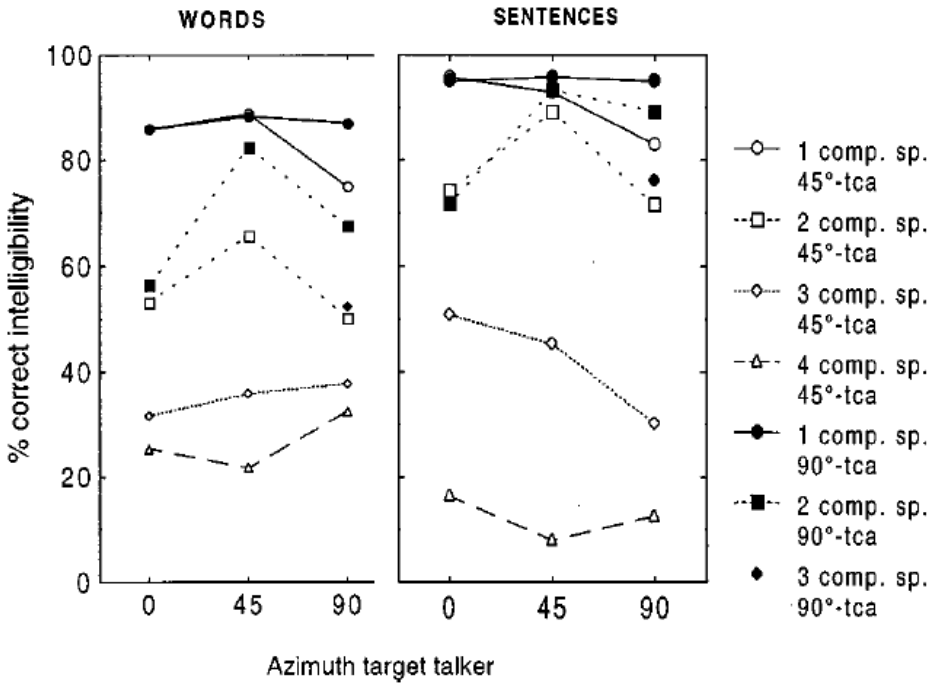



**Figure 6-18: The influence of angular separation between a target and competing sound source (TCA) in experiments by Drullman and Bronkhorst (2000). The plots show the mean intelligibility scores for words and sentences as a function of the azimuth of the target talker. Taken from Drullman and Bronkhorst (2000).**

It may be seen that the best location for the target signal, in terms of intelligibility, was for an azimuth of 45º, and not on the extreme right or left (+/- 90º). There is also a tendency for intelligibility to rise when the interference has a greater angular separation from the target. This is indicated by the filled symbols in Figure 6-18 being above the empty symbols.

Best *et al*. (2006) considered the influence of spatial separation of speech sound sources for a divided listening task in which subjects must pay attention to the content of more than one speech sound source. A listener's ability to extract information simultaneously from multiple sound sources is made more difficult when they are spatially separate. It has been discussed in Section 4.3 that increasing the angular separation of interferers from a target provides a binaural advantage. However, it follows that the converse is true, i.e. when there are multiple targets, they should be spatially collocated. Best *et al*. (2006) relate this to the "spotlight model" in visual perception and investigate its analogue in auditory perception through listening tests. Subjects were asked to attend to

either a single target speech sound source within the mixture, or both of the speech sound sources simultaneously. The sound sources were presented using loudspeakers positioned 1 m from the listener at the azimuths shown in Figure 6-19. The same talker was used for each sound source. A pre-test calibration exercise was performed on each listener to determine a noise level that resulted in 85% correct word scores when attending to a single sound source. During this process, two sentences were presented, one at each loudspeaker, with white noise from both loudspeakers. The noise levels were adjusted for each listener and each spatial configuration.

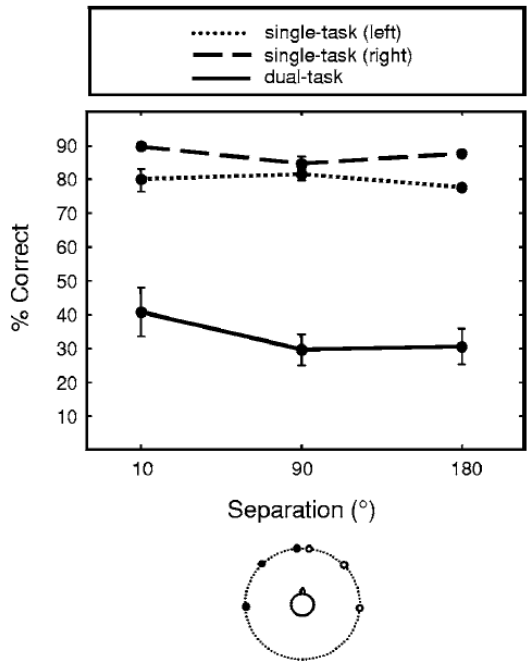



**Figure 6-19: The impact of spatial separation for single and dual attention configurations. Two sources were presented with angular separations of 10°, 90° or 180°. Subjects correct word scores are shown when attending to left (dotted), right (dashed) or both (solid) sound sources. Taken from Best *et al.* (2006).**

The results show a significant intelligibility impact when subjects were asked to attend to two sound sources instead of one. The effect worsens with large angular separations, confirming the theory of an "auditory spotlight". Typically, it would be expected that percentage correct word scores would increase with angular separation when listeners focus on a single target sound source. However, the use of background white noise to force a baseline performance condition has removed the binaural advantage for that listening configuration.

## 6.4 Summary

Two areas related to improving the intelligibility of a target speech sound source have been discussed; the spectral content of the sound sources and their spatial separation. Both of these attributes of the auditory scene have a significant influence on the ability of a listener to understand a target speech sound in the presence of multiple interfering sounds.

This section has highlighted the frequency bands that are most important for the intelligibility of speech. The frequency region from 1100 Hz to 2100 Hz contains most of the information in speech. If this band is missing, intelligibility is greatly reduced. If other frequency bands are available to a listener, the hearing system is able to combine them to provide an intelligibility improvement. This suggests that any processing of the spectral content of a target sound in frequency bands where it contains useful energy should be performed consistently.

The hearing system can tolerate an inconsistent ITD of 500 μs across the frequency bands at each ear. Intelligibility performance is similar to that achieved when the same ITD is applied across all frequency bands. It has been shown that the lower bands, where ITD is dominant, are used to achieve a binaural advantage.

There is evidence that listeners make use of a "better ear" listening strategy when multiple sound sources are presented. That is, the ear receiving the dominant target signal components is used to provide an intelligibility advantage. This suggests that the target sound should be processed to take advantage of this strategy.

According to the peek theory, listeners are able to make use of spectral lulls in the interferer sounds to give them a glimpse of the target sound. Interferers such as speech allow this to happen, whereas white or spectrally shaped noise does

not. Short temporal intervals of frequent exposure to a target sound can provide adequate information to achieve intelligibility.

It is clear there is an advantage in having the target and interferer sound sources spatially separated and it is also beneficial to have the interferers grouped together, away from the target sound source. The advantage can be up to 6dB when a speech-shaped noise interferer is relocated from the same position as the target speech to an angular separation of 105°. However, the advantage reduces to 2dB if there is a similar interferer at the same angular separation on the other side of the head. There is a rapid improvement in intelligibility of a target as interferers are moved beyond an angular separation of 30°. Furthermore, the hearing system can more easily process information from two target sound sources if they have an angular separation of up to 10°. Intelligibility performance reduces as the angular separation increases indicating a narrow focal region or auditory spotlight.

The intelligibility advantage is not restricted to free-field signals or the use of personalised HRTFs. Generic HRTFs used to simulate spatial separation have also been shown to provide significant intelligibility improvements. This is due to the basic similarity between all HRTFs below 4 kHz, where the majority of speech energy is contained.

# Chapter 7 Binaural processing of multiple sound sources

## 7.1 Introduction

The intelligibility of speech can fall dramatically in natural situations where it is contaminated by interfering sounds. The problem is exacerbated for people suffering from a hearing deficit. Conventional hearing aids tend to concentrate on improving residual hearing in a monaural sense by applying equalisation, compression and noise rejection (Oticon, 2008a). This strategy does not exploit our innate ability to understand speech better using both ears in synchrony. In fact, any signal processing that is applied in isolation and monaurally is likely to disrupt the auditory scene for the listener even further. A signal processing strategy that is able to make appropriate use of information at both ears simultaneously will certainly have an advantage over standard monaural systems.

The literature review in Chapters 4 to 6 has highlighted some interesting results specific to enhancing speech intelligibility. It is clear that there are particular auditory cues that can be exploited to improve the intelligibility of a target speech signal in the presence of multiple interfering speech sources. The key results that have been highlighted in the literature review are summarised in Table 7-1.

| *First author* | *Factor* | *Presentation* | *Target* | *Interferer* | *Significant results* |
|---|---|---|---|---|---|
| Shinn-Cunningham (2001) | masking - distance and angular separation | generic HRTFs | speech | speech | Increase the angular separation between sounds to increase the binaural masking level difference (BMLD).<br>Increase the (radial) distance of the masker from the listener compared to the distance of the target, to increase the BMLD.<br>If the target and the masker are both at 0° and 1 m from the listener, either move the masker to 45° or move the target to 90° to obtain a 6 dB advantage. |
| Durlach (1964) | masking - frequency | - | - | - | BMLDs are greater for low frequencies than for high frequencies. |
| Kohlrausch (1990) | masking - duration | dichotic | tone | noise | Shorter signal durations produce smaller BMLDs. Approximately 300-500 ms masker duration is required to detect an interspersed target signal. |
| Kidd (1998) | masking - masker type | freefield | tone | noise | Spatial separation of a target from an energetic masker can achieve advantages between 5 and 10 dB. For an informational masker it can be over 20 dB. |
| DeSimio (1996) | masking - angular separation | generic HRTFs | speech | noise | Angular separation improves phoneme recognition in the presence of a white noise masker. |
| Noble (1997) | masking - angular separation for hearing impaired | freefield | speech | noise | Listeners with high frequency hearing loss are unable to use pinna cues and so can not accurately localise sounds and therefore have a limited binaural advantage when sounds are spatially separated. |
| Culling (2000) | masking - articulation rate | dichotic | speech | noise | Rapid speech is harder to detect and understand than slow speech when presented with a noise masker.<br>The binaural advantage of having spatially separated speech and noise is about the same, irrespective of the articulation rate. |
| Best (2007) | masking - grouping | dichotic | tone | tone | Lateral position of target tone is identifiable when interfering tone is extracted into a separate auditory stream. |

| *First author* | *Factor* | *Presentation* | *Target* | *Interferer* | *Significant results* |
|---|---|---|---|---|---|
| Hall (1984) | masking - flanking masker | - | tone | noise | When a tone is masked by a narrow band of noise, the amount of masking can be decreased by simultaneously presenting an additional flanking masker that has a different centre frequency to the main masker and target signal |
| van de Par (1998) | masking - flanking masker with amplitude modulation | dichotic | tone | noise | Release from masking occurs if low frequency amplitude modulation is applied to the noise bands |
| van de Par (1998) | masking - flanking masker | dichotic | tone | noise | Release from masking occurs if the target signal and on-frequency noise band are presented to one ear and an additional flanking noise band is presented to the other ear. |
| Gallun (2007) | masking - contralateral masker | dichotic | speech | noise | Listeners are unable to isolate the SNR in one ear from a contralateral interferer. |
| Noble (1997) | masking - angular separation | freefield | speech | noise | Lower intelligibility scores for hearing impaired listeners compared to normal hearing listeners may be due to a loss of high frequency pinna cues |
| Brungart (2007) | masking - talker gender | dichotic | speech | speech | There is a complicated relationship between the comparison of a within-ear masker relative to the target, and an across-ear masker relative to the target. |
| Iyer (2007) | masking - interruption rate | diotic | speech | speech/noise | The interruption rate of masking sound affects intelligibility of the target sound. Speech interferers have less impact when continuous, whereas noise interferers have less impact when interrupted. |
| | | | | | |
| McAdams (1998) | continuity - signal level | diotic | tone | tone | If a signal induces continuity in another signal a residue signal will be heard.<br>The type and level of the inducer signal influences the residue signal. |
| Drake (1999) | continuity - signal duration | diotic | tone | tone | In general an inducer signal needs to have the same or shorter duration as the inducee signal before and after it.<br>For inducer durations up to 300 ms the inducee can be shorter. |
| Kashino (1996) | continuity – interaural phase differences | dichotic | tone | noise | Continuity thresholds are lower if the inducer and inducee are at different spatial locations. |

| *First author* | *Factor* | *Presentation* | *Target* | *Interferer* | *Significant results* |
|---|---|---|---|---|---|
| Darwin (2002) | continuity – binaural | dichotic | Huggins pitch | complex noise | A diotic sound more easily induces continuity in a binaural sound than a monaural sound. |
| Bashford (1996) | continuity – phonemic restoration | diotic | speech | noise | There are two levels of continuity for a speech target signal<br>(i) the induction of the apparent continuity of the signal<br>(ii) the grouping of broken segments of speech into sentences<br>It is therefore easier to understand sentences interrupted with noise than lists of random words. |
| Bashford (1996) | continuity - interferer type | diotic | speech | different noises | An interrupting signal of speech-modulated noise improves intelligibility more than white noise or silent gaps. |
| Bashford (1987) | continuity – interruption duration | diotic | speech | noise | For a target of normal discourse speech, a noise interruption of broadband noise can be detected if its duration exceeds about 300 ms.<br>For narrowband noise with a centre frequency the same as the speech, the duration for detection is the same. |
| Bashford (1987) | continuity - target context | diotic | speech | noise | Normal discourse can suffer a longer noise interruption duration than backward discourse or random word lists. |
| Cherry (1967) | continuity - interferer level | diotic | speech | noise | White noise interferer signal levels to restore intelligibility of band limited speech covered a 40dB range across subjects. |
| Bashford (1992) | continuity - interferer frequency band (phonemic restoration) | diotic | speech | noise | Band limited speech is restored better with an interrupting noise that has a matching frequency range, than a different frequency band. I.e. the spectral requirement for inducing continuity is applicable to the restoration of intelligibility. |
| Bashford (1992) | continuity - context | diotic | speech | noise | Interrupted high predictive sentences are restored more easily than random word lists. |
| Elfner (1967) | continuity - frequency separation | dichotic | tone | tone | Continuity duration thresholds increase as the frequency separation between the inducer and inducee is reduced. |
| Elfner (1971) | continuity - angular separation | free-field | tone | tone | Continuity duration thresholds increase as the angular separation between the inducer and inducee is reduced. |

| *First author* | *Factor* | *Presentation* | *Target* | *Interferer* | *Significant results* |
|---|---|---|---|---|---|
| Kelly (2002) | continuity - spectral redundancy | generic HRTFs | tone | noise | Continuity is achieved with a dominance-only selection criteria for target and interferer sounds. |
| Smith (2006) | continuity - phoneme restoration | diotic | speech | noise | Spectral warping of energy into sidebands around spectral holes increases intelligibility. |
| Warren (1976) | continuity - contralateral induction | dichotic | tone | noise | Continuity induced at each ear produces a fused image in the middle of the head. |
| Bashford (1988) | continuity - spectral gaps | diotic | speech | noise | Wideband noise inserted into a spectral gap in speech improves intelligibility compared to leaving the gap filled with silence. |
| Samuel (1981) | continuity - phoneme restoration | diotic | speech | noise | Noise is better for restoring fricatives tones are better for restoring vowels. |
| Warren (1997) | continuity - phoneme restoration | diotic | speech | noise or silence | Intelligibility increases when noise is inserted into gaps in the target speech. The effect increases as the sentences become more predictable |
| Shinn-Cunningham (2008) | continuity - phoneme restoration | diotic/dichotic | speech | noise or silence | Speech-modulated noise provides the best intelligibility restoration. Its effectiveness reduces if it is spatially separated from the target. |
| | | | | | |
| Freyman (1999) | Intelligibility - Spatial separation/Interferer type | free field | speech | noise or speech | Spatial separation of a speech target and noise masker does not improve intelligibility. If the interferer is another speech signal, there is an advantage. |
| Freyman (2001) | Intelligibility - Spatial separation | free field | speech | speech | Spatial separation of two interfering speech sources produces a greater advantage relative to only one interfering speech source. |
| Freyman (2001) | Intelligibility - Spatial separation | KEMAR HRTFs | speech | speech | The advantage of spatially separate interferers when presented binaurally, processed using generic HRTFs, is still significant, but not of the same level as free field listening. |
| Freyman (2001) | Intelligibility - Spatial separation/noise type | free field | speech | noise | There is almost no advantage to spatially separating temporally or spectrally modulated noise, from a speech target. |

| *First author* | *Factor* | *Presentation* | *Target* | *Interferer* | *Significant results* |
|---|---|---|---|---|---|
| Hawley (1999) | Intelligibility - location and number of interferers | KEMAR HRTFs / freefield | speech | speech | The results from soundfield and virtual listening conditions are similar.<br>The proximity of the competing talker to the target talker impacts intelligibility more than the number of simultaneous competing sentences. |
| Drullman (2000) | Intelligibility - angular separation | KEMAR / personalised HRTFs | speech | speech | The greater the angular separation between the target and interferer, the better the intelligibility level. The favoured target direction is 45°. |
| Drullman (2000) | Intelligibility - number of talkers | KEMAR / personalised HRTFs | speech | speech | The more interfering talkers the worse is the intelligibility level. This can be improved by increasing the angular separation between the target and the interferers. |
| Best (2006) | Intelligibility - spatial separation | freefield | Speech | Noise | Multiple targets should be collocated.<br>It is difficult to pay attention to 2 speech sources simultaneously.<br>There is a slight advantage when the target is on the right compared to the left. |
| Johnstone (2006) | Intelligibility - spatial separation/interferer type | free field | speech | noise or speech | There is a spatial release from masking which depends on the type of interferer used. A 90° separation helps. |
| Peissig (1997) | Intelligibility - spatial separation/location for normal and impaired hearing | generic HRTFs | speech | noise or speech | Normal hearing listeners found that having the interfering sounds at the same spatial location is better than each having a different spatial location.<br>The best azimuth angles for two interferers at different locations were found to be 105º and 255º, with the target at 0º<br>Hearing impaired listeners showed a slight advantage to having interfering sounds on the same side of the head as their hearing loss.<br>Hearing impaired listeners showed almost no binaural advantage when presented with multiple interfering sounds. |
| Brungart (2001b) | Intelligibility - speaker gender | diotic | speech | speech | Speaker gender does not produce significant differences when used as a target or a masker. However, some speakers are better targets than others. |
| Brungart (2001b) | Intelligibility - speaker amplitude | diotic | speech | speech | Intelligibility improves if the target is louder than the masker. |

| *First author* | *Factor* | *Presentation* | *Target* | *Interferer* | *Significant results* |
|---|---|---|---|---|---|
| Brungart (2001b) | Intelligibility - masker type | diotic | speech | speech | Intelligibility is better with an amplitude-modulated noise masker than wideband noise, due to the 'peek' theory. |
| Brungart (2001a) | Intelligibility - target and interferer levels | diotic | speech | speech | Increasing the level difference between the target and the masker assists intelligibility. |
| Warren (1995) | Intelligibility - target speech spectral content, spectral redundancy | diotic or dichotic | speech | none | The centre frequency of band-limited speech should be close to the range containing the majority of the spectral information, i.e. 1100 Hz to 2100 Hz. |
| Edmonds (2006) | Intelligibility - spectral content | monaural/diotic<br>dichotic<br>Spectral split | speech | speech | SRTs are significantly worse for the swapped condition than the dichotic condition, which suggests listeners are using the better-ear rule rather than the better-bands rule.<br>The ear with the majority of the target signal, and therefore least interference signal, provides a better intelligibility advantage. |
| Binns (2007) | Intelligibility - fundamental frequency of target | diotic | speech | speech shaped noise<br>speech | Manipulation of the fundamental frequency contour did not provide much of an advantage.<br>Performance was significantly better (SRT dropped by about 7dB) when the interferer is speech rather than speech-shaped noise, due to peek theory |
| Li (2007) | Intelligibility - peek theory | diotic | speech | Babble (20 talker) | It is better to have multiple small windows of the speech removed, rather than a single long window.<br>Confirmation that most of the energy in speech is below 3 kHz. |
| Warren (1997) | Intelligibility - spectral restoration | diotic | speech | noise | Inserting noise into a spectral gap within the target can almost double intelligibility levels, compared to having silence in the spectral gap. |

**Table 7-1: Summary of key points from literature review**

Binaural hearing is an example of a system which operates more effectively than the simple sum of its parts. It bestows a sense of spatial location on a sound source which is absent when listening through either ear individually. Generally, the greater the angular separation between a target speech source and an interfering sound, the easier it becomes to understand the target speech, a phenomenon known as the binaural advantage. The bilateral fitting of hearing aids potentially restores the binaural advantage in a passive sense, but does not enhance it. In the same way that a conventional hearing aid enhances monaural attributes of a speech signal to improve intelligibility, this research aims to enhance the binaural attributes of an auditory scene. The goal is to increase the binaural advantage and so raise intelligibility above what is achievable using a typical monaural or bilateral hearing aid configuration.

The more obvious ways of achieving this goal turn out to have problems. In this chapter, each of these is considered in turn to identify their shortcomings and these then inform the design of a new signal processing approach.

## *7.2 Direct sound source respatialisation*

The first step in investigating how to increase the apparent angular separation between two sound sources requires suitable test signals to be created. This involves spatialising mono sound sources by the application of different pairs of HRTFs. Presenting the resulting binaural signals over headphones causes each sound to be perceived at a different spatial location outside the listener's head. The process of spatialising a single mono sound source was discussed in Section 3.2.4. It is shown there, and repeated here for convenience, that an efficient method of achieving spatialisation is by multiplying the sound source spectrum with the left/right pair of HRTFs for the desired direction

For all the technical work performed in this investigation a set of HRTFs from the CIPIC database (CIPIC Interface Laboratory (1998)) was used. A detailed analysis of the HRTF sets has been carried out by Shoji (2007) who selected set

number 009 for his work. The nature of this research is subtly different from Shoji's, however, and for this reason set 021 was selected. Specifically, the spectral and temporal content for corresponding left/right HRTF pairs are well matched, which is desirable for the types of processing considered here. Informal listening showed that set 021 provided the author with equally good localisation cues. A block diagram illustrating the application of HRTF filters to a mono sound source to produce a binaural output signal is shown in Figure 7-1.

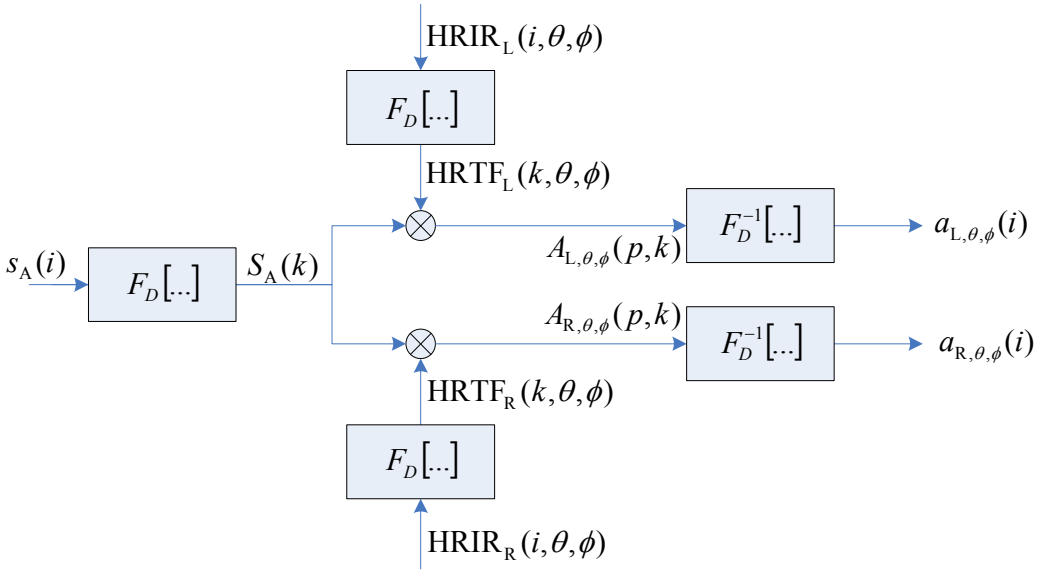



**Figure 7-1: The application of a pair of HRTFs to a mono sound to produce a pair of spatialised signals.**

To illustrate this, consider the spatialisation of a monophonic sound source **A** at some distance from a listener. Acoustic pressure variations from **A** can be described as a real-valued time-domain signal $S_A(i)$ , where $i$ is the sample time index over the signal's total duration. The naming convention used is summarised in Appendix E. The frequency domain representation is given by

$$S_A(k) = \mathrm{F_D}[s_A(i)] \qquad \textbf{Eq. 7-1}$$

where $0 \leq i \leq M\text{-}1$ and $M$ is the total number of samples in the input signal. Where $\mathrm{F_D}$, denotes the discrete Fourier transform (DFT) and $S_A(k)$ is the Fourier transform of the time domain samples $s_A(i)$. In practice, for example to perform

spectro-temporal analysis and reduce latency, the time domain signal $s_A(i)$ is split into $P$ temporal frames, indexed by $p = 0, 1, 2, \ldots, P-1$, each containing $N$ samples. Neighbouring frames are separated by $N/2$ samples. After transformation by the DFT, this results in a $N$ point frequency domain signal, $S_A(p, k)$, with $N/2$ distinct frequency points up to the Nyquist frequency, for each frame.

$$S_A(p,k) = \mathrm{F_D}\left\{ w(n)\, s_A\left[ n + \frac{pN}{2} \right] \right\} \quad \textbf{Eq. 7-2}$$

where:

$k = 0, 1, 2, \ldots, \frac{N}{2} - 1$ and $n = 0, 1, 2, \ldots, N-1$

$w(n)$ is an $N$-sample Hanning window function that is applied to the frame of time domain samples prior to transformation into the frequency domain. To spatialise the spectral components of the signal in the direction specified by azimuth $\theta$ and elevation $\phi$, it is necessary to multiply the frequency domain signal $S_A(p, k)$ by the left and right HRTF data, $\mathrm{HRTF_L}(k, \theta, \phi)$ and $\mathrm{HRTF_R}(k, \theta, \phi)$, respectively. The spatialised spectral components are hence

$$A_{\mathrm{L},\theta,\phi}(p,k) = \mathrm{HRTF_L}(k,\theta,\phi)\, S_A(p,k) \quad \textbf{Eq. 7-3}$$

$$A_{\mathrm{R},\theta,\phi}(p,k) = \mathrm{HRTF_R}(k,\theta,\phi)\, S_A(p,k) \quad \textbf{Eq. 7-4}$$

They are converted back to the time domain using the inverse DFT (IDFT) and 50% overlap-add (discussed in Section 3.2.4.2). This produces two spatialised output signals, $a_{\mathrm{L},\theta,\phi}(i)$ and $a_{\mathrm{R},\theta,\phi}(i)$ for the left and right channels, respectively:

$$a_{\mathrm{L},\theta,\phi}(i) = \mathrm{F_D^{-1}}\left[ A_{\mathrm{L},\theta,\phi}(p,k) \right] \quad \textbf{Eq. 7-5}$$

$$a_{\mathrm{R},\theta,\phi}(i) = \mathrm{F_D^{-1}}\left[ A_{\mathrm{R},\theta,\phi}(p,k) \right] \quad \textbf{Eq. 7-6}$$

where $p\frac{N}{2} \le i \le N\left(\frac{p}{2}+1\right) - 1$.

The spatialisation process is effectively a complex multiplication operation. In general, provided there are no zero values in the HRTF, the operation can be reversed by dividing out the HRTF data to recover the mono sound source. The

left and right spatialised signals should reduce to two identical mono signals when divided by the HRTF spectral filter coefficients that were used to spatialise the original mono sound, as shown in Figure 7-2.

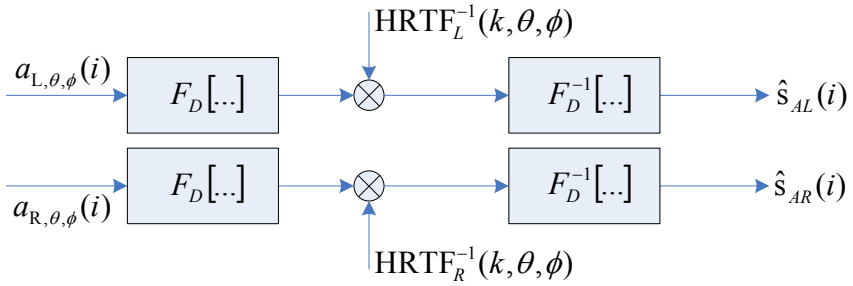



**Figure 7-2: The application of the inverse HRTFs (HRTF$^{-1}$) to divide out the spatialisation processing. Ideally, with perfect reconstruction, this produces two identical mono sounds.**

Assuming the spatialised sound source does not change direction, despatialisation of the left channel output $a_{\mathrm{L},\theta,\phi}(i)$ to form an estimate $\hat{s}_{\mathrm{AL}}(i)$ of the original input $s_{\mathrm{A}}(i)$ is achieved according to,

$$\hat{s}_{\mathrm{AL}}(i) = \mathrm{F}_{\mathrm{D}}^{-1}\left[\frac{\mathrm{F}_{\mathrm{D}}[a_{\mathrm{L}}(i)]}{\mathrm{HRTF}_{\mathrm{L}}(\theta,\phi)}\right] \qquad \textbf{Eq. 7-7}$$

where, in practice, Hanning windowing is again applied to the spatialised time domain signals to create a series of frames of data, as described above in equation Eq. 7-2. Similarly, for the right channel:

$$\hat{s}_{\mathrm{AR}}(i) = \mathrm{F}_{\mathrm{D}}^{-1}\left[\frac{\mathrm{F}_{\mathrm{D}}[a_{\mathrm{R}}(i)]}{\mathrm{HRTF}_{\mathrm{R}}(\theta,\phi)}\right] \qquad \textbf{Eq. 7-8}$$

Ideally, with perfect reconstruction, $\hat{s}_{\mathrm{AL}}(i)$ and $\hat{s}_{\mathrm{AR}}(i)$ will each be in-distinguishable from $s_{\mathrm{A}}(i)$. Once the original mono sound source has been retrieved, it can, if required, be respatialised to a new location, using a different HRTF pair, by applying the spatialisation processing technique shown in Figure 7-1 again.

## 7.3 Direction detection using a cross correlogram

In a real-world situation, spatialisation of a sound source occurs naturally, for example when a talker speaks in a room. It is entirely separate from the despatialisation and respatialisation processes which can be used to artificially increase the angular separation between a pair of sounds. In particular, the direction of the source to be moved is not known directly. Therefore a method is generally required for estimating it. This section considers how the spatial direction of a sound source can be determined from just the signals arriving at the ears. We are primarily concerned with speech sources. Since these tend to lie in, or close to, the horizontal plane, only sources in the horizontal plane will be considered.

### 7.3.1 Single sound source direction detection

Section 2.4 reviewed how, amongst other cues, the hearing system uses differences in the sounds arriving at each ear to determine the direction of a sound source. It therefore follows that one of the processing elements for binaural signal manipulation should be the comparison of time and frequency information between the left and right signals reaching the ears. The FFT, discussed in Section 3.2.4, is a popular means of converting a set of time domain data into the equivalent frequency domain data. Its use for analysing and processing binaural audio is considered later in Section 7.5. An alternative method is to use a set of overlapping bandpass filters that split the time-domain signal into multiple concurrent time-domain signals, one for each filter. These filters form an auditory filterbank. The individual signals from a filterbank can be recombined, or mixed, to form a single time-domain output signal. Auditory filterbanks can be designed to mimic the overlapping critical band filtering observed in the hearing system. One method for generating such overlapping responses uses bandpass gammatone filters with centre frequencies on the equivalent rectangular bandwidth (ERB) scale (Slaney, 1993). An example of 32 overlapping filters is shown in Figure 7-3.

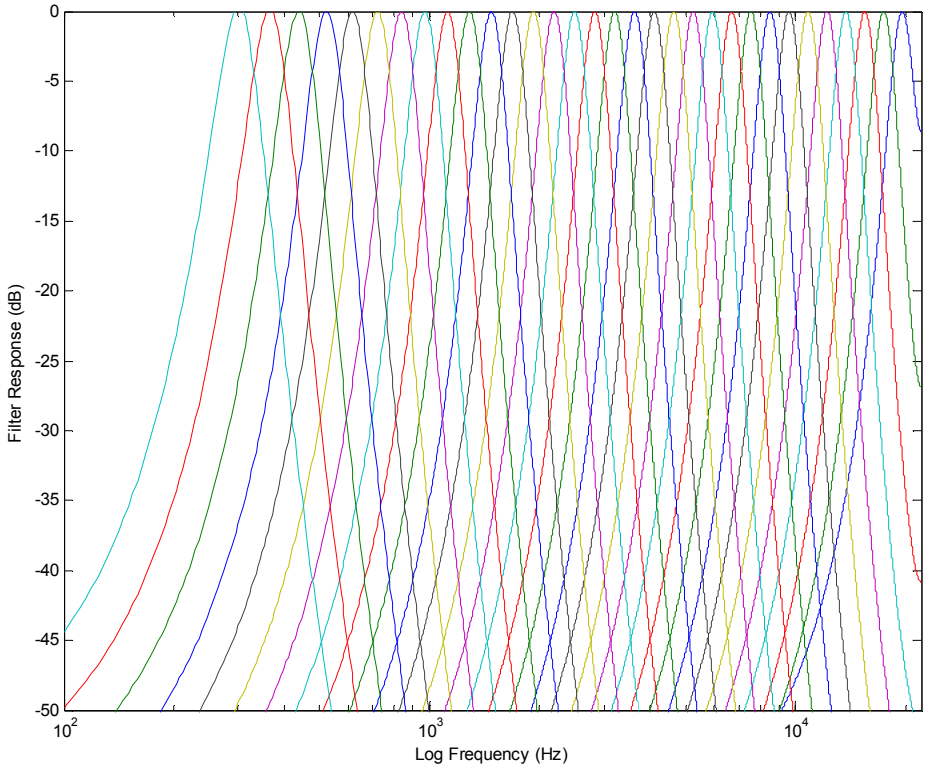



**Figure 7-3: 32 overlapping gammatone filter responses separated by their equivalent rectangular bandwidths (after Slaney, 1993).**

The processing architecture for analysing a general incoming pair of binaural signals, $x_L(i)$ and $x_R(i)$, is shown in Figure 7-4. Each temporal window (frame) of samples is passed through the filterbank to produce individual output signals which form a matrix of data containing both frequency and time information.

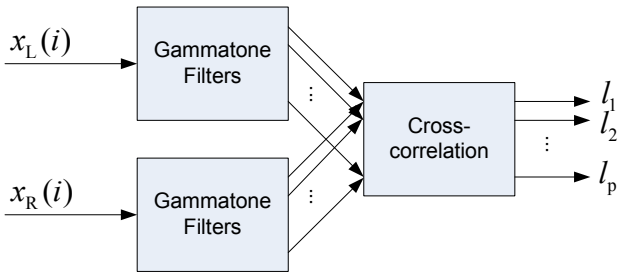



**Figure 7-4: The signal processing blocks for calculating ITD using cross-correlation. Left and right spatialised signals $x_L(i)$ and $x_R(i)$ pass through a bank of gammatone filters. Cross-correlation between the filter outputs over a time shift of $N$ samples produces $P$ temporal lags $l_p$ ($1 \leq p \leq P$).**

Each data point in the left channel time-frequency matrix can be compared with the corresponding point in the right matrix. Specifically, the left and right audio streams $x_L(i)$ and $x_R(i)$ can be analysed to determine if there are any temporal differences. A cross-correlation function is used to determine the $P$ frequency-dependent time lags $l_p$ ($1 \leq p \leq P$) between the left and right data. The concept of interaural cross-correlation has been investigated by Lindemann (1986) and is a key component of research by Wang and others (Wang and Brown (1999), Palomaki *et al*. (2001)), as part of a method for improving the intelligibility of a sound source. Offsets in the peak of the cross-correlation output from the midpoint which fall within the temporal analysis window of $N$ samples, give the time difference in samples between the left and right signals. The lag is computed for each of the data sets in the time-frequency matrix and can be converted into an ITD for each frequency band.

An example of the calculated ITD using cross-correlation is given in Figure 7-5. A mono speech sound has been artificially spatialised to 40° azimuth using HRTFs. The resulting left and right signals were processed using an auditory filterbank followed by the cross-correlation method, described by the block diagram in Figure 7-4.

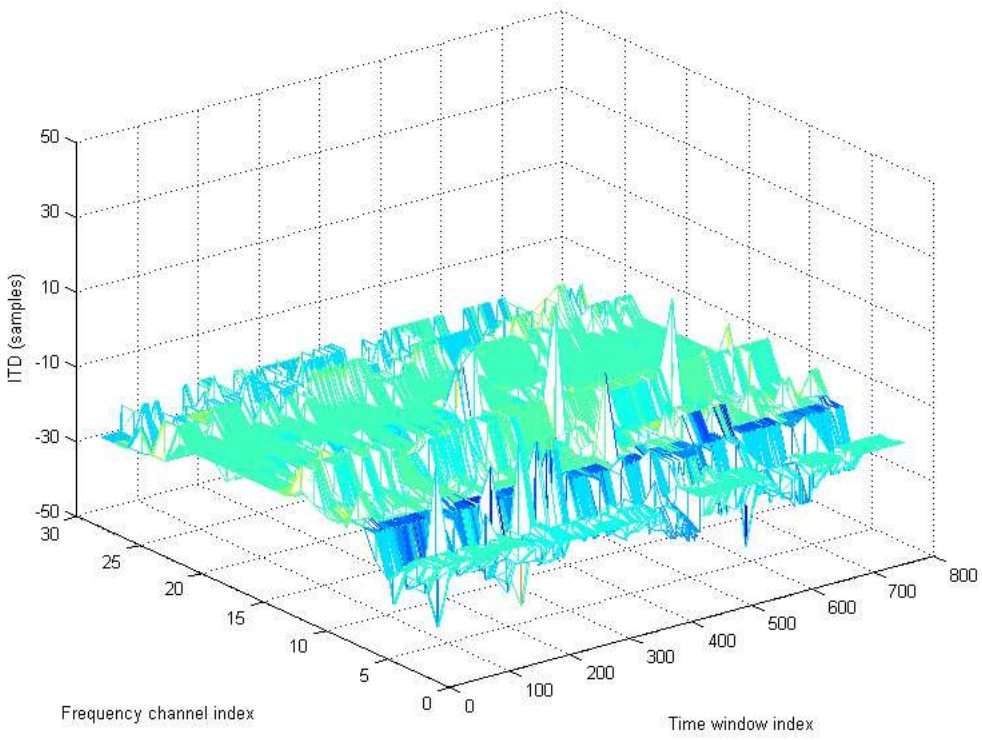



**Figure 7-5: Estimated ITD using the cross-correlation of $P$ = 28 frequency channels from a gammatone filterbank. The frequency channels cover the range from 100 Hz to 1500 Hz. Each temporal analysis window has $N$ = 128 samples and there is a 50% overlap with the next window.**

A perfect ITD calculation would be represented by a flat horizontal plane at the corresponding ITD value on the vertical axis of Figure 7-5. A few spurious lags have been calculated due to the mismatch in phase changes between the left and right signals spatialised using HRTFs and the estimated ITD using cross-correlation. However, in general, the ITD is predominantly consistent across both time and frequency dimensions. It may be concluded that, for a single sound source, the cross-correlation method is reasonably robust for estimating ITDs.

Roman *et al*. (2003) have developed an improved direction detection method based on a cross-correlogram algorithm. Their approach is analysed in the context of its potential suitability for the despatialisation process within the

respatialisation algorithm. The direction detection algorithm developed by Roman *et al*. is shown in Figure 7-6.

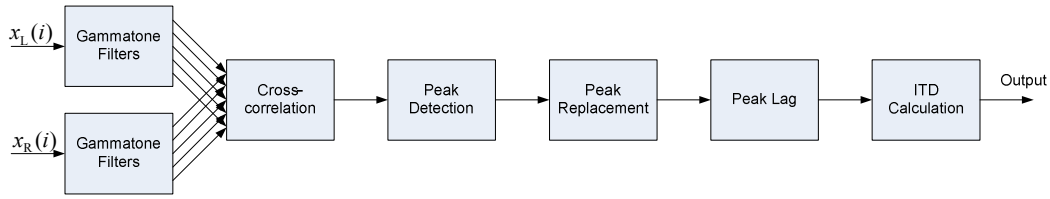



**Figure 7-6: Block diagram based on the system by Roman *et al.* (2003) for calculating ITD for a binaural spatialised signal.**

Binaural signals $x_L(i)$ and $x_R(i)$ are presented to the gammatone filters. The creation of a cross-correlogram begins by performing repeated *N*-sample cross-correlation between left/right outputs from a pair of gammatone filterbanks. This produces one time lag per frame for each frequency channel. However, the peak lag for each frequency channel for a particular frame is replaced with a Gaussian-shaped curve. This produces a "skeleton" (Palomaki *et al.*, 2001) of data which can be pooled across the frequency dimension to produce a clear peak for determining the ITD of each sound source. An example of this is shown in Figure 7-7.

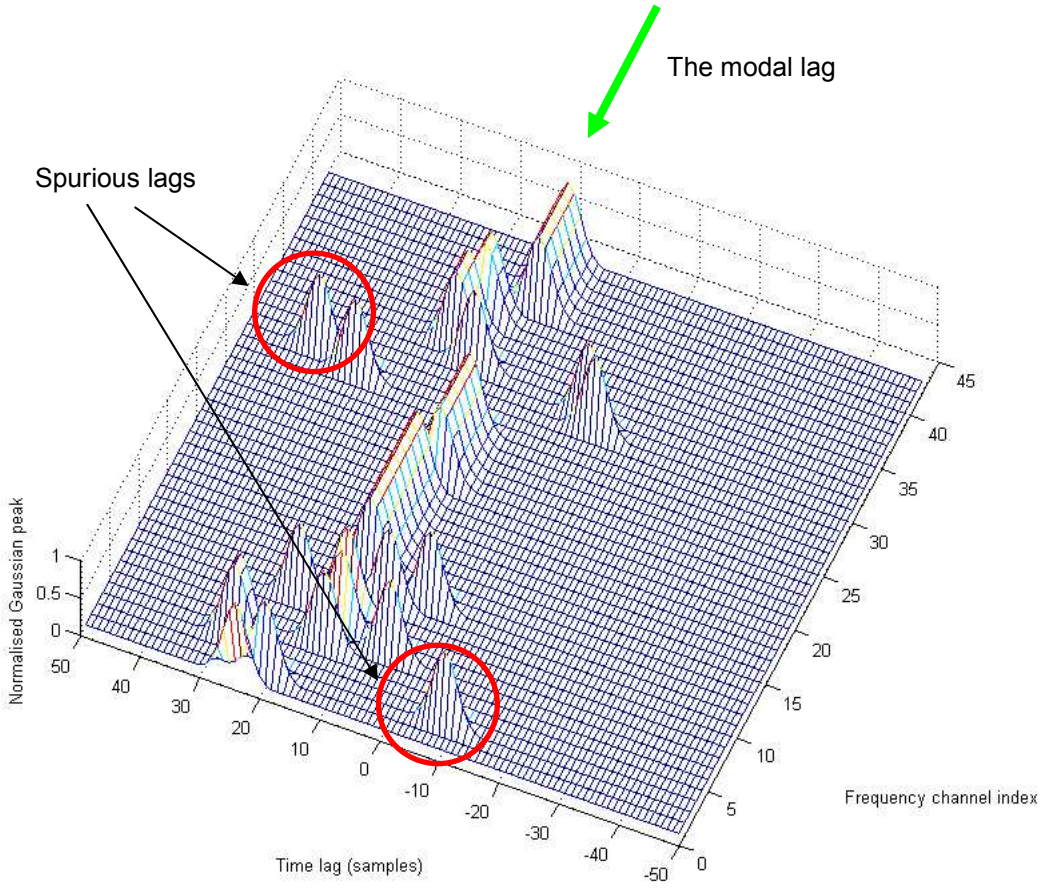



**Figure 7-7: Plot of the peak lags for a single frame and a single sound source spatialised to -40° azimuth for frequencies up to approximately 1500 Hz after Palomaki *et al.* (2001).**

This shows a trend in the detected lag of approximately 14 or 15 samples, marked by the arrow. Several frequency channels have radically different lags, for example, channels 3 and 29, which are highlighted by red circles. The variation in the lag is chiefly due to the phase responses of the HRTFs, which do not have a precisely linear relationship between the left and right channels. Therefore, the cross-correlation analysis of the gammatone filterbank outputs will potentially produce a different peak lag for each filter.

To provide more robust direction information a level of confidence is assigned to each peak time lag. This is achieved by averaging across frequencies within a time window to determine the dominant lag. Each individual lag component is compared to the average lag to determine whether it is likely to be correct or not. This removes the outliers, as they are ignored in favour of the directions that

have a higher confidence value. To illustrate how this improves the ITD estimate, the plot in Figure 7-8 shows the average calculated time lags for a single sound source, spatialised to -40° azimuth.

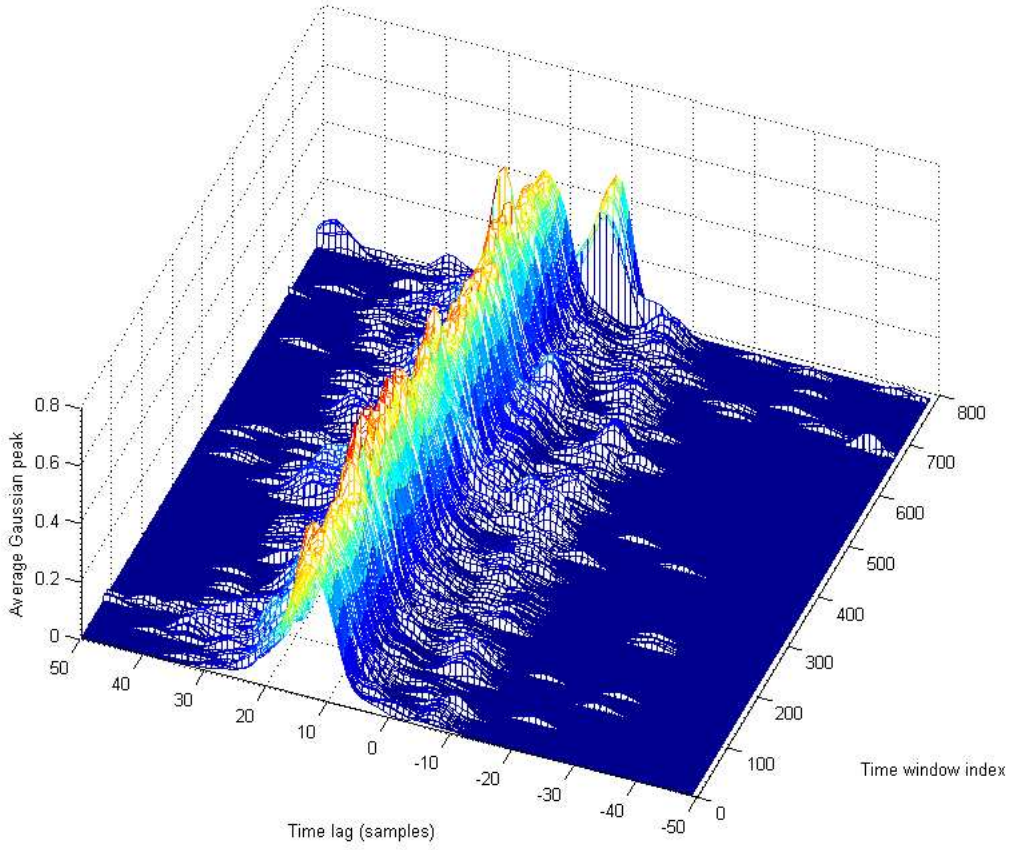



**Figure 7-8: Cross-correlogram for a single sound source spatialised to -40° azimuth.**

There is a dominant lag of approximately 14 samples (320μs) across the majority of time windows. For comparison, the peak lag across time is shown in Figure 7-9 along with the actual ITD calculated from the HRTF set used to spatialise the sound. The HRTF ITD has been rounded to the nearest sample time lag. This shows that the detected ITD is reasonably consistent and close to the actual ITD from the HRTFs, although, there is not a perfect match between the estimated and actual ITD. It should be noted that the time windows at the right of the plot have a calculated ITD of 2 samples (i.e. 0° azimuth), this discrepancy is due to the signal at the end of the file being almost silent.

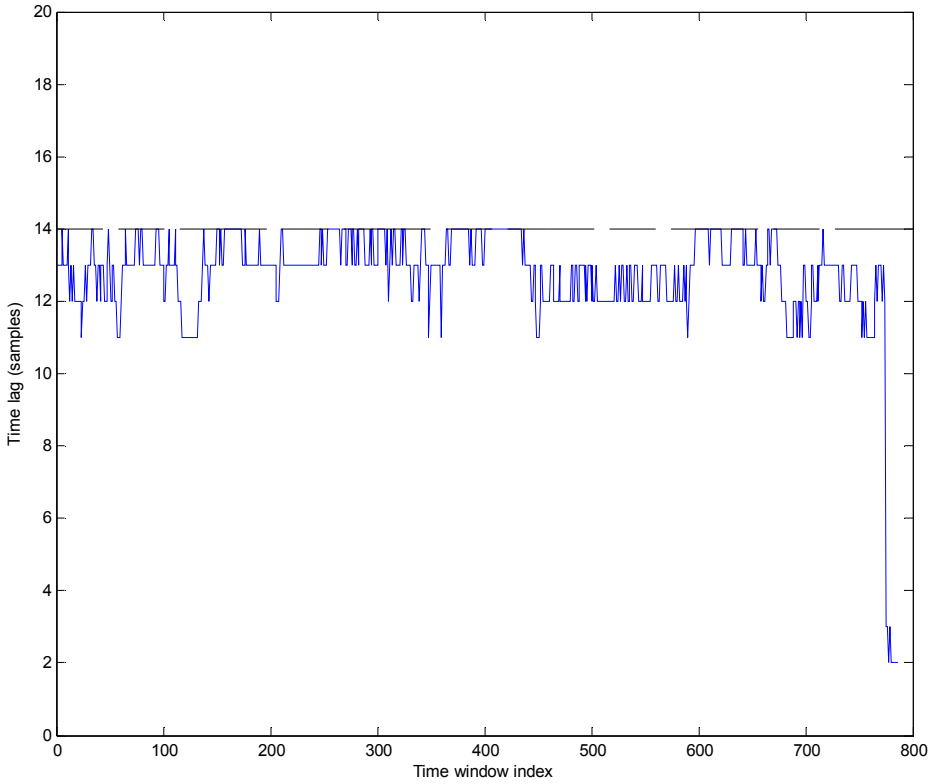



**Figure 7-9: The dominant time lag in each time window for a single sound source spatialised to -40° (blue solid line). The estimated ITD for -40° azimuth for the HRTF data set is also shown (black dashed).**

### 7.3.2 Respatialisation involving two sound sources

So far, we have considered only one active sound source at a time. This research, however, is concerned with increasing the intelligibility of a speech target by increasing the spatial separation between the speech and an interferer. Hence, we next simulate two sound sources, spatialised in different directions, and attempt to respatialise one of them.

When two spatialised sound sources, **A** and **B**, are presented to a listener, one of the following will be heard at each ear; no sound; only sound **A**; only sound **B**; a mixture of sound **A** and sound **B**.

Direct-sound activity at the two ears is highly correlated, i.e., in general, if only sound **A** is heard at the left ear, then only a very similar copy of sound **A** is heard at the right ear, etc. The binaural frequency-domain signals, $X_{\mathrm{L}}(k)$ and $X_{\mathrm{R}}(k)$, arriving at the left and right ears, respectively, can be written as simultaneous equations in two variables:

$$X_{\mathrm{L}}(k) = S_A(k).\mathrm{HRTF}_{\mathrm{L}}(k,\alpha) + S_B(k).\mathrm{HRTF}_{\mathrm{L}}(k,\beta) \qquad \textbf{Eq. 7-9}$$

$$X_{\mathrm{R}}(k) = S_A(k).\mathrm{HRTF}_{\mathrm{R}}(k,\alpha) + S_B(k).\mathrm{HRTF}_{\mathrm{R}}(k,\beta) \qquad \textbf{Eq. 7-10}$$

where:

$\mathrm{HRTF}_{\mathrm{L}}(k, \alpha)$ and $\mathrm{HRTF}_{\mathrm{R}}(k, \alpha)$ denote the left and right HRTF data, respectively, for direction $\alpha$;

$\mathrm{HRTF}_{\mathrm{L}}(k, \beta)$ and $\mathrm{HRTF}_{\mathrm{R}}(k, \beta)$ denote the left and right HRTF data, respectively, for direction $\beta$;

$S_A(k)$ and $S_B(k)$ are the spectra associated with the original mono sounds **A** and **B**, respectively.

When the direction of each sound source is known, it is generally possible to solve the equations exactly to obtain the original mono signals $A$ and $B$.

$$S_A(k) = \frac{X_L(k) - S_B(k).\mathrm{HRTF}_{\mathrm{L}}(k,\beta)}{\mathrm{HRTF}_{\mathrm{L}}(k,\alpha)} \qquad \textbf{Eq. 7-11}$$

$$S_B(k) = \frac{X_R(k) - S_A(k).\mathrm{HRTF}_{\mathrm{R}}(k,\alpha)}{\mathrm{HRTF}_{\mathrm{R}}(k,\beta)} \qquad \textbf{Eq. 7-12}$$

$$S_A(k) = \frac{\left(\dfrac{X_{\mathrm{L}}(k)}{\mathrm{HRTF}_{\mathrm{L}}(k,\alpha)} - \dfrac{X_{\mathrm{R}}(k)}{\mathrm{HRTF}_{\mathrm{R}}(k,\beta)}\right)}{\left(\dfrac{\mathrm{HRTF}_{\mathrm{L}}(k,\alpha)}{\mathrm{HRTF}_{\mathrm{L}}(k,\beta)} - \dfrac{\mathrm{HRTF}_{\mathrm{R}}(k,\alpha)}{\mathrm{HRTF}_{\mathrm{R}}(k,\beta)}\right)} \qquad \textbf{Eq. 7-13}$$

$$S_B(k) = \frac{\left(\dfrac{X_{\mathrm{R}}(k)}{\mathrm{HRTF}_{\mathrm{R}}(k,\beta)} - \dfrac{X_{\mathrm{L}}(k)}{\mathrm{HRTF}_{\mathrm{L}}(k,\alpha)}\right)}{\left(\dfrac{\mathrm{HRTF}_{\mathrm{R}}(k,\beta)}{\mathrm{HRTF}_{\mathrm{R}}(k,\alpha)} - \dfrac{\mathrm{HRTF}_{\mathrm{L}}(k,\beta)}{\mathrm{HRTF}_{\mathrm{L}}(k,\alpha)}\right)} \qquad \textbf{Eq. 7-14}$$

This analytic approach quickly breaks down when more sources exist, which is likely to occur in a real listening environment. The simultaneous equations become underdetermined and so the solution cannot be easily generalised.

## 7.4 Respatialisation using the dominance method

In recent years an alternative approach to respatialising binaural signals, based upon signal dominance, has met with considerable success (Faller and Baumgarte (2001), Kelly and Tew (2002) and Shoji (2007)).

With two simultaneous sound sources, **A** and **B**, there will generally be energy from more than one source in each time-frequency (T-F) component or unit (TFU). In the dominance detection method, introduced in Section 5.2.1, signals from the source arriving at the ear with the greatest energy in a particular TFU are considered to be the only source present. If the dominant component belongs to a source requiring respatialisation, the HRTFs are applied to the entire mix of signals in that TFU, meaning that the subdominant components will be wrongly processed. Perceptually, it is found that refinements of this method are capable of high quality respatialisation with very few artefacts (Shoji, 2007).

Signal dominance is computed using Eq. 7-15 and Eq. 7-16. Every frequency component within a frame defines two binary-valued spectral masks, $F_{\mathrm{MA}}(n, k)$ and $F_{\mathrm{MB}}(n, k)$, based on the relative dominance of the spectral components of signals $A(n, k)$ and $B(n, k)$. There are two pairs of these masks, one pair based on the relative source energies in each TFU for the left channel and a similar pair for the right channel. Eq. 5-2 and Eq. 5-3 in Section 5.2.1, for determining the dominance for two sound sources, are adapted here to apply specifically to the left channel:

$$\text{If } |A_{\mathrm{L}}(n,k)| > |B_{\mathrm{L}}(n,k)| \text{ then } F_{\mathrm{MAL}}(n,k) = 1, F_{\mathrm{MBL}}(n,k) = 0 \qquad \textbf{Eq. 7-15}$$

$$\text{If } |A_{\mathrm{L}}(n,k)| \leq |B_{\mathrm{L}}(n,k)| \text{ then } F_{\mathrm{MAL}}(n,k) = 0, F_{\mathrm{MBL}}(n,k) = 1 \qquad \textbf{Eq. 7-16}$$

where:

$A_L(n, k)$ and $B_L(n, k)$ denote the temporal-spectral components for time-domain signals $a_L(n)$ and $b_L(n)$, respectively, in a particular frame,
$F_{MAL}(n, k)$ is the binary dominance spectral mask for signal A,
$F_{MBL}(n, k)$ is the binary dominance spectral mask for signal B,
$n$ identifies a temporal sample in a particular frame,
$k$ identifies a frequency point in the same frame,
$0 \leq n \leq N-1$, $0 \leq k \leq N-1$ and $N$ is the size of temporal window.

Each unity-valued frequency component in the binary dominance mask $F_{MAL}(n, k)$ indicates that signal **A** is dominant for this TFU. Conversely, each unity-valued frequency component in $F_{MBL}(n, k)$ indicates that signal **B** is dominant. The unity-valued entries in each mask define which TFU signal components will pass through the mask, whereas the zero-valued entries indicate which TFU signal components will be blocked. Two similar masks operate in the right channel.

A diagram illustrating the application of the dominance method to two spatialised sound sources, **A** and **B**, is shown in Figure 7-10. The sources are described by two $M$-sample mono signals, $s_A(i)$ and $s_B(i)$, $(0 \leq i \leq M-1)$. These are segmented into $P$ temporal frames and spatialised to different fixed directions $\alpha$ and $\beta$ to form signal spectral pairs $A_{L\alpha}(p, k)$ and $A_{R\alpha}(p, k)$, and $B_{L\beta}(p, k)$ and $B_{R\beta}(p, k)$, respectively $(0 \leq p \leq P-1)$. In its present simplified form, this method of dominance detection assumes there is access to the unmixed mono TFUs for **A** and **B**, that is, $S_A(p, k)$ and $S_B(p, k)$, respectively. Hence, in each frame the spectral magnitudes for the left channels of each spatialised sound are compared to determine which is dominant.

With reference to the left channel in Figure 7-10, the dominant TFUs in $A_{L\alpha}(p, k)$ pass unchanged directly to the output, $Y_L(p, k)$. On the other hand, the dominant TFU components in signal $B_{L\beta}(p, k)$ are respatialised before being passed to output $Y_L(p, k)$. This is accomplished by using $HRTF_L(k, \beta)$ to divide out the appropriate components and applying the corresponding component of $HRTF_L(k, \gamma)$, the HRTF for the new direction, $\gamma$ (see Section 7.2). Note that this process

corrupts all subdominant components of $A_{L\alpha}(p, k)$ (i.e. the spectral magnitude components of $A_{L\alpha}(p, k)$ which are not dominant in each TFU).

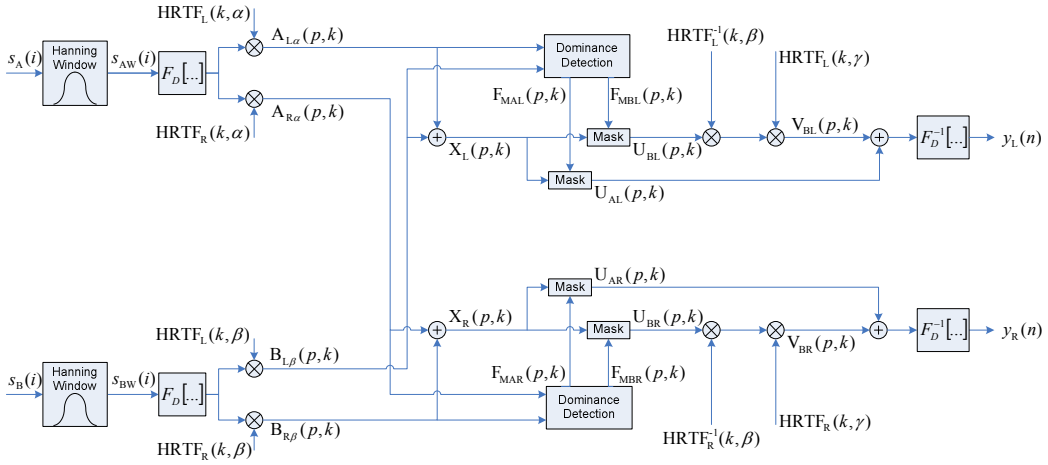



**Figure 7-10: The processing system for determining the dominant left and right spatialised signals for sounds A and B, spatialised using HRTFs for directions *α* and *β*, respectively. Then dominant components of B are respatialised to a new location using the HRTF for direction *γ*.**

For each frame, the spatialised spectral signals in Figure 7-10, $A_{L\alpha}(p, k)$, $A_{R\alpha}(p, k)$, $B_{L\beta}(p, k)$ and $B_{R\beta}(p, k)$ are given by:

$$A_{L\alpha}(p,k) = \mathrm{F_D}[s_{AW}(n)]\mathrm{HRTF_L}(k,\alpha) \qquad \textbf{Eq. 7-17}$$

$$A_{R\alpha}(p,k) = \mathrm{F_D}[s_{AW}(n)]\mathrm{HRTF_R}(k,\alpha) \qquad \textbf{Eq. 7-18}$$

$$B_{L\beta}(p,k) = \mathrm{F_D}[s_{BW}(n)]\mathrm{HRTF_L}(k,\beta) \qquad \textbf{Eq. 7-19}$$

$$B_{R\beta}(p,k) = \mathrm{F_D}[s_{BW}(n)]\mathrm{HRTF_R}(k,\beta) \qquad \textbf{Eq. 7-20}$$

where $n$ and $k$ have their usual meanings and $s_{AW}(n)$ and $s_{BW}(n)$ are the windowed time domain samples for signals **A** and **B** respectively.

The mixed signals, $X_L(p, k)$ and $X_R(p, k)$ are given by the sum of the appropriate spatialised signal components:

$$X_L(p,k) = A_{L\alpha}(p,k) + B_{L\beta}(p,k) \qquad \textbf{Eq. 7-22}$$

$$X_R(p,k) = A_{R\alpha}(p,k) + B_{R\beta}(p,k) \qquad \textbf{Eq. 7-24}$$

The dominance decision for each spectral component is given by:

$$If\ |A_{L\alpha}(p,k)| > |B_{L\beta}(p,k)|\ then\ F_{MAL}(p,k)=1,\quad F_{MBL}(p,k)=0 \qquad \textbf{Eq. 7-25}$$

$$If\ |A_{L\alpha}(p,k)| \leq |B_{L\beta}(p,k)|\ then\ F_{MAL}(p,k)=0,\quad F_{MBL}(p,k)=1 \qquad \textbf{Eq. 7-26}$$

$$If\ |A_{R\alpha}(p,k)| > |B_{R\beta}(p,k)|\ then\ F_{MAR}(p,k)=1,\quad F_{MBR}(p,k)=0 \qquad \textbf{Eq. 7-27}$$

$$If\ |A_{R\alpha}(p,k)| \leq |B_{R\beta}(p,k)|\ then\ F_{MAR}(p,k)=0,\quad F_{MBR}(p,k)=1 \qquad \textbf{Eq. 7-28}$$

This generates four masks, two for each signal stream. The masks for each signal stream are the inverse of each other. The four masked signals, $U_{AL}(p,k)$, $U_{BL}(p,k)$, $U_{AR}(p,k)$ and $U_{BR}(p,k)$, are produced by applying the appropriate mask to the mixed signals:

$$U_{AL}(p,k)=X_L(p,k)F_{MAL}(p,k) \qquad \textbf{Eq. 7-29}$$

$$U_{BL}(p,k)=X_L(p,k)F_{MBL}(p,k) \qquad \textbf{Eq. 7-30}$$

$$U_{AR}(p,k)=X_R(p,k)F_{MAR}(p,k) \qquad \textbf{Eq. 7-31}$$

$$U_{BR}(p,k)=X_R(p,k)F_{MBR}(p,k) \qquad \textbf{Eq. 7-32}$$

This operation has the effect of multiplying all the dominant spectral components by 1 and all the non-dominant spectral components by 0. The final stage of processing is to respatialise all the TFUs that are assigned as a dominant interferer and to leave the TFUs alone where the target is dominant so that they retain their existing spatial cues. This produces the four signals

$$V_{AL}(p,k)=U_{AL}(p,k) \qquad \textbf{Eq. 7-33}$$

$$V_{BL}(p,k)=\frac{U_{BL}(p,k)}{\mathrm{HRTF}_L(k,\beta)}.\mathrm{HRTF}_L(k,\gamma) \qquad \textbf{Eq. 7-34}$$

$$V_{AR}(p,k)=U_{AR}(p,k) \qquad \textbf{Eq. 7-35}$$

$$V_{BR}(p,k)=\frac{U_{BR}(p,k)}{\mathrm{HRTF}_R(k,\beta)}.\mathrm{HRTF}_R(k,\gamma) \qquad \textbf{Eq. 7-36}$$

The combined spectral components in each frame can then be returned to the time domain.

$$y_L(n) = \mathrm{F}_\mathrm{D}^{-1}\left[V_\mathrm{AL}(p,k) + V_\mathrm{BL}(p,k)\right] \quad \textbf{Eq. 7-37}$$

$$y_R(n) = \mathrm{F}_\mathrm{D}^{-1}\left[V_\mathrm{AR}(p,k) + V_\mathrm{BR}(p,k)\right] \quad \textbf{Eq. 7-38}$$

To evaluate the effectiveness of respatialisation using the dominance method described above, a binaural test signal, $X_\mathrm{S}$, was created by spatialising the sources **A** and **B** in directions 0° azimuth and -80° azimuth, respectively, using the standard method shown in Figure 7-1. A second signal, $X_\mathrm{D}$, was prepared using the dominance method. To create it, the two sound sources **A** and **B** were initially spatialised with **A** at 0° azimuth and **B** at 80° azimuth, respectively. Each of the dominant **B** components were then respatialised using the dominance method to -80° azimuth. This change in direction represents an extreme movement of the dominant spectral components for sound source **B**. Signals $X_\mathrm{S}$ and $X_\mathrm{D}$ were then informally compared perceptually. The listening analysis focussed on whether the audio processed using the dominance method matched the directly spatialised audio. Specifically, consideration was given to whether any of the subdominant **B** components (which had not been processed) could still be heard at 80°, i.e. in their original direction, and whether the processed **B** components had successfully been moved across the median plane to -80°. When the signal $X_\mathrm{D}$ created using the dominance method was informally compared with the directly spatialised signal $X_\mathrm{S}$ they were found to be perceptually very similar and the speech sounds used for sound source **B** were successfully respatialised. Furthermore, sound source **A** remained at 0°.

### 7.4.1 Source direction estimation of two directional sources

The success of the dominance mask method for respatialising a single source in a mixture of two spatialised sources depends on knowing the direction of the source to be moved. In the previous example the direction was fully specified. However, as for the real-life case of respatialising a single source in isolation,

discussed in Section 7.2, in practice, the direction of the source will have to be estimated using the incoming binaural signal.

Once again, the cross-correlogram method (Section 7.3) is applied, but this time in a two-source scenario. The method begins by computing the cross-correlation of the outputs from a gammatone filterbank for two simultaneous mono speech signals, artificially spatialised to 20° and -40°, respectively. These are presented in Figure 7-11. It can be seen that the cross-correlation process alone is unable to produce clean peaks due to the interference between the sound sources. The result is a chaotic estimate of ITD, which is unsuitable for the respatialisation process using HRTFs.

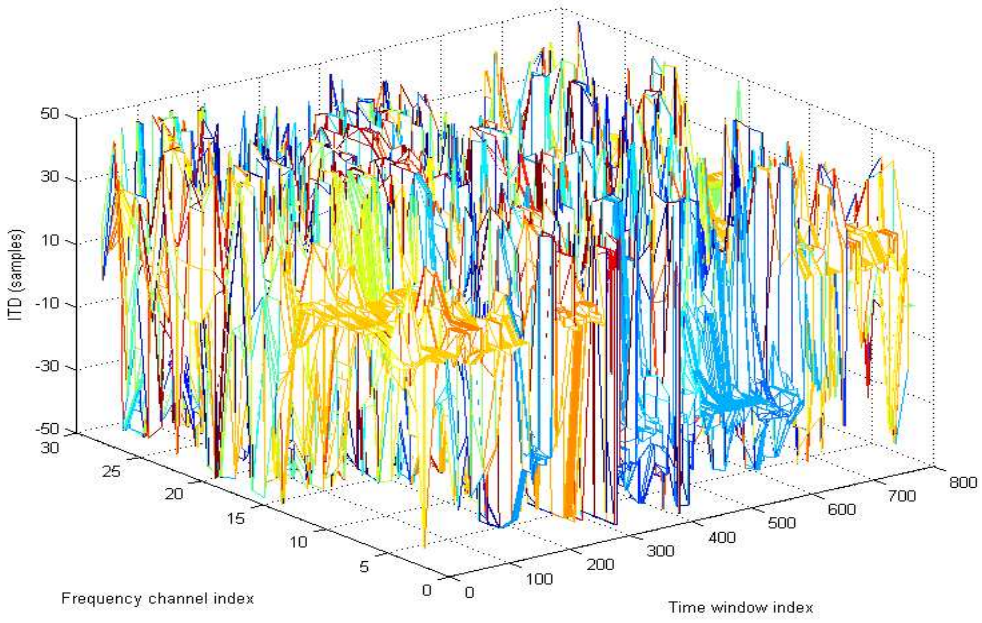



**Figure 7-11: The chaotic estimate of ITDs calculated for two speech sounds, artificially spatialised using generic HRTFs. Sound A is a male spoken word "crowded" spatialised to 20° azimuth. Sound B is a male spoken word "friends" spatialised to -40° azimuth.**

By contrast, when this result is processed according to the method by Roman *et al.*, two distinct lags are revealed, indicating two dominant directions. These are highlighted in Figure 7-12 by the green arrows. Compared with the corresponding result for a single source in Figure 7-8 there is an increase in the

number of incorrect directions being detected, indicated by the large number of smaller random peaks.

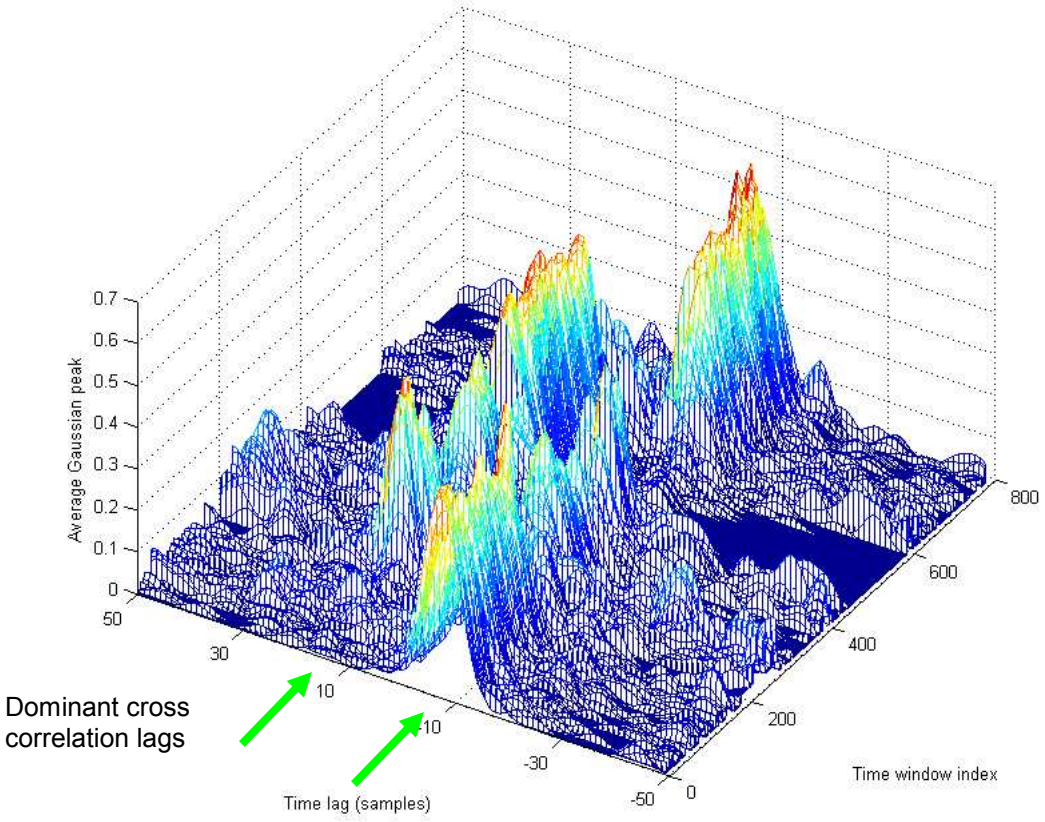



**Figure 7-12: The cross correlogram for two sound sources spatialised to -40° and 20° azimuth.**

The time lags for each of the peaks are shown in Figure 7-13, with the actual ITDs for the azimuths used superimposed with dotted lines. The ITD values are rounded to the nearest sample time lag. There are time intervals where there is an accurate estimate of one of the dominant directions. However, the errors are more frequent than in the single-source case and the estimated ITDs are significantly different from the actual directions the sounds are spatialised to. A particularly bad region at around time window index 250 is highlighted with a red oval.

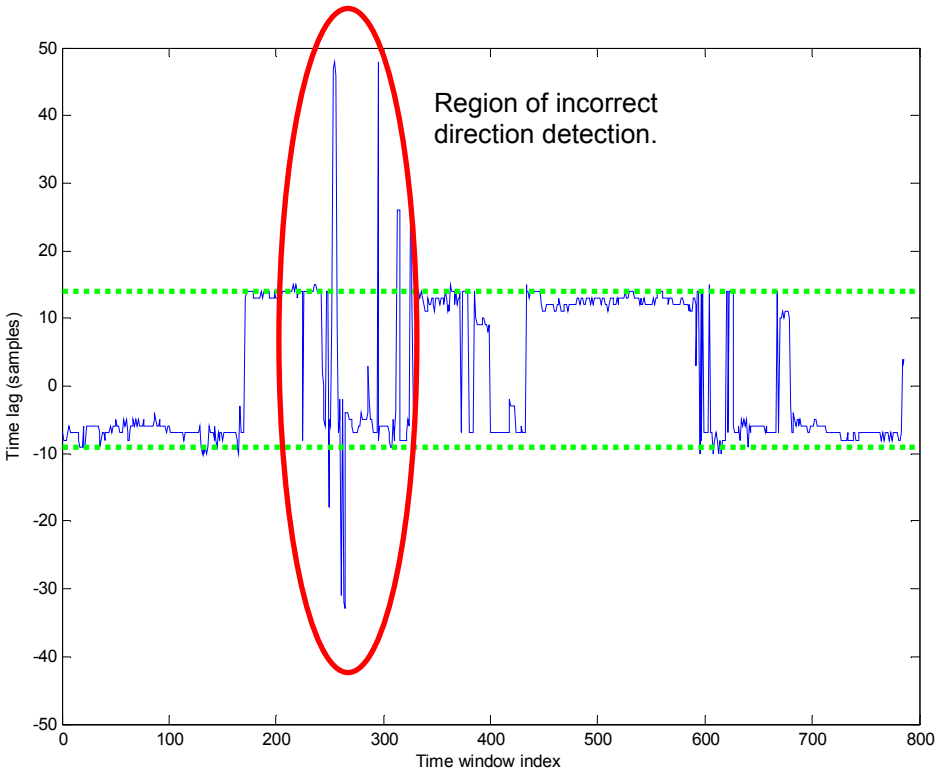



**Figure 7-13: The dominant directions detected for two sound sources spatialised to -40° and 20° (blue solid line). The actual ITDs for the HRTFs used are also shown (green dotted lines).**

The cross-correlogram method is much more successful at estimating the dominant direction of a sound within a single time window than cross-correlation alone. There still remain, however, a number of issues with this approach to respatialisation using direction estimation. Firstly, the direction is determined as an average over all frequency bands within the gammatone filterbank for frequencies below the ITD/ILD threshold. In the examples shown, this frequency is set to approximately 1500 Hz. The act of averaging across the frequency dimension dilutes the direction accuracy for each individual spectral component. Secondly, although ITD is acknowledged to be the dominant localisation cue at lower frequencies, phase difference ambiguities mean that it cannot provide directional information for the higher frequencies where ILD becomes an important localisation cue. The respatialisation technique using HRTFs requires direction information for every frequency band being processed. This is not available with the single estimated ITD value per time window.

Thirdly, assigning the same direction to every frequency channel risks disrupting the auditory scene if those time-frequency components are processed with the incorrect HRTF filters. Finally, as the level of the target source, **A**, relative to the interferer, **B**, becomes lower, it becomes subdominant in more TFUs and hence becomes disrupted when the dominant interferer in these TFUs is respatialised.

A further practical problem arises. In the proposed hearing aid application the intention is to move interferer **B** so as to increase the angular separation between it and the target sound, **A**. It seems reasonable to consider the impact of not only moving **B**, but also of attenuating it, to reduce the masking effects of **B** on **A**. The processing system used to achieve this is given in Figure 7-14. In this example the attenuation is represented in its extreme form by a gain of zero.

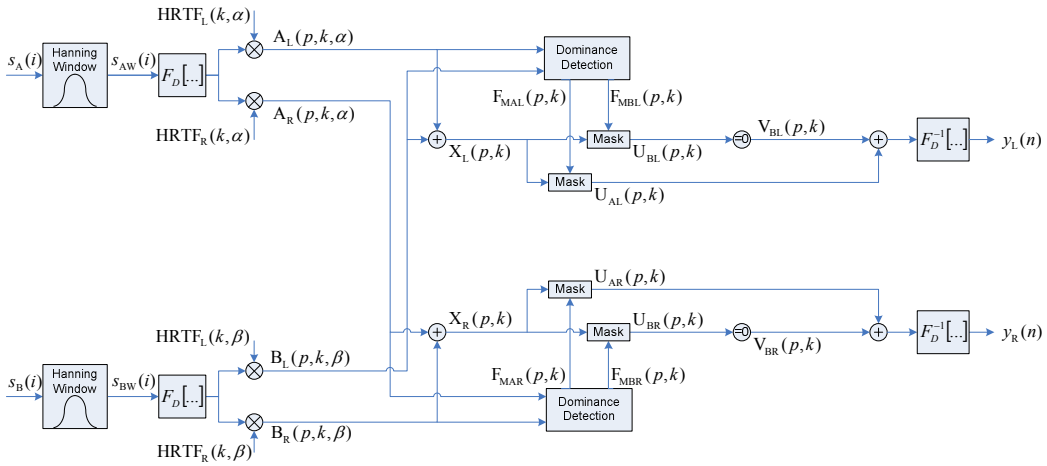



**Figure 7-14: The processing system used for setting all the spectral components to zero where the interferer is dominant**

The signal processing required to achieve this is very similar to that used in the system described in Figure 7-10. The only difference is that the mask calculations in Eq. 7-25 to Eq. 7-28 are altered to become:

$$If\ |A_L(p,k,\alpha)| > |B_L(p,k,\beta)|\ \ then\ \ F_{MAL}(p,k)=1,\quad F_{MBL}(p,k)=0 \qquad \textbf{Eq. 7-39}$$

$$If\ |A_L(p,k,\alpha)| \le |B_L(p,k,\beta)|\ \ then\ \ F_{MAL}(p,k)=0,\quad F_{MBL}(p,k)=0 \qquad \textbf{Eq. 7-40}$$

$$If\ |A_R(p,k,\alpha)| > |B_R(p,k,\beta)|\ \ then\ \ F_{MAR}(p,k)=1,\quad F_{MBR}(p,k)=0 \qquad \textbf{Eq. 7-41}$$

$$If\ |A_{\mathrm{R}}(p,k,\alpha)| \le |B_{\mathrm{R}}(p,k,\beta)|\ \ then\ \ F_{\mathrm{MAR}}(p,k)=0,\quad F_{\mathrm{MBR}}(p,k)=0 \qquad \textbf{Eq. 7-42}$$

This illustrates that all of the non-dominant TFUs are set to zero and therefore only the TFUs where the target is dominant pass through.

When the dominant interference signal components are reduced, all the subdominant components of **A** will forcibly be reduced too. This will cause spectral holes to appear in the target signal. The effect on the spectral content of the left and right processed output signals is illustrated in Figure 7-15, which shows the dominant spectral components of the target that remain when all of the dominant components of the interferer signal have been removed by setting them to zero.

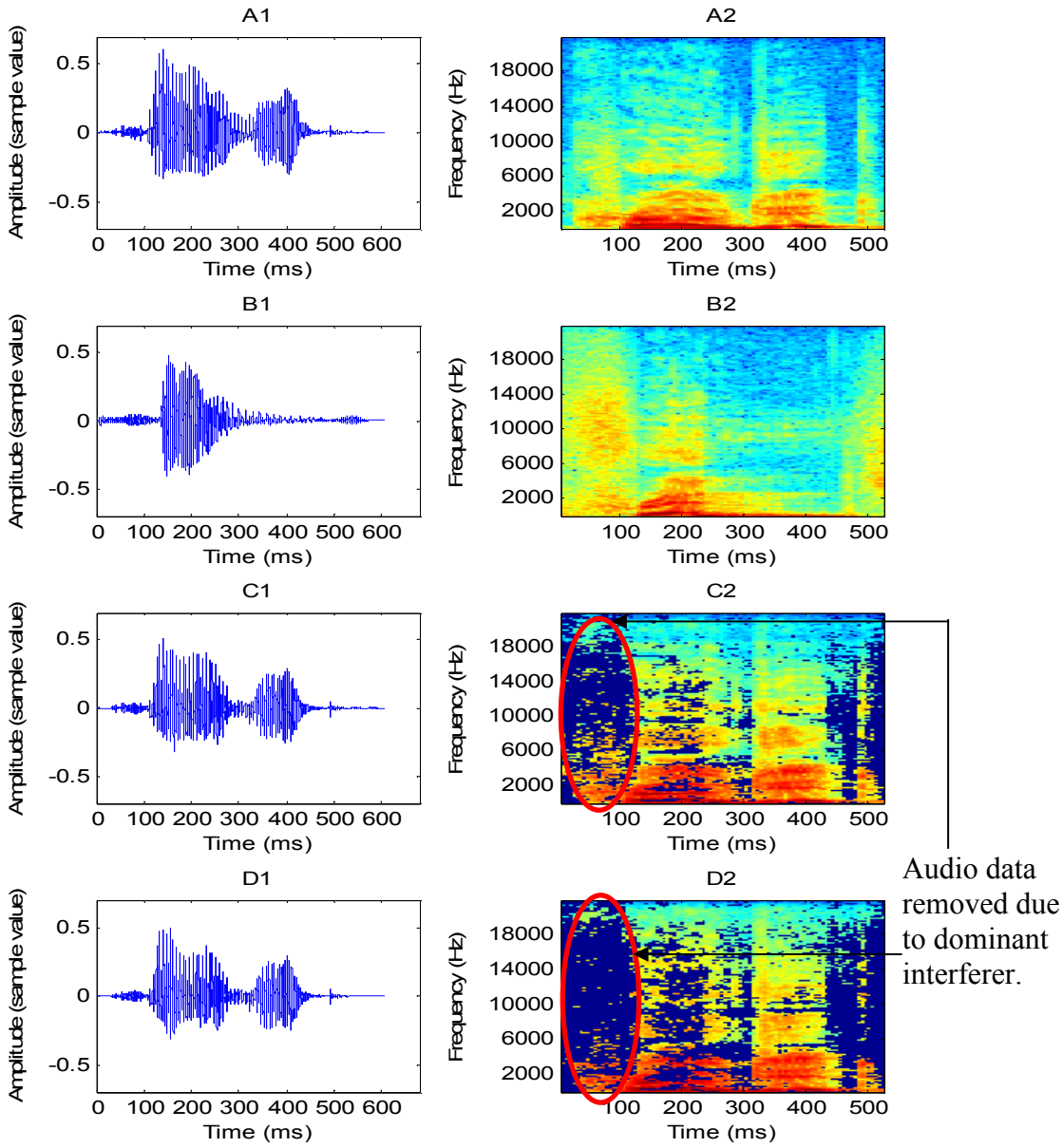



**Figure 7-15: Time domain and spectral plots of target signal A (a) and interferer signal B (b). (c) and (d) are the left and right channel outputs $Y_L(p,k)$ and $Y_R(p,k)$, respectively, of the spatialised and mixed output when the target-only binary-masks, $F_{MAL}[p, k]$ and $F_{MAR}[p, k]$, are applied. The red ovals highlight the areas where audio data has been removed, corresponding to the interferer sound being dominant.**

Many of the dominant interferer components mask the subdominant components of the target. This induces auditory continuity in the target signal, which is the

hearing system's way of moderating the masking effect. When the dominant interferer components are removed, so are the corresponding subdominant components of the target. Hence, continuity of the target speech is disrupted by the spectral holes created in it, which in turn reduces its intelligibility.

On the other hand, a partial attenuation of the dominant interferer components when they are respatialised may lead to reduced masking of the target without unduly affecting auditory continuity. Hence, there may be an optimal degree of interferer attenuation which provides some improvement in intelligibility.

### 7.4.2 Combining dominance and direction

So far, it has been shown that the relative amplitudes of the target and interfering sounds provide a good cue for determining which spectral components to respatialise. These components are processed to produce an auditory scene that is similar to directly spatialising the original mono sounds. However, this requires knowledge of the individual sources, which is impractical in a real world system where they reach the listener already mixed. It has also been shown that the dominant direction within a time window can be determined using a cross-correlogram. However, this does not provide discrete direction information for each spectral component.

Were it possible to determine a directional mask from the mixed signal, the respatialisation processing based on dominance could be applied to the dominant interferer frequency components. The generation of this mask from an analysis of the binaural input signal has proven to be extremely difficult. For example, by combining the dominance and direction detection methods, a mask of "dominant directions" could potentially be produced. This would show the spectral components that are correctly identified as being allocated to the target or the interferer and whether they are dominant or subdominant. The mask for two sound sources, a target spatialised to -40° and an interferer to 20° is shown in Figure 7-16.

Figure 7-16**(a)** shows the left target dominant mask and **(e)** is the left interferer dominant mask, so (a) is the inverse of (e). Similarly, for the right channel, (c) is the inverse of (g). The masks in the second column are black when the dominant signal matches the direction that was detected using the cross-correlogram method for that frequency component. Therefore, (b) shows the spectral components where the direction was detected as the target and the target is dominant. It can be concluded from Figure 7-16 that very few spectral components are correctly identified as being both dominant and from the correct direction. Therefore, either a large number of spectral components have the wrong direction detected for them, or they are not dominant for that direction. An alternative way of considering this is that only the components shown in black in column 2 of Figure 7-16 will be correctly processed by the respatialisation algorithm. Informal listening tests have highlighted the poor performance of the direction detection method for determining the masks for the respatialisation algorithm. It was found that even small amplitude target signals can disrupt the direction detection of the interferer components. The tests also indicated that a significant portion of the interferer sound remains localised in its original direction.

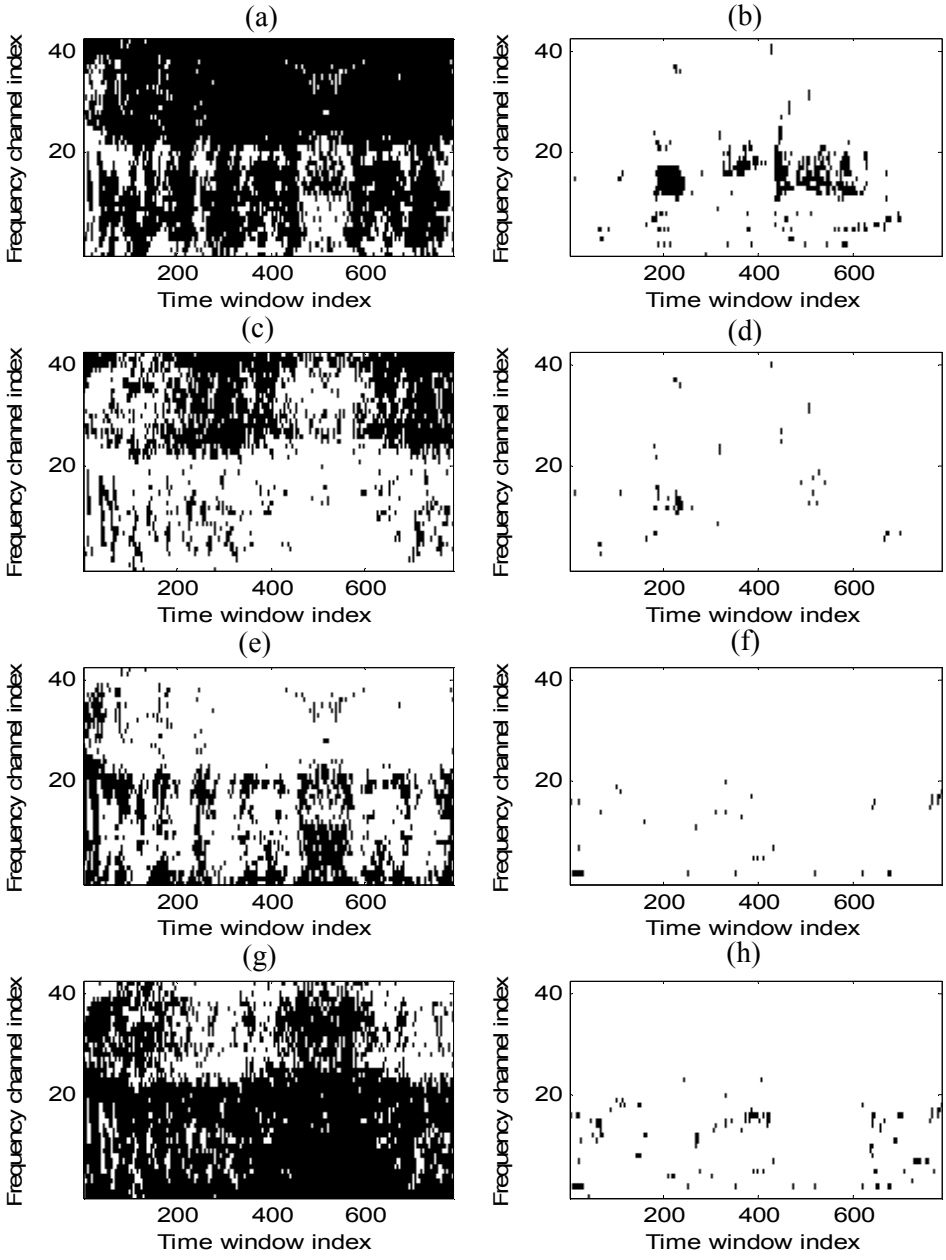



**Figure 7-16: The binary mask created when direction and dominance are combined. In (a), (c), (e) and (g), black regions signify a mask value of 0 and white regions a value of 1. Mask (a) is the left target dominant mask, (c) is the right target dominant mask, (e) is the left interferer dominant mask, (g) is the right interferer dominant mask. The black regions in (b), (d), (f) and (h) in the right-hand column show where dominance and direction match for each signal.**

Its poor ability to respatialise a single interferer sound when combined with a target sound suggests that this approach would completely fail when processing more complex auditory scenes, such as babble noise. This was confirmed to be the case, again using informal listening tests.

To meet the objective of applying the processing system to a low power digital hearing aid an alternative solution is required.

## *7.5 An auditory lens*

### 7.5.1 Introduction

Three binaural processing techniques have been discussed. Their essential characteristics, advantages and disadvantages are summarised in Table 7-2.

| **Processing method** | **Advantages** | **Disadvantages** |
|---|---|---|
| Direct | Using accurate HRTFs this produces the most numerically accurate binaural signals. | Requires prior knowledge of individual sound sources.<br>Requires accurate direction information for sound sources.<br>Requires accurate HRTF data. |
| Dominance | Perceptually authentic respatialisation | Requires prior knowledge of individual sound sources to determine dominance.<br>Requires accurate direction information.<br>Requires accurate HRTF data. |
| Cross-correlogram | Provides good estimate of a single source direction for each time window. | Does not provide direction information for every TFU. |

**Table 7-2: A summary of the advantages and disadvantages of the three techniques associated with respatialisation which have been discussed: direct, dominance and cross-correlogram.**

The main shortfall of the direct and dominance respatialisation techniques is the requirement for prior information relating to the direction and spectral content of the original sound sources or a means of estimating them from the composite binaural signal which is to be manipulated. *A priori* knowledge about the individual sound sources is impractical in a real listening scenario and in any case would remove the need for processing in the first place. Whilst the cross-correlogram provides a good estimate of the directions of the sound sources, it does not provide sufficient information for its application in respatialisation.

It would be highly beneficial if the requirement to have prior knowledge of the sound sources could be lifted. As discussed in Section 6.3, it is unnecessary to have accurately respatialised sounds to achieve a binaural advantage capable of raising the intelligibility of a target speech source. Therefore, attention was turned towards the development of a more generic system that did not rely on the calculation of precise directions for every spectral component. Furthermore, to simplify the processing and hopefully to reduce errors, a more direct respatialisation strategy was considered. That is, instead of despatialising and then respatialising each spectral component of an interferer sound, the algorithm should attempt to respatialise the spectral components to the new locations in one step. One-step respatialisation is shown diagrammatically in Figure 7-17. Ideally, the interferer sound sources are relocated without affecting the target sound source, enhancing binaural unmasking. Once in their new positions, the levels of the interferer sounds can be carefully reduced relative to the target. As was discussed in the case of the dominance method (see Section 7.4), the higher the level of the interferer sounds, the more they will mask the target. On the other hand, if the level of the interferer sounds is reduced too far, and depending on exactly how the movement of the interferers is achieved, spectral holes may appear in the target as a result of the interferer ceasing to induce auditory continuity in it. In both cases, intelligibility will be reduced and so there is assumed to be an optimal degree by which the interferers can be attenuated to yield the greatest improvement in intelligibility.

To assist the design of such a signal processing system some assumptions are made in order to reduce its complexity without having an excessive impact on its intended performance.

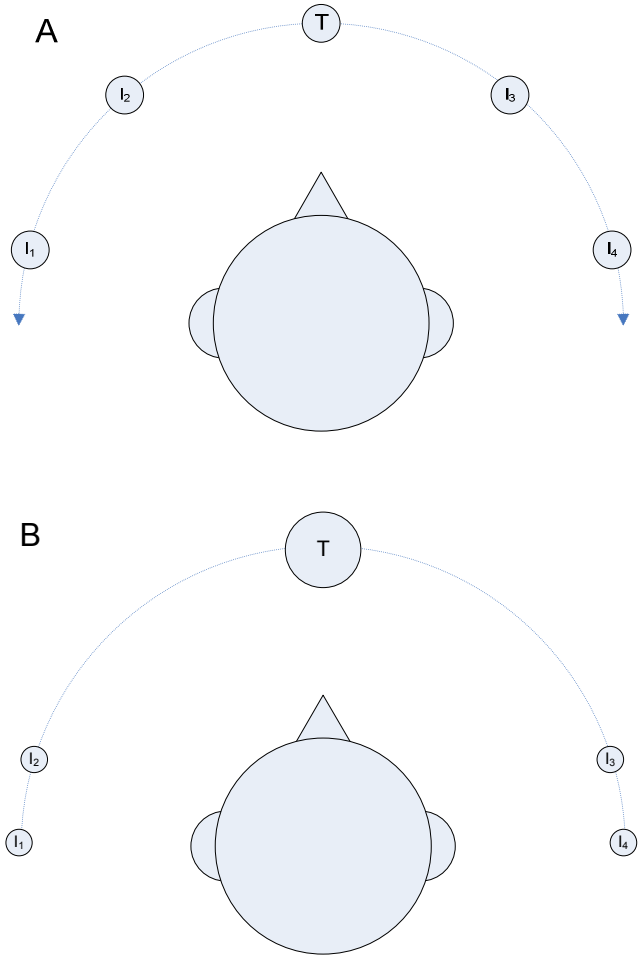



**Figure 7-17: (a) is an example distribution of target 'T' and interferer 'I' sound sources around a listener and (b) is the ideal result of the proposed sound source processing. The interferers are respatialised to new locations further away from the target. The use of smaller circles for the interferers signifies that their amplitudes have also been reduced relative to the target.**

The first assumption is based on the fact that the dominant cues for localisation in the horizontal plane are ITD and IID (Butler and Humanski, 1992). Furthermore, it is widely accepted that ITD is dominant for lower frequencies and IID for higher frequencies (Section 2.4.2). ITD and IID can be determined using the phase and amplitude differences between the corresponding TFUs in

the spectra for the left and right binaural signals. Respatialisation will focus on these properties by increasing ITD for frequencies below 1500 Hz and IID for frequencies above this. For simplicity, the fine spectral differences, principally the pinna cues, which have their strongest role in source localisation outside the horizontal plane, will not be considered. It is proposed, therefore, that simply by increasing the two binaural difference cues it will be possible to create for the listener the illusion of the interferer sound sources being relocated further away from the target sound and, in this way, raise the intelligibility of the target speech source.

Secondly, to distinguish between a target sound and an interfering sound a set of criteria will be defined. It will be assumed that the target sound is directly in front of the listener and is characterised by its ITD and IID both being close to zero. A tolerance is necessary, creating a target region rather than one specific direction, since in practice a target sound source will most likely not be at precisely 0° azimuth. All sounds outside this region will be treated as interferers and their non-zero ITD and IID will be amplified to increase their perceived angular separation from the target region. This is shown diagrammatically in Figure 7-18.

We term this form of binaural processing an auditory lens. It is analogous to a visual lens in that it has a focal region within which the sound objects are enhanced and a peripheral area where sound objects are increasingly pushed to the side and obscured, but not completely removed. It is noted that in its simplest form, as proposed, there is no distinction made between sounds that are in front or behind the listener. This is due to the imposed limitation of only considering the ITD and IID localisation cues.

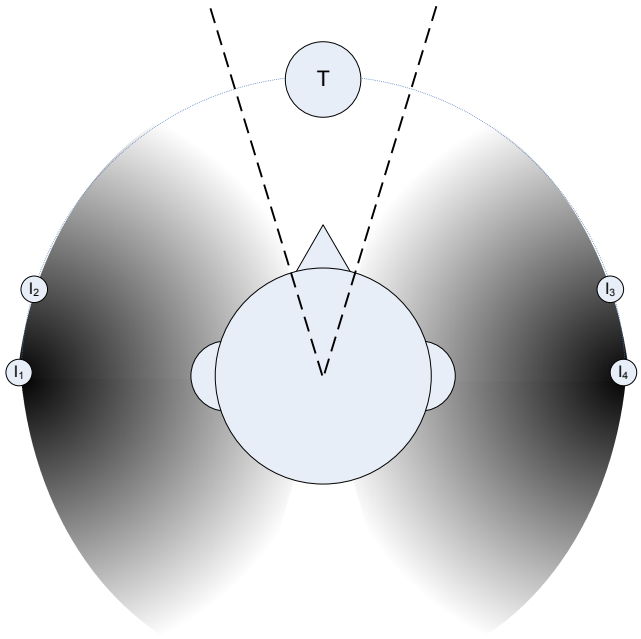



**Figure 7-18: The concept of an auditory lens applied to auditory scene processing. The aim of the processing is to move interfering sound sources from the lighter shaded areas to the darker areas without affecting the focal region marked with the dashed lines.**

It is proposed that the auditory lens will be adjustable in the following dimensions:

- The angular range of the focal region;
- The degree to which interferers can be attenuated;
- The amplification of the target.

This will allow a listener wearing a device that processes audio using an auditory lens to alter the settings depending on the listening environment they are in. For example, in a very noisy "cocktail party"-type scenario they may wish to have a narrow focal range and significant target amplification with aggressive attenuation of interferers. However, in more typical listening conditions the focal range and attenuation can be more relaxed.

Auditory lens processing has an important property. Interferer sound sources remain on the same side of the listener after they have been respatialised. This means that a listener's attention can be switched to an interferer sound source, if required, simply by the natural action of turning their head towards it.

The theoretical benefits of the auditory lens are encouraging. The practical implementation and validation of the signal processing algorithms are discussed next.

### 7.5.2 Auditory lens system architecture

This section describes the processing system architecture used for the auditory lens. Figure 7-19 illustrates the principal processing steps that are used. The spatialised signals are generated in an identical method to the dominance method as described in Eq. 7-17 to Eq. 7-22.

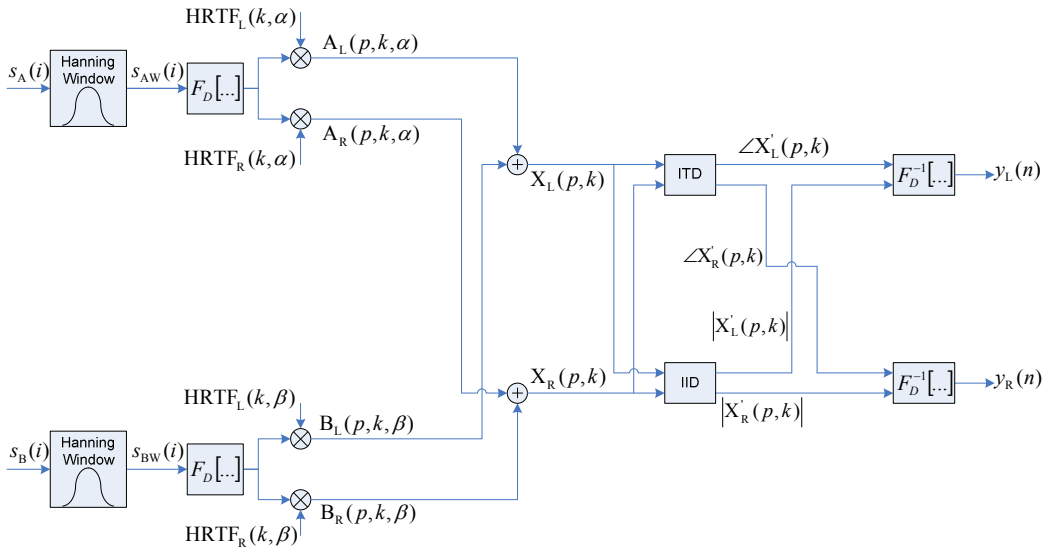



**Figure 7-19: Block diagram of the auditory lens system architecture**

The blocks labelled "ITD" and "IID" refer to the different types of augmentation used to enhance the spatial cues. These are covered in detail in Sections 7.5.3 and 7.5.5.

### 7.5.3 ITD augmentation

As discussed in Section 2.4.2, ITD is the dominant cue for localisation of a sound source and so increasing the ITD of a source will cause its perceived direction to become more lateralised. In Section 2.4.2.1 it is shown that the ITD can be approximately modelled by the Woodworth formula. This is repeated below:

$$\Delta T(\theta) \approx \frac{r(\theta + \sin\theta)}{c}$$ **Eq. 7-43**

where

$r$ is the radius of a 2-D circular head approximation;

$\theta$ is the azimuth of the source direction;

$c$ is the speed of sound in air.

The ITD for a pair of HRTFs from the CIPIC database (set 21, θ= -80° to +80°, $\phi$ = 0°) is compared in Figure 7-20 to the Woodworth approximation. The figure shows that the Woodworth formula provides a surprisingly good fit to the ITD encapsulated in a real HRTF as a function of azimuth. The ITD from the HRTF data is calculated using Eq. 7-17 which is derived from the phase angle, $\delta_{\mathrm{HRTF}}(k)$, between the left and right channels.

$$\delta_{\mathrm{HRTF}}(k) = \angle \mathrm{HRTF}_{\mathrm{L}}(k) - \angle \mathrm{HRTF}_{\mathrm{R}}(k)$$ **Eq. 7-44**

where $\angle$ denotes the phase angle of the complex HRTF data. Provided the frequency lies below a certain threshold, there is no ambiguity in the relationship between phase difference and ITD.

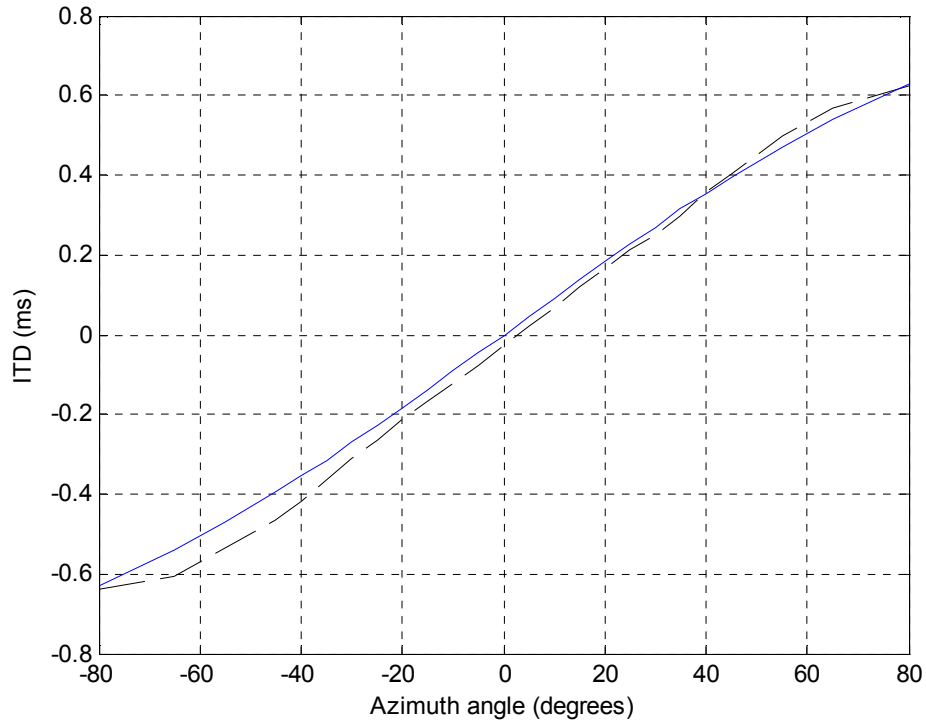



**Figure 7-20: Comparison of the Woodworth approximation for $r$ = 9 cm and $c$ = 340 m/s (blue solid line), with ITD computed for θ= -80° to +80°, $\phi$ = 0° from set 21 in the CIPIC database (black dashed line)**

For the purpose of moving an interferer sound, the following criteria for ITD processing are specified:

- Target sounds are defined as those that have an ITD that places them on or within a few degrees azimuth of the frontal median plane according to the Woodworth formula (for sound sources at zero elevation this translates to being in front of, or behind, the listener).
- The perceived location of the target sound should remain untouched. Therefore, ITDs determined to lie within the focal range will remain unprocessed.
- The perceived location of an interferer sound should be altered such that it is moved further to the side of the listener. Therefore, ITDs outside the focal range will be increased. Sounds that are closer to the target sound will receive a greater increase in ITD than sounds that are already farther away.

- The relative positions of sounds will be maintained. That is, the order of sounds around a listener will remain intact to limit disruption of the auditory scene.

The auditory lens does not distinguish between dominant and non-dominant frequency components. It simply has a focal region where the ITD for each TFU remains unprocessed and all other ITDs are increased.

The processing architecture for the ITD manipulation is given in Figure 7-21.

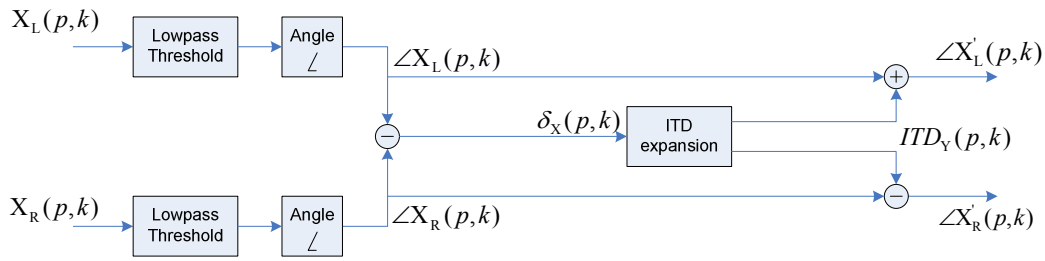



**Figure 7-21 Functional architecture of the lens processing for expanding the ITD**

A binaural signal mixture, $x_L(i)$ and $x_R(i)$, is windowed and transformed to the frequency domain where the resulting signals, $X_L(p, k)$ and $X_R(p, k)$ are lowpass-filtered. The phase difference in radians, $\delta_X(p,k)$, obtained from each left/right TFU pair below the frequency threshold, $k_{LP}$, of the filter is converted into an ITD, as shown in Eq. 7-45.

$$\mathrm{ITD}_X(p,k) = \left(\delta_X(p,k)\frac{N}{2\pi k}T_S\right) \qquad 0 \le k < k_{LP} < N-1 \qquad \textbf{Eq. 7-45}$$

where $T_S$ is the sampling interval and $N$ is the number of points in the Fourier transform of the windowed input signal. The lens processing increases the original phase difference $\delta_X(k)$ to the augmented phase difference $\delta'_X(k)$, or equivalently the ITD, as described in Eq. 7-46.

$$\mathrm{ITD}_Y = \mathrm{ITD}_{PK} - \left[\left(\frac{\mathrm{ITD}_{PK}}{\mathrm{ITD}_{PR}^{\,2}}\right)\cdot(\mathrm{ITD}_X - \mathrm{ITD}_{FR})^2\right] \qquad \textbf{Eq. 7-46}$$

(Note: time and frequency variables have been omitted for clarity.)

where: $\mathrm{ITD}_x$ is the calculated ITD of the binaural input signal,

$\mathrm{ITD}_y$ is the ITD of the processed signal,

$ITD_{PK}$ is the peak ITD offset,

$ITD_{PR}$ is the ITD processing range,

$ITD_{FR}$ is the ITD focal range.

The augmented ITD is then converted back into a phase angle using Eq. 7-47.

$$\delta_X'(p,k) = \frac{\mathrm{ITD}_Y(p,k)2\pi k}{NT_S} \qquad \textbf{Eq. 7-47}$$

The new phase angle is applied to the original angle for each corresponding TFU, as given in Eq. 7-48 and Eq. 7-49.

$$\angle \mathrm{X}_L'(p,k) = \angle \mathrm{X}_L(p,k) + \frac{\delta_X'(p,k)}{2} \qquad \textbf{Eq. 7-48}$$

$$\angle \mathrm{X}_R'(p,k) = \angle \mathrm{X}_R(p,k) - \frac{\delta_X'(p,k)}{2} \qquad \textbf{Eq. 7-49}$$

The behaviour of the mapping from $ITD_x$ to $ITD_y$ is illustrated in Figure 7-22 and the terms are defined graphically in Figure 7-23.

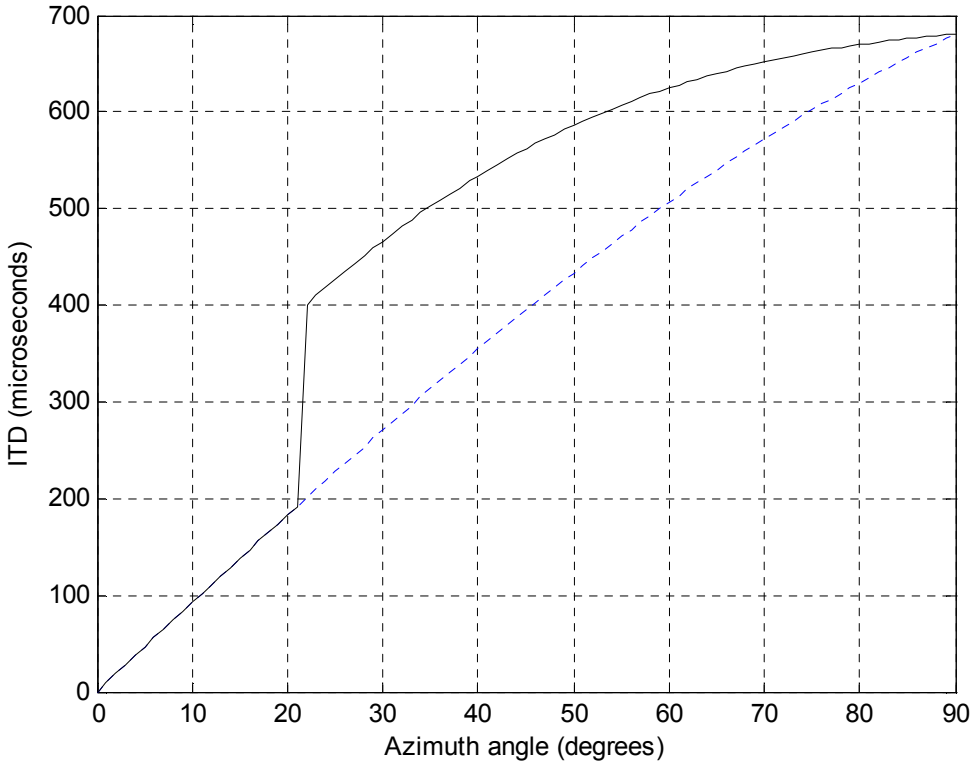



**Figure 7-22: The Woodworth approximation of ITD (dashed blue line), and the processed ITD (solid black line) for the auditory lens.**

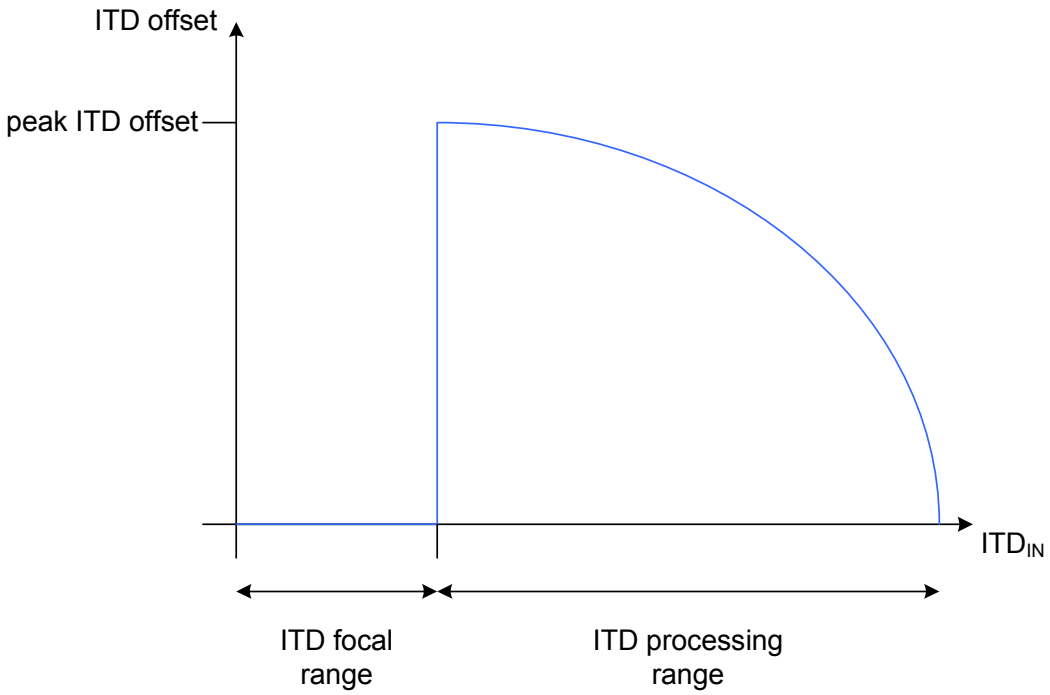



**Figure 7-23: Plot of the ITD mapping for increasing the angular separation between sound sources.**

The $ITD_{FR}$ is the region where no ITD processing is applied. Any sound source within this range will remain at the same perceived location. The $ITD_{PR}$ is the region of ITDs where the ITD is increased. This has the effect of remapping the sound to a location further to the side of the listener from its original location. $ITD_{PK}$ is the maximum ITD increase that is applied. Figure 7-23, shows how the ITD offset drops to zero as the ITD approaches its maximum processed value. Increasing the value for $ITD_{PK}$ creates a more aggressive ITD mapping, as shown in Figure 7-24 (b).

The parameters of the function above can be adjusted to produce various processing configurations, selections of which are shown in Figure 7-24 (a) to (c).

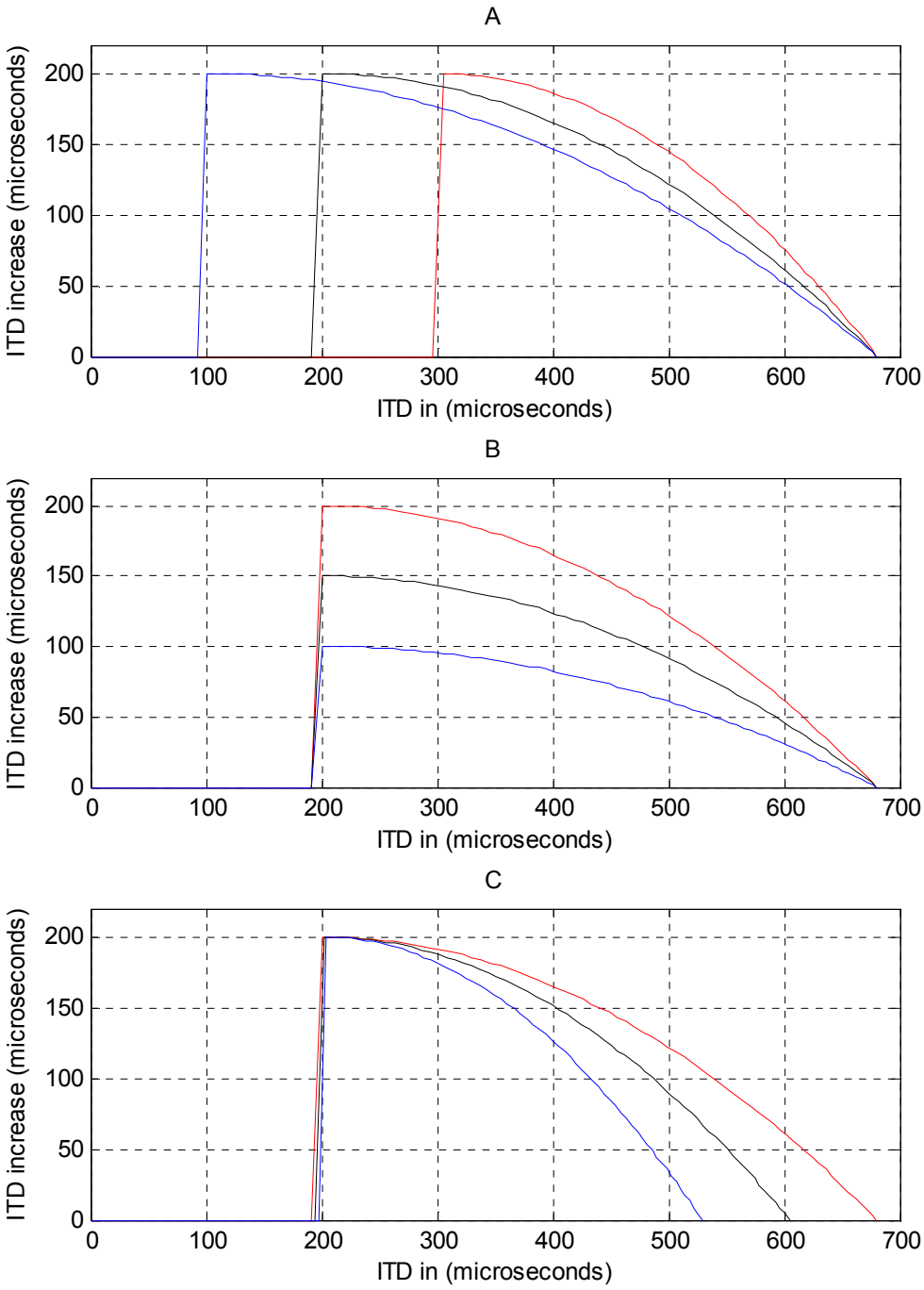



**Figure 7-24: Different tunings for the ITD augmentation function. (a) zero limit tunings for 100 $\mu$s (blue), 200 $\mu$s (black) and 300 $\mu$s (red) ; (b) peak ITD offset tunings for 100 $\mu$s (blue), 150 $\mu$s (black) and 200 $\mu$s (red) ; (c) ITD range tunings based on different head radii for the Woodworth formula of 70 mm (blue), 80 mm (black) and 90 mm (red).**

The phase and magnitude are then recombined to produce the complex FFT data which is finally inverse transformed to reconstruct the modified output waveform.

This section has introduced a means for applying ITD augmentation to a binaural signal using a flexible mathematical mapping function. The validation of the ITD processing is now discussed, where it is shown that it provides an effective illusion of a sound source being perceived at a more lateralised location. This component of the auditory lens provides a generic method for simulating an increase in angular separation between interferer and target sounds.

### 7.5.4 Validation of ITD augmentation

The ITD processing discussed in Section 7.5.3 was validated to ensure that the ITD adjustments applied to mixtures of signals are as expected. Firstly, a single sound was spatialised and the data passed through the lens respatialisation process. The sound sampled was a male talker speaking the word “crowded”. The mono sound was spatialised to 10° azimuth using the direct method described in Figure 7-1. Figure 7-25(a) shows the time lag between the left and right channels of the sound. Figure 7-25 (b) shows the effect of the auditory lens processing on this signal graphically. The auditory lens has increased the time lag between the left and right signals whilst substantially maintaining the shape of the waveform. The alteration in the spectral levels is due to the IID processing, which is discussed in Section 7.5.5. Perceptually, the sound appears to move further to the side of the listener.

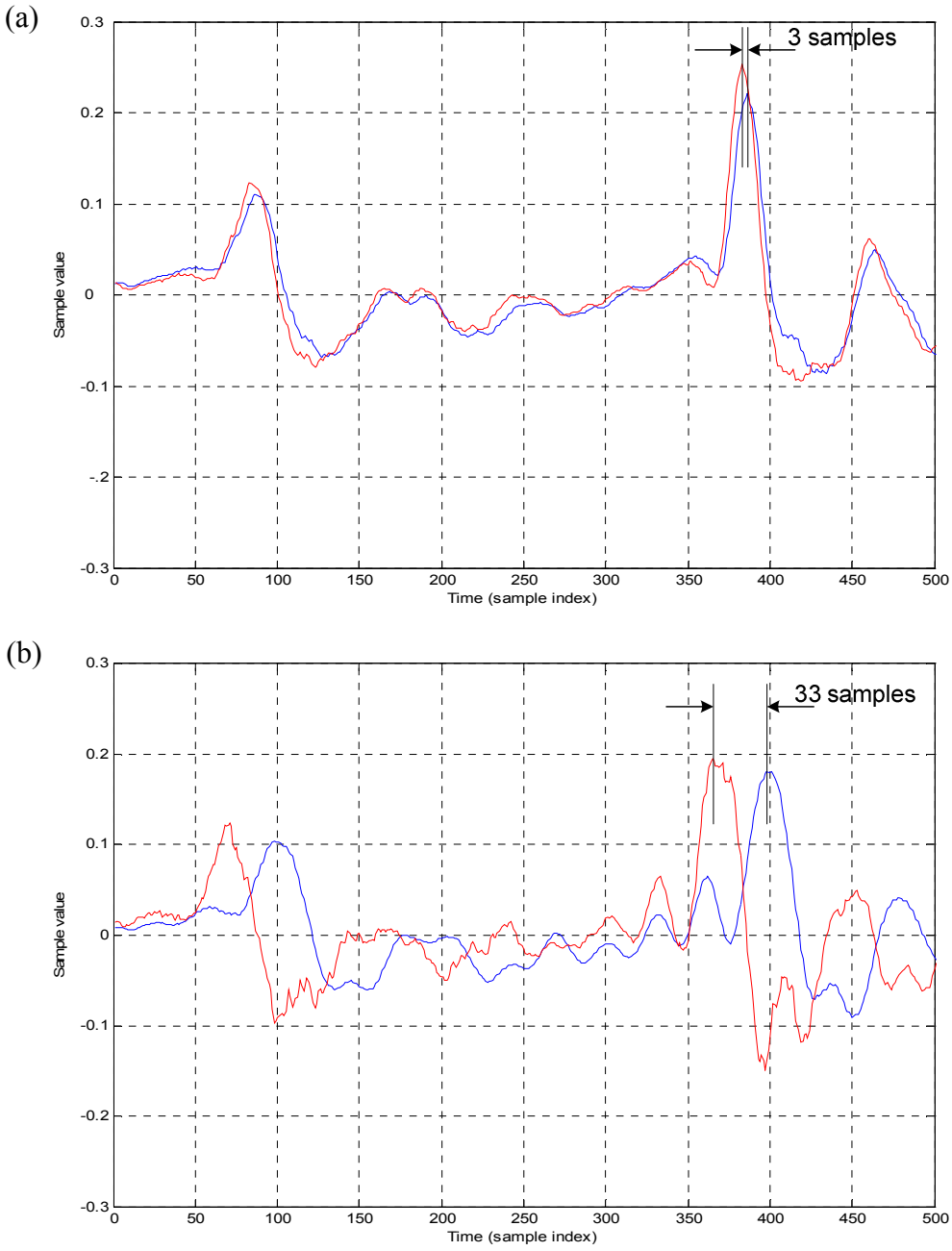



**Figure 7-25: Time domain plots of left (blue) and right (red) binaural signals. (a) is the signal for a mono sound source spatialised to 10° azimuth, with an ITD of 3 samples (68 µs). (b) shows the signal after processing with the auditory lens. The ITD has been extended to 33 samples (748 µs).**

To confirm that spectral components that fall within the focal range are left unprocessed, the lens was tested using a sound spatialised to 0° azimuth. The output signals had no additional phase shift applied and as a result the perceived spatial location of the sound remained unchanged.

### 7.5.5 IID augmentation

The general format of IID augmentation is taken directly from the methods already discussed in relation to ITD augmentation. That is, IIDs determined to lie within the zero IID limit are left untouched and all IIDs outside this limit are increased. The processing architecture for the IID manipulation is shown in Figure 7-26.

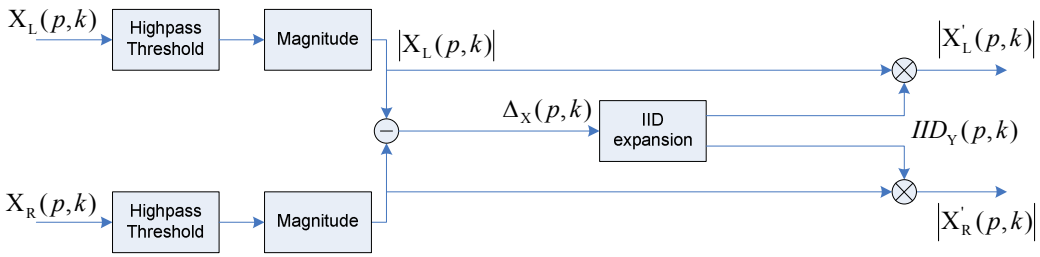



**Figure 7-26 Functional architecture of the lens processing for expanding the IID**

The IID increases with frequency (see Section 2.4.2.3) and so the IID augmentation induced by the lens must reflect this. The equation used to perform the IID augmentation is given in Eq. 7-50.

$$IID_Y = IID_{PK} - \left[\left(\frac{IID_{PK}}{IID_{PR}^{\ 2}}\right).(IID_X - IID_{FR})^2\right] \qquad \textbf{Eq. 7-50}$$

(Note: as for ITD (Eq. 7-46), the time and frequency variables have been omitted for clarity.)

where:

$IID_X$ is the calculated IID of the binaural input signal,

$IID_Y$ is the IID of the processed signal,

$IID_{PK}$ is the peak IID offset,

$IID_{PR}$ is the IID processing range,

$IID_{FR}$ is the IID focal range.

The parameters of the IID remapping function are defined graphically in Figure 7-27.

The new IID is then used to alter the original magnitude for each corresponding TFU, as given in Eq. 7-51 and Eq. 7-52.

$$\left|X_L^{'}(p,k)\right| = \left|X_L(p,k)\right| \cdot \frac{IID_Y(p,k)}{2} \qquad \textbf{Eq. 7-51}$$

$$\left|X_R^{'}(p,k)\right| = \left|X_R(p,k)\right| \cdot \frac{2}{IID_Y(p,k)} \qquad \textbf{Eq. 7-52}$$

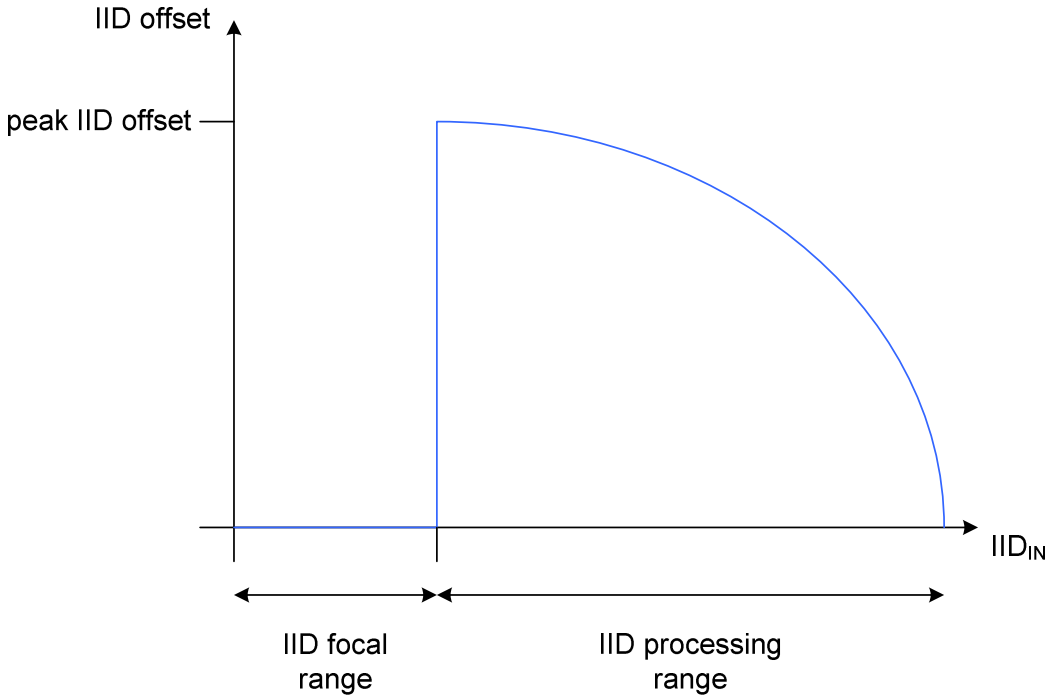



**Figure 7-27: Plot showing the parameter definitions for the IID remapping part of the auditory lens.**

Figure 7-28 shows a remapping function for a typical augmentation of IID. This uses parameter values of 5 dB to 25 dB for *peak IID offset*, 4dB for *IID focal range* and 20 dB to 45 dB for *IID processing range*.

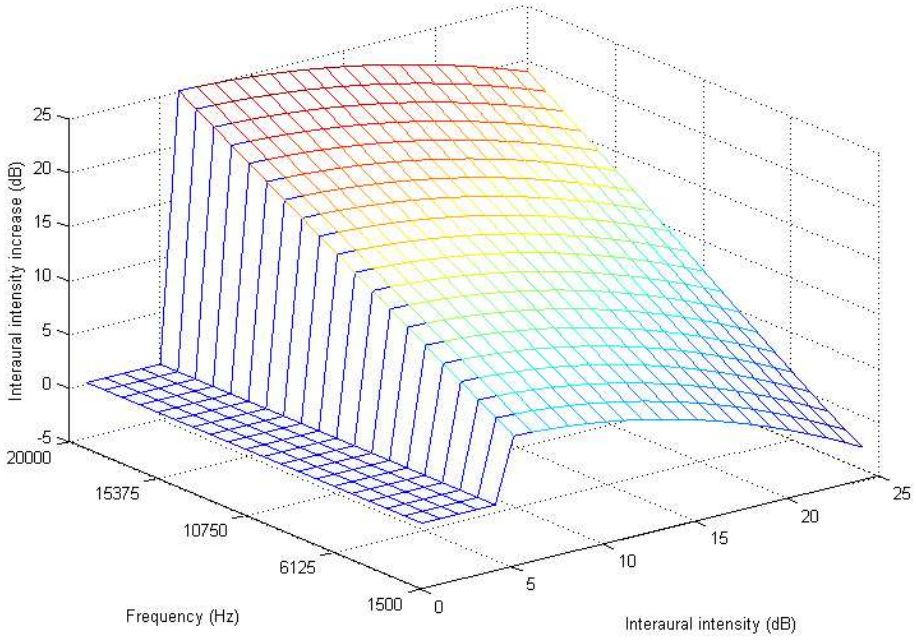



**Figure 7-28: A typical augmentation of IID.**

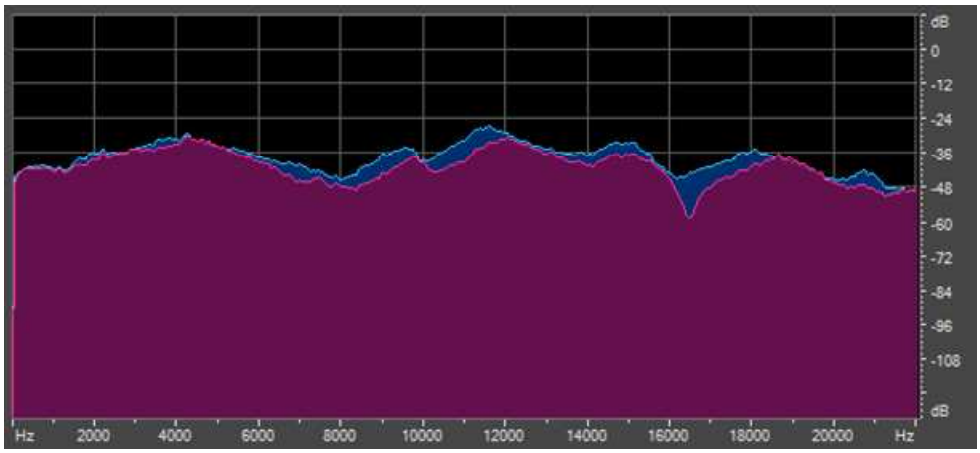
The frequency range covers the frequencies where IID is dominant. For the example shown in Figure 7-28 this is from 1500 Hz to 20 kHz. The IID augmentation for 1500 Hz and above is applied alongside the ITD augmentation below 1500 Hz, previously described.

### 7.5.6 Validation of IID augmentation

The IID processing to augment the intensity differences between spectral components of the left and right channels was evaluated perceptually using informal listening tests as well as numerically. The plots in Figure 7-29 show the IID processing that has been applied to a white noise signal, spatialised to -5° azimuth. Plot (A) shows the frequency content of the directly spatialised white noise signal, the left channel is shaded blue, the right channel shaded red with the purple regions indicating an overlap of the plots. The difference between the two channels is given by the blue region at the top of the plot. Plot (B) shows the frequency content after the lens processing. The blue region has increased in area, indicating a greater difference between the left and right channels. The

region below approximately 1500 Hz has been left untouched. There are some further spectral regions that have not been processed as the IID was small enough to be placed in the focal range.

(a)




(b)

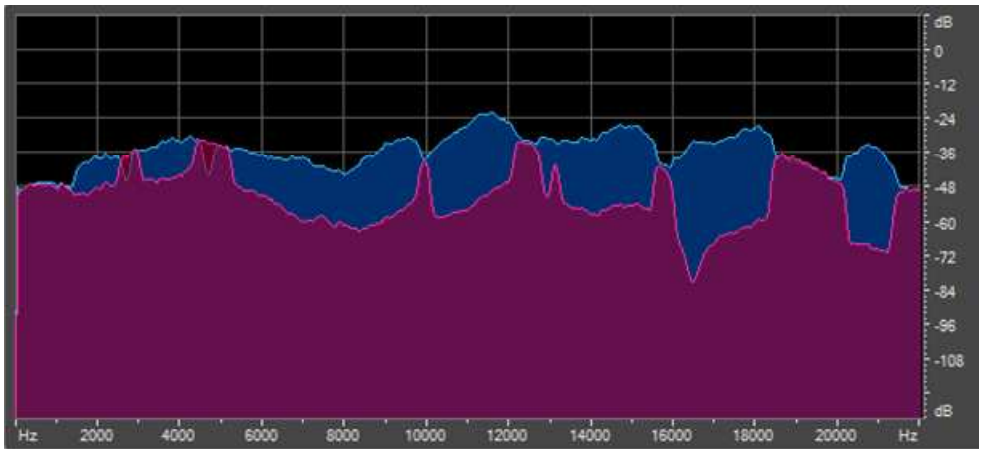
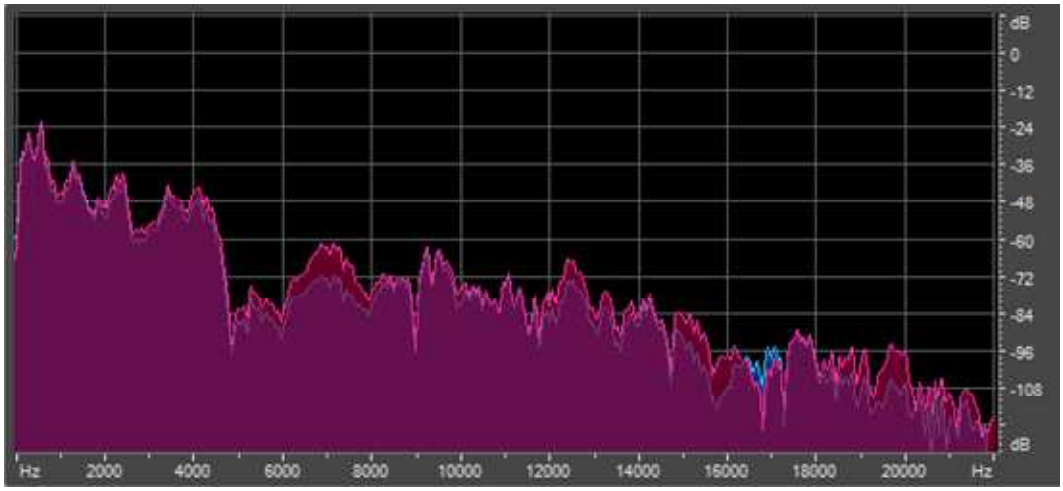



**Figure 7-29: Two spectral plots of a white noise signal highlighting the general increase in IID. (a) is directly spatialised to -5° azimuth and (b) processed to increase lateralisation using the auditory lens. The left signal of the binaural pair is shown in blue, the right in red.**

Figure 7-30 shows the intensity change applied to the left and right channels of a segment of speech (a) pre- and (b) post-lens processing. Initially, plot (a), the level difference of the left (blue) and right (red) spectral energy plots is relatively small. Plot (b) shows the result of passing the speech signal through the auditory lens. A general increase in the intensity level difference is observed, which gradually rises with frequency. Signal energy below 1500 Hz is unchanged.

(a)




(b)

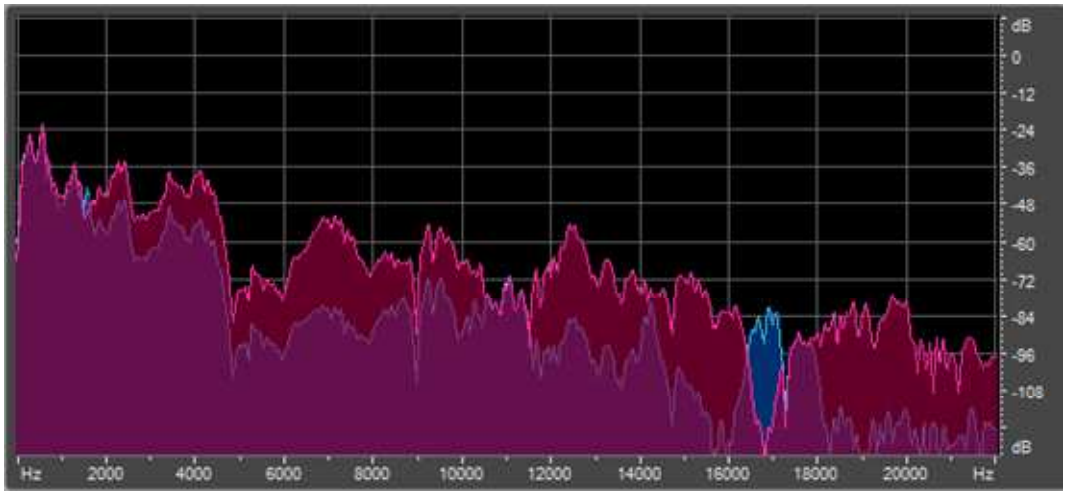



**Figure 7-30: Comparison of the intensity difference between the left and right channels of a segment of spatialised speech (a) pre-lens processing and (b) post-lens processing. The input signal (a) is speech spatialised using HRTFs to an azimuth of 10°.**

Informal listening tests highlighted that the processing of IID had only a minor effect on the perceived location of the interferer signal. The majority of the movement is caused by the ITD processing. However, the IID manipulation adds consistency to the sounds, and tends to keep the low and high frequency components spatially fused. The impact of the processing is affected by the type of interferer sound used. For example, where the interferer is a second male speech source, the majority of the energy resides in the lower frequency bands and so is largely processed by the ITD augmentation part of the system.

### 7.5.7 Amplitude manipulation

It was highlighted in Section 6.2 that some form of overall attenuation of the interferer sound components might benefit the intelligibility of the target sound. It is neither possible nor wise to attempt to reduce the interferers too much. For one thing, it is important to preserve the binaural scene for social and safety reasons. For example, if an interesting sound source becomes active to one side of the listener, he/she will not only want to hear it, but also will need to know which way to turn their head to make it become the new target. A more pragmatic reason not to attenuate the interferer components too far is because they also contain some of the target signal. Attenuating them will create spectral holes, which, if the attenuation is too great, will cause a loss of auditory continuity which will disrupt the target signal, reducing its intelligibility.

There exists an optimum amount of interferer attenuation that will sufficiently expose the target sound without disrupting auditory continuity of the target induced by the interferer. The optimum attenuation will depend on the interferer's spectral content and the respatialised location of its components. Informal listening tests using a variety of target speech sentences and speech-based noise interference resulted in good exposure of the target speech with a static attenuation of 6 dB. The target speech and speech-based noise interferer were pre-processed such that the amplitude of the speech was 17 dB lower than that of the speech-based noise. Due to the spectrally near-stationary nature of the noise, the fixed attenuation was effective and did not noticeably affect continuity, although artefacts in the target were apparent. Increasing the attenuation further produced stronger artefacts, due to spectral holes, which became more objectionable.

### 7.5.8 Summary

This chapter has presented a novel signal processing algorithm for enhancing the intelligibility of a target speech sound source in the presence of interfering

sounds. This is achieved through the manipulation of the spectral spatial cues for the interferer sounds.

Spatialisation of sound sources is most effectively achieved by filtering a monophonic signal using a pair of head-related transfer functions. Under highly restricted conditions, this process can be extended to allow one of a pair of binaurally spatialised sources to be relocated. However, to do so requires prior knowledge of the spectral content of the individual sounds in the auditory scene and their spatial location. In practice this is not information that is available to a listener and, because of this, methods of estimating the direction of each sound source were considered. A cross-correlogram technique was demonstrated to provide a good estimate of the dominant direction across all frequencies within a single temporal frame. However, the respatialisation processing requires an accurate direction for each individual TFU and this information was not available using this method.

It was shown that the problem could be simplified by processing only the spectral components for which the interferer sound was dominant. Combining HRTF spatialisation with this dominance method, again with prior knowledge of the original signals, it is possible to respatialise only the interferer signal with results perceptually very close to those obtained by spatialising the interferers directly to a more lateral position.

Estimation of the dominant components in a binaural mix of sources has proven to be very difficult and is compounded by the difficulties of estimating source directions. Together, these problems render an HRTF-based respatialisation method ineffective due to the large number of estimation errors in the binary masks used for indicating source dominance.

It is clear that a signal processing scheme that requires neither *a priori* information nor sophisticated information to be extracted from the binaural signal, would be beneficial if it could attain a similar level of binaural unmasking. One such candidate is the proposed auditory lens. Based on the premise that to create the illusion of moving a sound source in azimuth it is

sufficient to process just the ITD and IID, the lens offers a simple and effective solution. The ITD is determined at low frequencies by comparing the phase information for the left and right binaural signals across all implicated TFUs. Beyond a small guard band for a narrow region of locations in front of the listener, where the target sound is assumed to lie, any temporal differences are amplified to increase ITD. At higher frequencies, the IID is determined by comparing the spectral energy information between the appropriate left and right pairs of TFUs. As for the ITD, any differences in IID are amplified in a way which approximates the effect of human head shadow.

Informal listening tests reveal that the auditory lens is capable of creating a powerful illusion of respatialisation. As such, its ability to provide binaural unmasking and an intelligibility improvement for a target speech source heavily obscured by an interferer was felt to be worth investigating. To determine its effectiveness on a wider audience it was essential to evaluate it through a tightly controlled perceptual listening experiment. The design and execution of the experiment is the subject of the next chapter.

# Chapter 8 – Psychoacoustic evaluation of the auditory lens

## *8.1 Introduction*

The literature review, Chapters 4 to 6, has highlighted the factors that influence the intelligibility of a target sound when presented simultaneously with multiple interfering sounds. The angular separation between sound sources has a direct impact on the intelligibility of a target sound. A method for improving the intelligibility of a target speech sound source, based on manipulating angular separation, was introduced in Chapter 7. The algorithm warps the binaural soundfield by remapping ITD and IID to increase the angular separation between the target and interferer sound sources. One of the essential features of the algorithm is its low computational load, allowing it to be used in low power portable devices, such as hearing aids.

Up to this point, the validation of the algorithm has been performed either by modelling and simulation, or through informal listening. It was shown in Section 7.5.4 that auditory lens processing is capable of inducing a perceived respatialisation of a sound source, increasing angular separation from a target sound source. Furthermore, interferer spectral components can be attenuated to increase SNR. Intelligibility can be estimated algorithmically, for example, using automatic speech recognition, but the models used are based on the type of signal processing applied. The auditory lens is a novel approach and as such, existing intelligibility models may not be compatible. While detailed analysis of the signals provides an indication of the expected performance of the algorithm, perceptual testing of the algorithm is seen as the only way to evaluate the system rigorously. This will ensure there are no design limitations due to testing with a small, unrepresentative set of individuals. This section discusses the design and execution of a listening experiment to assess the ability of the algorithm to improve intelligibility for a typical listener with normal hearing.

## 8.2 Experiment design

The aim of the listening experiment is to determine whether the auditory lens processing can successfully improve the intelligibility of a target sound in a mixture of interfering sounds. It is noted that the typical listener is capable of understanding speech in very challenging listening environments. Therefore, to evaluate the intelligibility improvement induced by the algorithm it is necessary to start from an extreme baseline listening condition where intelligibility is very low. This was achieved using a simulated cocktail party effect with unilateral multi-talker, co-located speech interference. This has spectral and temporal properties of speech providing a harsh interferer when spatialised to an azimuth close to the target speech sound. Bilateral and spatially diffuse babble interferers are subjects of planned future work based on the findings of the current listening configuration.

The emphasis of the listening test is to quantify intelligibility and not audio quality. It is expected, based on informal listening, that the auditory lens will not impact audio quality to a point where it becomes annoying or distracting to the listener.

The experiment is aimed at listeners with no known or only minor hearing loss. Ultimately, the auditory lens is intended for people who would benefit from wearing bilateral hearing aids and who therefore have some level of binaural hearing ability. Bilateral hearing aids are becoming more popular, even for listeners with monaural hearing loss (Köbler *et al*., 2001). It is anticipated that the auditory lens will further improve the listening experience for users of bilateral hearing aids by adding another level of signal enhancement to that used to compensate for the monaural deficits in each ear.

### 8.2.1 Experiment protocol

The listening experiment was designed with a strategy of testing many subjects once only. The advantage to the listener is that the time they devote to participating in the experiment is kept to a minimum. This approach also produces experimental results which are a better reflection of algorithm performance for the general population. The listening experiments discussed in the literature review reveal the wide range of test durations which have been employed in the past. These are listed in Table 8-1.

| First author | Approximate duration of listening tests |
|---|---|
| Warren (1995) | 30 mins |
| Peissig (1997) | 4 hours |
| Freyman (1999) | 1 hour (a break half way through) |
| Freyman (2001) | 1 hour (a break half way through) |
| Hawley (1999) | 6.5 hours over multiple sessions. |
| Hawley (1999) | 6 hours over multiple sessions. |
| Best (2006) | 4 or 5 sessions. No more than 1 hour per day. |
| Warren (1976) | 3 sessions of 45 minutes on separate days |
| Bashford (1987) | 35 minutes per session (including training) |
| Bashford (1987) | 45 minutes per session (including training) |
| Bashford (1996) | 20 minutes per session |
| Warren (1997) | 30 minutes |
| Shinn-Cunningham (2008) | 4 sessions. One session per day |

**Table 8-1: Summary of listening test durations for experiments discussed in literature review (Chapters 4 to 6).**

Based on these figures and the practicalities of listener availability, it was decided that the experiment should aim for a total duration of approximately 30 minutes. The experiment requires sustained concentration and 30 minutes was felt to represent a reasonable limit before listener fatigue becomes an issue. Furthermore, it allows subjects to complete the experiment in a single visit without taking a break.

Each subject was given training to acclimatise them to the type of task they would be exposed to during the main experiment and to help them to focus on

the target speech. There was also the opportunity for the listener to adjust the overall amplitude of the audio to a comfortable level. The use of a training session meant that the bulk of the learning process would take place prior to the main experiment, leading to more stable and significant results.

The experiment was configured to suit the analysis of variance (ANOVA) method for determining the statistical significance of the data. ANOVA allows the means of more than two data sets to be compared. Each independent variable within the test is known as a *factor*. The *treatment* of a factor describes how it is altered during the test. For the listening test described in this section there are two factors, sentence predictability and spatial processing type. The predictability has two treatments, low and high. The spatial processing type has four treatments, 'spatialised 5', 'spatialised 80', 'dominance' and 'lens', which are described in detail in Section 8.2.6.

The subjects were required to listen to each audio segment, comprising a simple sentence, and type in the words they heard. Subjects were prevented from listening more than once to a particular presentation of a sentence, which ensured that they could not gain an advantage by replaying the sentence to improve their intelligibility score. Each subject was exposed to 100 sentences, given in Appendix C, and all the sentences presented to one listener were different to avoid any learning effects due to familiarity with the sentences (McNaughton *et al.*, (1994)). Each subject was presented with sentences which were all processed using just one of the four spatial conditions. This prevented any interaction between processing conditions. Furthermore it kept the experiment design simple and provided just one variable, the intelligibility of the sentence. Each subject heard the sentences in a different random order to minimise any interaction between the high and low predictability sentences.

### 8.2.2 Target sentence selection

The target sentences were selected from the speech perception in noise (SPIN) set of sentences (Kalikow, 1977). These are phonetically balanced and classified

as high and low predictive sentences, based on whether the final (key) word in the sentence could be predicted from the preceding words. “The boat sailed along the coast” and “Jane was interested in the stamp” are examples of high and low predictive sentences, respectively. The SPIN sets used are ‘Form 2.1’ and ‘Form 2.2’, providing 50 high predictive and 50 low predictive sentences. Meaningful sentences were chosen as opposed to random word lists or isolated words in order to create listening trials which sounded natural and realistic. This decision was also taken in the light of the findings by Bashford *et al.* (1996), who reported that the intelligibility of high predictive sentences was restored when a noise interrupter was used to induce auditory continuity, whereas intelligibility was not restored under similar conditions when using random word lists.

The average duration of each sentence is approximately 5 seconds. Informal testing suggested that a typical subject would be able to type their response in approximately 10 seconds. 100 sentences would therefore take approximately 1500 seconds or 25 minutes. The training phase was designed to take no longer than 5 minutes, giving a total experiment time of around 30 minutes, as required.

The sentences were recorded into a PC in an anechoic chamber using a Canford MBC-550 high quality omnidirectional microphone and a Focusrite Safire soundcard. Each file was stored in mono using 16 bit PCM samples at a sampling rate of 44.1 kHz.

### 8.2.3 Choice of Interferer

The evaluation of target speech intelligibility under challenging conditions requires an interferer sound source that is ideally temporally and spectrally stable. This reduces variation in performance due to temporal or spectral holes in the interferer that may or may not align with words in the target sentence. This would occur if, for example, a single different sentence to the target sentence were used as the interferer, whether spoken by the same talker or a different talker. Furthermore, an interferer that has a long term spectral average which is

speech shaped would provide acute masking of the target sentence, which is desirable in this context.

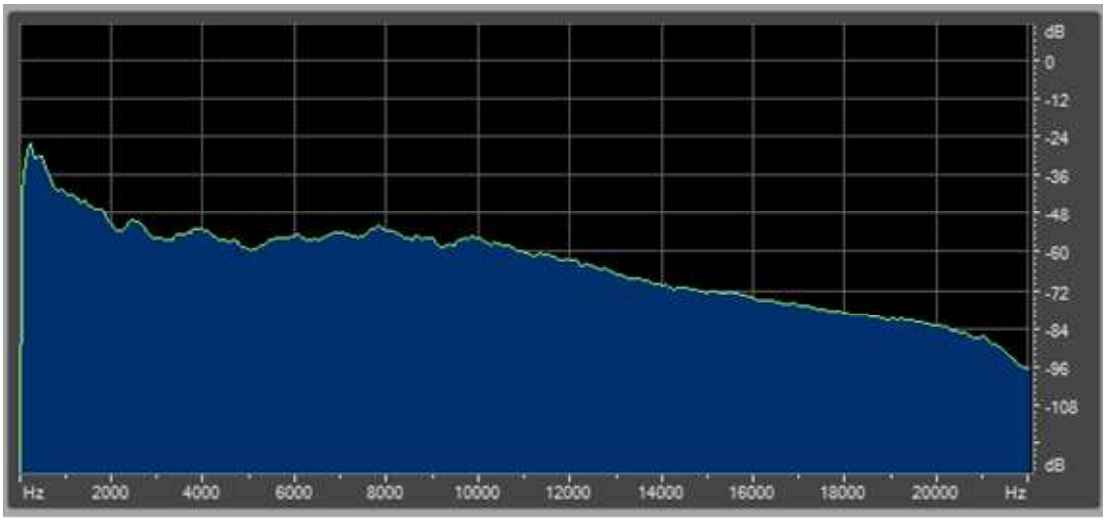



**Figure 8-1: The average spectral content of the mixed mono multi-talker interferer sound**

Because of its naturalness and improved spectral and temporal stability compared with a single talker, a multi-talker, speech-based noise was chosen as the interferer sound . This was composed of 6 segments of German, French and English passages read by both male and female talkers. The segments were time-reversed to avoid the possibility of distracting the listener with recognisable words from the interferers. That is, the speech-based noise acts as an energetic masker and not an informational masker. The audio files were mixed into a single mono sound, a typical spectral plot of which is shown in Figure 8-1. This was spatialised to –5° azimuth. To ensure the subject did not become familiar with the interferer sound, the fragment of speech-based noise used in each trial was randomly selected from within the 14-second file.

### 8.2.4 HRTF selection

The purpose of the listening tests is to measure the effectiveness of manipulating angular separation and relative signal strengths for improving intelligibility. These manipulations are satisfactorily achieved by modifying only the ITD and IID of binaural signals and this meant that generic HRTFs could be used in the

synthesis of some of the test sentences (Drullman and Bronkhorst, 2000; Shinn-Cunningham *et al*. (2001)). A discussion about which generic HRTF data sets were appropriate for the spatialisation of sound sources as input signals to the auditory lens was presented in Section 7.2. HRTF set 021 from the CIPIC database was selected for the development and informal evaluation of the auditory lens. This set was also used for the spatialisation of all of the sound sources for the listening tests. The large number of sound sources needing to be spatialised were created using fast convolution, outlined in Section 3.2.4.1.

Since highly accurate spatial perception is unimportant in the context of improving intelligibility it was considered unnecessary to equalise the transfer function for the headphones used in the experiment.

### 8.2.5 Listening environment

Several different rooms were used for the tests, which one depended on the location of the subject. It has been assumed that the room would not affect the intelligibility scores, provided that it was quiet. All the signals were presented to the subjects using high quality Beyer DT990 headphones. The signals were processed offline and stored as 44.1 kHz, 16bit PCM sample files. Matlab was used to implement the experiment and Matlab sound playback libraries were used from within its GUI. The digital signals were transferred to a high quality external soundcard, which was either an E-MU 0202 audio interface module or a Focusrite Saffire Pro, depending on the measurement kit used in a particular test.

The SNR of the test signals for each trial were deliberately set such that the target speech was difficult to hear due to the relatively high level of interferer sound. Therefore, a small amount of additional external noise from the listening environment or the playback equipment was considered acceptable and to have a negligible impact on the intelligibility of the target speech.

### 8.2.6 Test conditions

The majority of real-world speech sources lie in or close to the horizontal plane. Therefore, the conditions under test involve sound sources at zero elevation, i.e. at the level of the subject's interaural axis.

The spatial processing treatments for the listening test are defined in Table 8-2. The table allows a direct comparison to be made between the four test conditions.

| Target azimuth direction | Interferer azimuth direction | Identifier | Description |
|---|---|---|---|
| 0° | -5° | Spatialised 5 (S5) | The target and interferer speech sounds are individually spatialised and then mixed together |
| 0° | -80° | Spatialised 80 (S80) | The target and interferer speech sounds are individually spatialised and then mixed together |
| 0° | -5° | Dominance (D) | The S5 condition is used as the starting point. The "dominance" method (Section 7.4) is used to respatialise interferer spectral components to –80°. |
| 0° | -5° | Lens (L) | The S5 condition is used as the starting point. The "lens" method (Section 7.5) is used to respatialise the interferer spectral components further away from the target. |

**Table 8-2: The four listening conditions evaluated through the listening test.**

#### 8.2.6.1 Condition S5

This condition spatialises the target speech at 0° azimuth and multi-source speech-based noise at –5° azimuth. The direct spatialisation of the sounds represents the ideal binaural soundfield for this angular separation. It is the most challenging scenario in the experiment and is the baseline condition for the listening test. It is expected that listener intelligibility scores for this condition will be low. This condition also provides a starting point for the spatial processing algorithms.

#### 8.2.6.2 Condition S80

This condition spatialises the target speech at 0° azimuth and the interferer at -80° azimuth. The direct spatialisation of the sounds represents the ideal binaural soundfield for this angular separation. When used in conjunction with condition S5 it acts as the experiment gold standard for indicating the intelligibility improvement achievable due to relocation of the interferer sound under ideal conditions.

#### 8.2.6.3 Condition D

This condition is used to determine the performance of an ideal binaural mask based on *a priori* knowledge of target/interferer sound source dominance. It attempts to transform condition S5 approximately into condition S80. That is, the aim of the algorithm is to achieve an intelligibility performance as close as possible to the gold standard condition S80. The perceptual comparison of respatialising the TFUs where the interferer is dominant to the direct spatialisation method, condition S80, is discussed in Section 7.4. This condition also provides an indication of the intelligibility performance that could potentially be achieved with the binaural mask estimation methods discussed in Section 7.3.

#### 8.2.6.4 Condition L

This condition is used to determine the performance of the auditory lens that has been developed as an efficient and practical means of improving the intelligibility of a target speech sound when presented in a harsh acoustic environment. Like condition D, it attempts to transform condition S5 approximately into condition S80. That is, the aim of the algorithm is to achieve an intelligibility performance as close as possible to the gold standard condition S80. The lens does not require any prior knowledge of the sound sources or the spatialisation that has been applied to them. Furthermore, it is listener independent in terms of its spatial processing and can be applied to any binaural signals.

### 8.2.7 Intelligibility measurement

Two intelligibility scores were calculated using the number of correctly identified words for the high and low predictive target sentences. The 100 sentences used in the experiment contain a total of 625 words; 327 in the high predictive sentences and 298 in the low predictive sentences. The number of correctly recognised words typed in by a subject were counted using an automated word comparison routine. All words that were marked as incorrect were checked manually. Punctuation and spelling errors which nevertheless sounded the same as the test sentence when spoken were accepted as correct, e.g. "we're" and "we are" and also "heard" and "herd". The addition or deletion of suffixes was disallowed, e.g. "play", "plays" and "played". These rules were applied carefully across all the data.

### 8.2.8 Setting baseline signal levels

The signal levels of the target and interferer sounds were set so that it was difficult to understand the target speech in the S5 condition. The aim was for subjects to achieve an average correct word score for the S5 condition of approximately 20% across all sentences. Informal listening evaluation of the audio was used to set the initial levels. The interferer was set at a comfortable level and the level of the target speech sound was reduced until it became very difficult to understand for an experienced listener. The levels were checked using a pilot test carried out by 5 listeners. After each pilot test the correct-word scores were calculated. The target speech level was reduced for all sentences if a subject scored above 30%. Furthermore, the level of the target speech was reduced by a further 2 dB for any individual sentences where 100% of the words were correctly identified. This ensures that the other spatial processing conditions being tested have scope for improvement across every sentence. This was then validated in subsequent pilot tests to ensure the intelligibility level for the attenuated sentences was below 100%. The outcome of the pilot tests resulted in the target speech level for the majority of the high predictive sentences being set to -17 dB and for low predictive sentences to –16 dB relative

to the speech-based noise level. The remaining high and low predictive sentences were set to –19 dB and -18 dB, respectively.

### 8.2.9 Subject selection

The listening test was conducted using 60 volunteers. Three subjects reported having a minor hearing impairment, the remainder had no known impairment. The subjects were distributed equally between each of the four spatial processing conditions, giving 15 in each category. 48 males and 12 females participated, covering an age range of 19 to 62 years old.

## *8.3 Experiment procedure*

### 8.3.1 Listening environment

A quiet listening room was used for each run of the experiment. To assist with the capture of test results, three similar sets of equipment were used. Each set used the same audio files and only differed in terms of playback equipment and room used. The equipment used consisted of a portable computer with the Matlab interface running on it. The audio files were played digitally over USB or FireWire to an external high quality soundcard. Either a Focusrite Safire Pro 10, or an EMU 0202 audio interface was used. Beyer DT990 headphones were connected to the analogue outputs of the soundcard.

### 8.3.2 Graphical user interface for subject responses

A graphical user interface (GUI) was designed to record personal data for each subject and their responses for each trial of the test. Two interfaces were used,

one for the training phase and one for the main experiment, screenshots of which are given in Figure 8-2 and Figure 8-3, respectively.

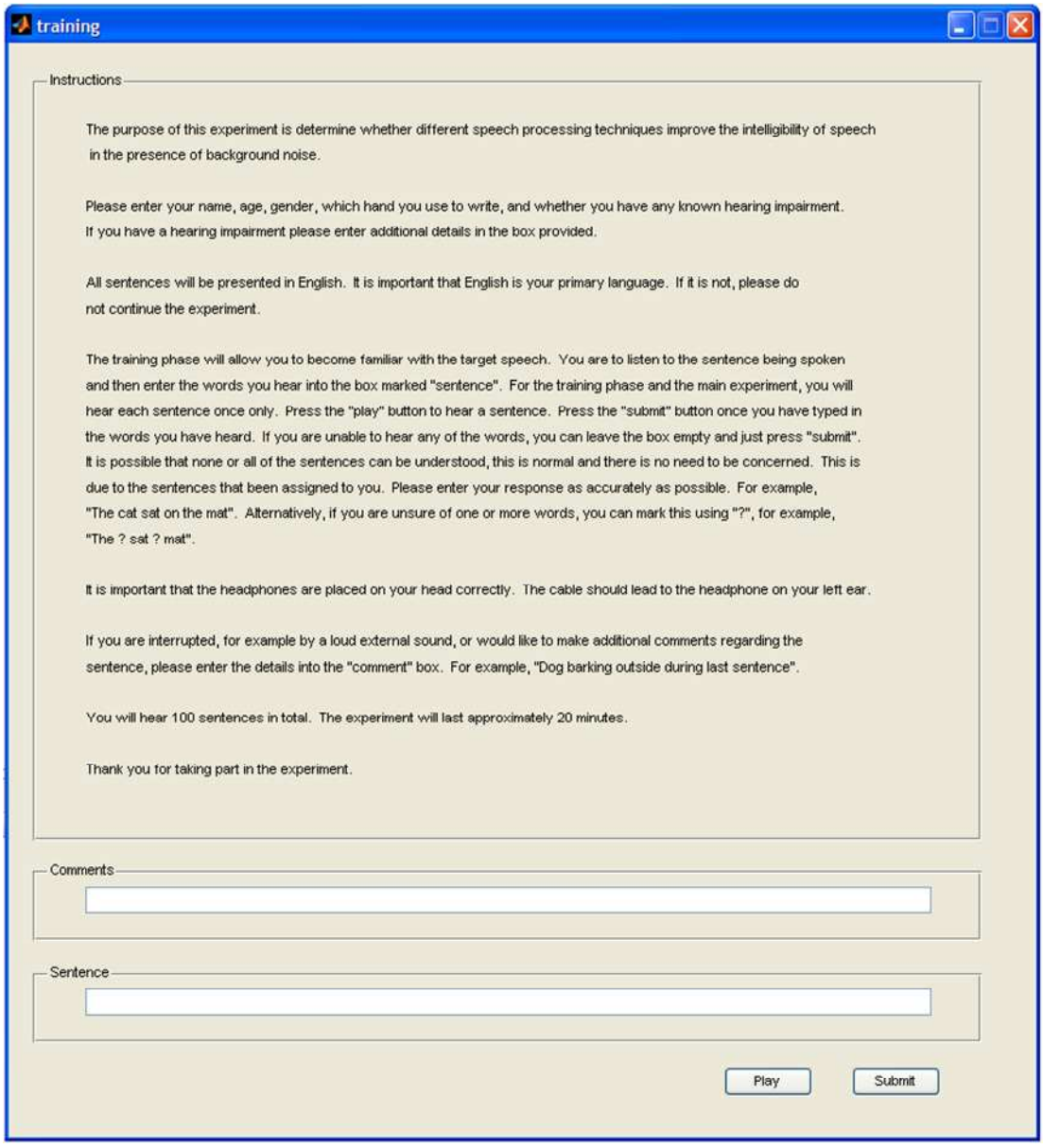



**Figure 8-2: The training GUI with instructions for the subject.**

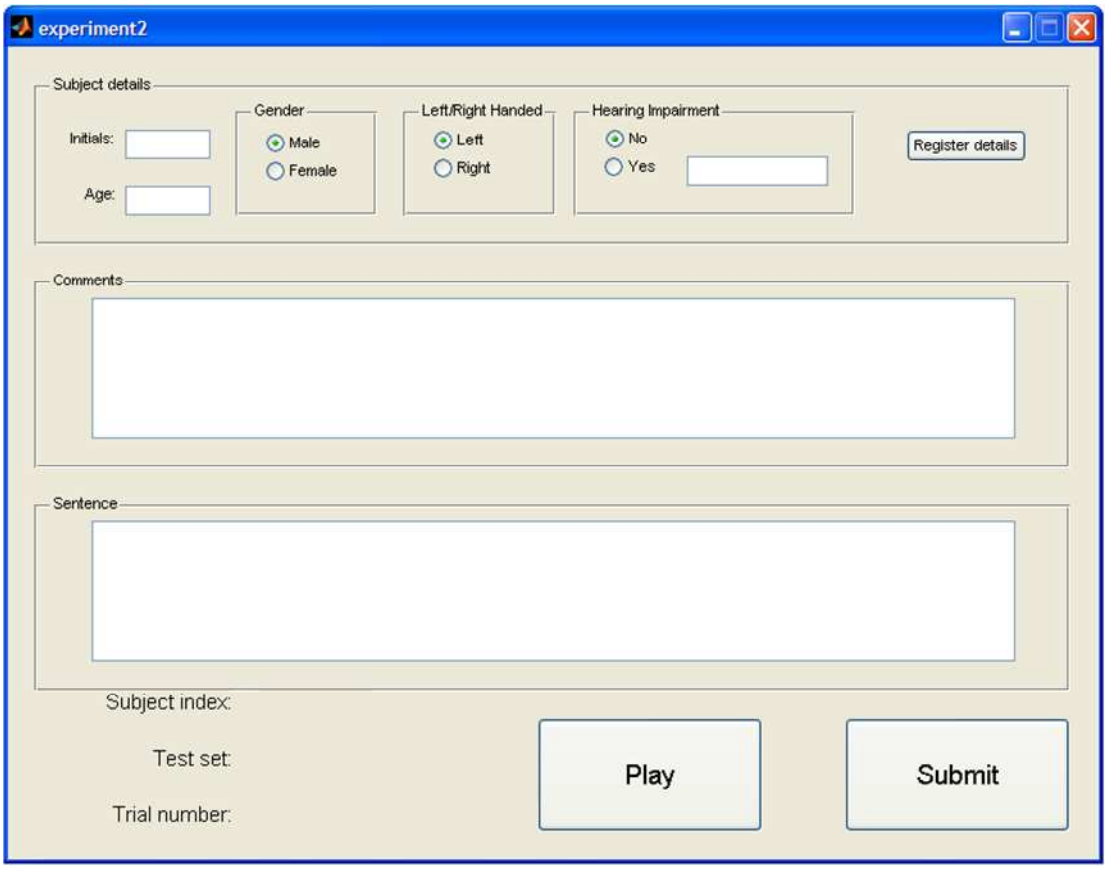



**Figure 8-3: The GUI used for capturing responses from subjects during the main experiment.**

The instructions given to the subjects are also provided as text in Appendix A. Subjects were warned that for some of the audio extracts the target speech would be extremely difficult to hear and that any failure to do so did not necessarily indicate they have a hearing impairment. To reduce anxiety, it was emphasised that the purpose of the experiment was to assess the signal processing algorithm and not their hearing.

### 8.3.3 Training

During the training phase, the subjects were presented with a variety of unprocessed and processed speech recordings. The subject pressed the 'play' button to hear a sentence. After each sentence the subject was tasked with typing the words they heard into the specified box in the GUI before clicking on the

'submit' button to proceed to the next sentence. Each sentence was played only once. No feedback was given as to how many words are correctly identified.

At one extreme in the training session, target speech was presented on its own without an interferer. This was used to give the subject practice in entering a complete sentence and so that they could become familiar with the user interface. At the other extreme, the interferer was presented without the target speech. In this situation the expected response was to leave the sentence entry box blank or type a single question mark in it. The remaining trials were characteristic of the main experiment and gave subjects practice in how to indicate one or more missing words. They also became familiar with the voice of the target speaker and the level of difficulty in understanding the target sentences that they would encounter. The subjects were allowed to adjust the volume of the signal to a comfortable level. The responses were informally checked prior to starting the main experiment. This was to ensure that the subject had no obvious hearing loss. If a severe hearing loss was detected or declared by the subject, the listening test would have been cancelled. This was not the case for any of the subjects tested.

### 8.3.4 Main experiment format

For the main part of the experiment, subjects were presented with 100 sentences in a unique random order. All the sentences presented to an individual subject were processed using only one of the four conditions listed in Table 8-2. In all other respects, the format of the main experiment was similar to that of the training session.

## *8.4 Results*

This section discusses the listening test results for the four conditions listed in Table 8-2. The correct-word scores for each subject are given in Appendix B. The results are analysed in terms of the impact of the spatial processing.

### 8.4.1 Spatial processing

The mean intelligibility scores for the high and low predictive sentences are plotted individually in Figure 8-4.

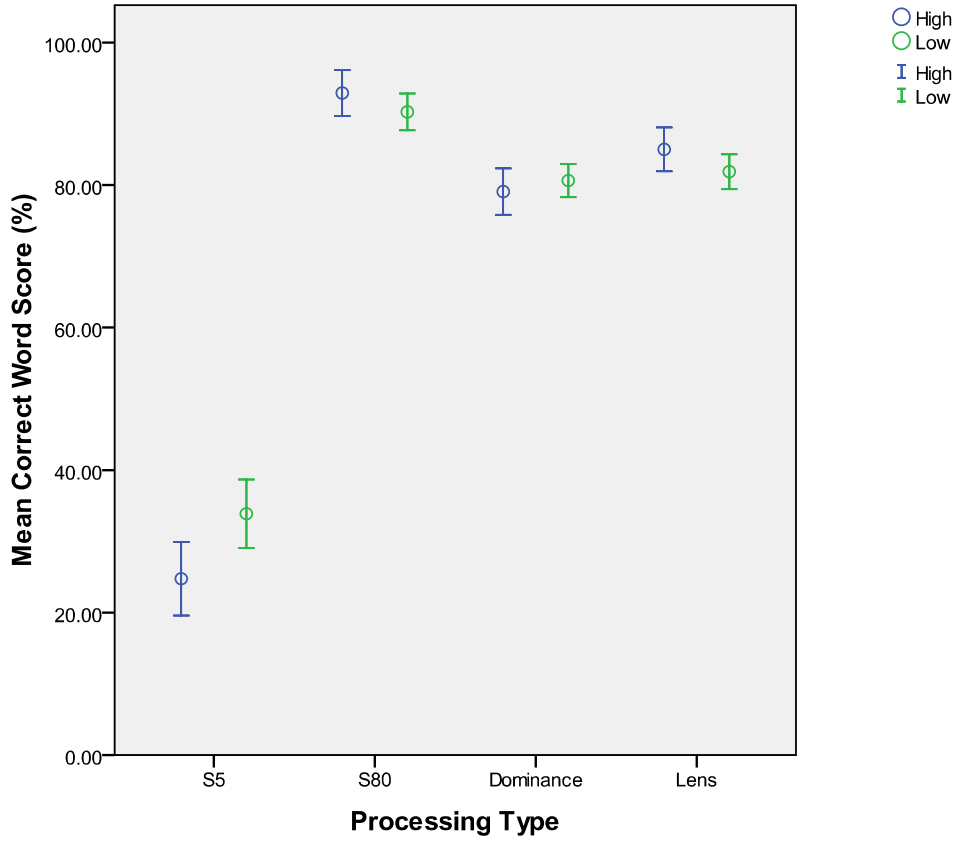



**Figure 8-4: The correct-word scores for each of the four configurations. Error bars are at the 95% confidence level.**

The ANOVA results for sentence predictability, are summarised in Table 8-3. The data show a significant main effect of predictability, $F(1,56) = 4.756$, MSE = 7.578, $p < 0.05$. There is also a significant interaction between the type of spatial processing and the predictability of the sentence, $F(3,56) = 32.265$, $p < 0.05$.

The results show there is an improvement in intelligibility when respatialising the interferer sound sources away from the target sound source.

| Source | Sum of Squares (SS) | Degrees of Freedom (df) | Mean square (MS) | *F*-ratio | Sig. |
|---|---|---|---|---|---|
| *Within subjects* | | | | | |
| Predictability | 36.042 | 1 | 36.042 | 4.756 | 0.033 |
| *Interaction* | | | | | |
| Predictability * Processing type | 733.518 | 3 | 244.506 | 32.265 | 0.0 |
| Error | 424.366 | 56 | 7.578 | | |

**Table 8-3: ANOVA summary for sentence predictability and processing type.**

### 8.4.2 Learning effect

The intelligibility scores are also considered as a function of sentence number in the experiment. The results are shown in Figure 8-5.

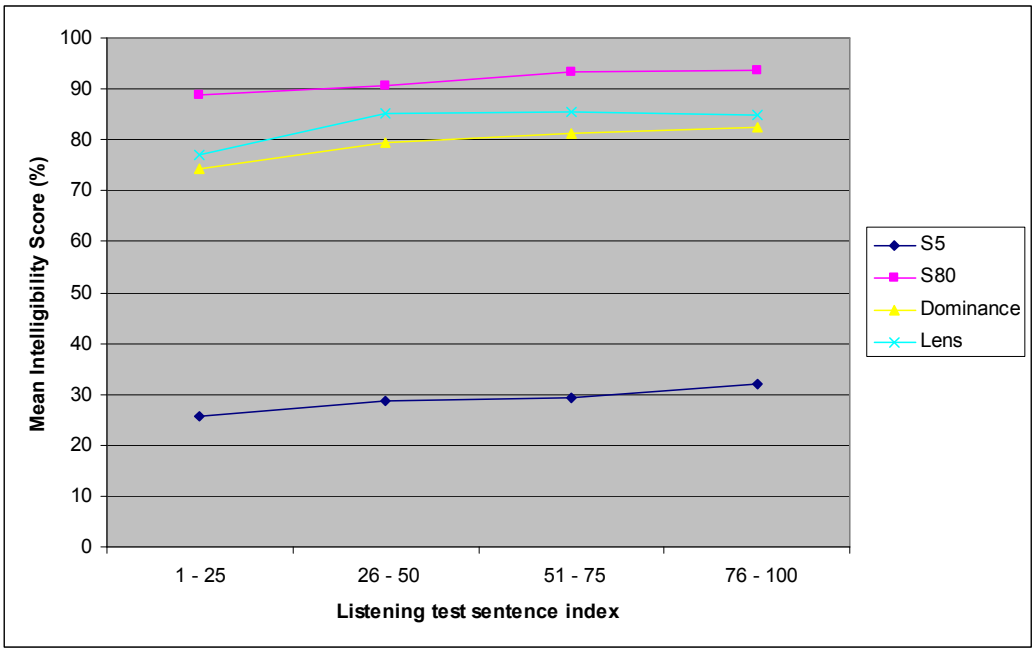



**Figure 8-5: The intelligibility scores for each quartile of the test sentences, showing a learning effect through the duration of the experiment.**

The ANOVA data for the effect on intelligibility during the experiment is summarised in Table 8-4.

| Source | Sum of Squares (SS) | Degrees of Freedom (df) | Mean square (MS) | *F*-ratio | Sig. |
|---|---|---|---|---|---|
| *Within subjects* | | | | | |
| Time | 1402.897 | 1 | 1402.897 | 43.281 | 0.0 |
| *Interaction* | | | | | |
| Time * Processing type | 52.999 | 3 | 17.666 | 0.545 | 0.654 |
| Error | 1815.154 | 56 | 32.413 | | |

**Table 8-4: ANOVA data for the learning effect on intelligibility for the duration of the experiment**

The ANOVA data indicates there is a significant main effect of time, $F(1,56) = 43.281$, $MSE = 32.413$, $p < 0.05$. There is no significant interaction between the type of spatial processing and the intelligibility improvement during the experiment. That is, there is an improvement, no matter which of the processing methods in Table 8-2 is used. The result is in line with a visual inspection of Figure 8-5, which generally demonstrates a slight, but consistent rise in intelligibility. The L (Lens) condition is the only exception, as for this condition intelligibility plateaus after the first quarter of the main experiment.

### 8.4.3 Other processing interactions

Each subject was asked to disclose their gender, age and which hand they write with. The scatter plot in Figure 8-6 shows the intelligibility scores with respect to age.

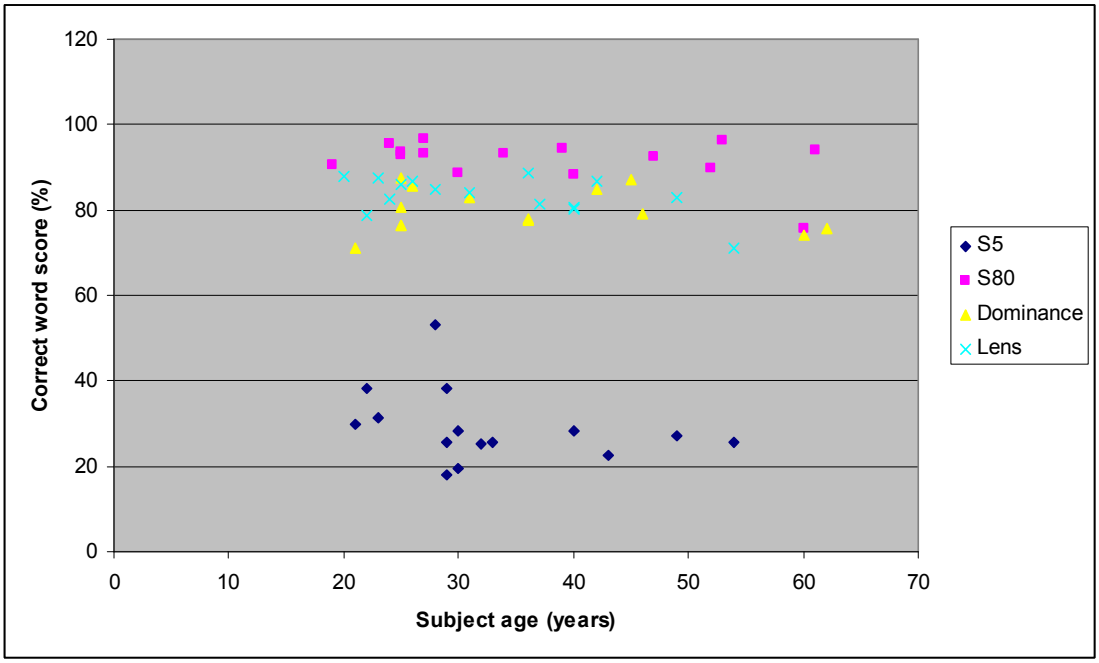



**Figure 8-6: Scatter plot of the relationship between age of the subject and their intelligibility score for each of the four spatial processing conditions.**

Visual inspection of the data shows there is no significant interaction between age and intelligibility performance. The data relating to gender and which hand the subjects use to write show no significant results and have therefore not been included.

# Chapter 9 Discussion

This chapter discusses the results of the listening experiment presented in Chapter 8, in which the intelligibility of target speech in interference was measured under four conditions. The outcome of the experiment is related back to the original hypothesis to consider the extent to which it is supported by our findings. Further work, based on the research reported in this thesis is also discussed.

## *9.1 The effect of spatial processing*

The results of the listening test reported in Chapter 8 have highlighted a number of important points. The most significant of these is confirmation that the intelligibility of a target speech sound source can, in appropriate circumstances, be improved by increasing the angular separation between it and interfering sound sources.

In the first pair of test conditions, S5 and S80, the target speech and interferer speech-based noise were spatialised using generic HRTFs. The target was always spatialised at 0° azimuth and in the S5 and S80 conditions the interferer was spatialised at -5° azimuth and -80° azimuth, respectively. For the sentences and test configurations used, the intelligibility scores increased from 25% to 93% for high predictive sentences and from 34% to 90% for low predictive sentences, when the angular separation was increased from 5° to 80°. It was anticipated that directly spatialising the interferer sounds at -80° would induce a binaural advantage and lead to a substantial improvement in intelligibility.

As a result of pilot listening tests, the dominance respatialisation method was expected to produce a perceptually similar auditory scene to the directly spatialised S80 condition and hence an intelligibility improvement which would match it closely. The intelligibility scores, 79% and 80% for high and low predictive sentences, respectively, confirm that the majority of the binaural

advantage due to angular separation is preserved. It is likely that respatialisation errors, due to significant portions of sounds being incorrectly processed, were responsible for reducing the effectiveness of the spatial separation. For example, when a TFU is marked as dominant for the interferer, it may nevertheless contain a significant amount of target energy. This will be incorrectly respatialised when that TFU is moved to 80° azimuth.

The lens respatialisation method has exceeded expectations in terms of intelligibility improvements. We had anticipated that it would perform significantly less well than the dominance method. In fact, the intelligibility scores, 85% and 82% for high and low predictive sentences respectively, are slightly greater than those achieved using the dominance method. This is believed to be due to smaller respatialisation errors using the lens method. A TFU is respatialised based on its instantaneous calculated location. The smooth mapping of processed locations ensures that the auditory scene is gradually widened, rather than switching between locations in a binary fashion, as is the case in the dominance method. It is possible that an even greater intelligibility performance could be realised using alternative parameter values for the auditory lens.

## *9.2 The effect of learning*

Subjects were presented with a training session lasting about 10 minutes. In comparison, the main experiment lasted approximately 30 minutes, though actual durations varied considerably from subject to subject, covering a range of 17 to 41 minutes. The training session was included with the intention that the bulk of subject learning would occur before the main experiment was attempted. In spite of this, the results shown in Figure 8-5 indicate a small learning effect for all four processing methods. That is, intelligibility improved during the experiment. The lens method showed the greatest increase of 8% from 77% to 85%. However, unlike the other processing methods, which increase reasonably linearly across all 100 test sentences, the lens method reaches its maximum intelligibility score after only 25 sentences and displays no further improvement. This suggests that

these subjects were able to adapt very quickly to the lens processing and exploit the binaural advantage it provides. It also indicates that the subjects exposed to the other processing methods were still adapting to the signals and may have improved further if they had been presented with more sentences. This is something that should be considered in future work.

## 9.3 Validation of hypothesis

The hypothesis for this research, as stated in Chapter 1, is:

*The intelligibility of a target speech source, in a binaural signal contaminated by spatially distinct interfering sounds, may be increased using an algorithm suitable for implementation in a hearing aid.*

The results of the psychoacoustic analysis clearly indicate that intelligibility is improved through spatial processing using the auditory lens algorithm. Therefore, the hypothesis is satisfied and validated. Importantly, the auditory lens achieves this improvement without having to identify and characterise the individual sound sources in the binaural soundfield in terms of their precise direction and spectral content. This is in complete contrast to the dominance method of processing the same signals, which suffers from having to determine the relative energies of target and interferer in each TFU.

## 9.4 Further work

It has been shown that the auditory lens provides an intelligibility improvement for the configurations used in the listening test. However, it will be essential to extend the scope of the configurations to reflect a wider range of listening environments. In addition, there are some algorithmic improvements that should be investigated.

### 9.4.1 Additional listening configurations

The listening tests described in Chapter 8 focus on whether the auditory lens can successfully improve intelligibility. For simplicity and convenience, the listening conditions were restricted to stationary sound sources and interferers on only one side of the listener. So that the auditory lens can be utilised in a real device such as a hearing aid or headset, further listening conditions must be considered.

One such condition involves the placement of interfering sound sources on both sides of the listener. This was considered as part of the current investigation. However, one of the limitations of the present algorithm is that for two interfering sound sources that have equal and opposite angular separation from the median plane there is an increased risk that they will be detected as lying in the focal region of the lens. This results in TFUs that are dominated by interferer sounds being incorrectly left unprocessed. Preliminary considerations suggest that the phase difference between corresponding left/right TFUs may behave differently over time for two laterally symmetrical interferers compared with these same two interferers when a target is present. It is envisaged that a measure of direction stability, estimated across both time and frequency, might assist the lens in reducing ambiguity in handling bilateral interfering sounds. Two approaches were briefly considered, however a detailed investigation and evaluation was not carried out and should be a topic of further work. The first method involved a simple weighted average of neighbouring TFUs within a 3x3 square to determine the direction of the TFU at the centre of the square. The intention of this method was to smooth out any large variances in the calculated direction across time and frequency. The second method involved applying a threshold of variance to a TFU across the time dimension. The average of the previous $L$ TFUs for a particular frequency bin is calculated and compared to the current TFU. If the current TFU is within a predetermined threshold from the average it is considered stable. It is expected that $L$ and the thresholds may be different for each frequency.

Another real listening condition that has not been investigated is reverberation. Using the existing auditory lens algorithm, it is expected that the diffuse reverberant tail, which contains weak random directional information, could cause target speech to appear more dispersed and attenuated. Reverberation can be considered as an extreme case of sounds on each side of the listener. Therefore, it is expected that direction stability estimation might also help with the processing of reverberant sounds.

A further listening condition that requires investigation is the use of moving sound sources. The investigation discussed in this thesis has only considered stationary sources. It is anticipated that the auditory lens algorithm in its current form will continue to perform well with moving sound sources. The changing ITD and IID will be processed irrespective of sound source movement. In processing terms, this is no different to the interaction between multiple spatialised sounds creating a varying ITD and IID at the listener's ears, although a target sound which moves out of the focal region of the auditory lens will become less intelligible. A simple head rotation tracker, based on miniature accelerometers, would enable a target to be tracked and remain in focus even in the presence of small head movements by the listener. Further listening tests will be needed to validate the performance of the auditory lens with multiple moving sound sources.

The auditory lens is based upon manipulation of ITD, with ILD manipulation above 1.5 kHz believed to be providing largely an aesthetic improvement to the processed signal rather than an improvement in intelligibility. Sources at elevations outside the horizontal plane will be processed by the auditory lens according to their ITD and not by the spectral features introduced by the pinnae. Another complication with the pinna spectral cues is their individual nature, since they are unique to a particular listener. Hence, it is suggested that issues associated with source elevation are a lower priority than the other factors which have been discussed. As yet, the capture and synthesis of HRTF data is not directly relevant to the technical work discussed in this thesis. However, it is expected that further advances in HRTF individualisation will benefit auditory lens processing in future.

### 9.4.2 Auditory lens improvements

Although the auditory lens algorithm has been shown to improve intelligibility, there are a number of improvements that should be considered. These adjustments were touched upon in the section describing the development of the current algorithm. However, it is necessary to carry out a detailed investigation into the specific performance improvements that can be achieved.

One component of the system that may benefit from further enhancement is the remapping of interaural differences. For the listening configurations used the aggressive transition between the focal region and the remapping curve works very well. The sound source locations have been selected such that they are either comfortably within the focal region, i.e. the target speech sound, or outside this region, where they will be processed. It is proposed that a smoother transition is implemented so that sound sources that are close to, or on the transition threshold are not excessively disrupted. Furthermore, for some users, it may be beneficial to provide a means of adjusting the position of the transition threshold and the aggressiveness with which it is applied to suit personal taste and the particular listening conditions.

As well as adjusting the ITD and IID of the signals arriving at each ear, the auditory lens also attenuates the TFUs that are not within the focal region. One improvement here could be to adjust the attenuation based on the calculated IID and ITD. For example, it may be preferable to attenuate sounds that are further away from the target sound source more than those that are closer. This will prevent any sharp transitions in level as a sound component moves in and out of the focal range. On the other hand, the smaller movement of some interferer components may lead to a reduction in binaural unmasking and hence reduce intelligibility gains. Furthermore, the maximum attenuation that is applied should be carefully controlled. The extreme case would be to set all TFUs outside the focal region to zero, thus completely removing all components classed as interference. It was illustrated in Section 7.4.1 that this could disrupt

the continuity of the target sound due to the spectral holes that would be introduced. Therefore, any attenuation of TFUs should consider the impact on the target speech, in terms of both intelligibility and quality.

### 9.4.3 Practical implementations

The audio processing algorithms discussed in this thesis have an additional, more practical, benefit beyond speech intelligibility enhancement. The processing architecture that has been proposed is quite simple and uses mathematical techniques, such as the FFT, which lend themselves to implementation on a digital signal processor (DSP). Although not a direct requirement for this research, the practical applications of the system have always been kept in mind. In particular, it is believed that this research will result in algorithms suitable for bilateral hearing aids and wireless audio devices such as Bluetooth headsets. Methods for improving speech processing algorithms within digital hearing aids have been investigated for many years. Those enhancements are now migrating to devices worn by normal hearing listeners, especially with the increase in popularity of wireless hands-free communication. Adapting algorithms for a small low power processor in a Bluetooth headset was a recent challenge, Spittle (2008). As hearing enhancement devices converge between those for hearing impaired and normal hearing listeners, the proposed auditory lens has increasing potential to provide a useful front-end processing system.

# Chapter 10 Conclusion

The aim of this research has been to develop a means of increasing the intelligibility of speech in interference using binaural signal processing. Targeted in the long term at people who have a hearing deficit amenable to binaural enhancement, our preliminary studies have been undertaken using subjects with normal hearing.

The investigation started with a detailed overview of the human hearing system in Chapter 2. This considers the localisation cues used by the auditory system to determine the direction and distance of a sound source from a listener. It is shown that ITD is the dominant cue below approximately 1500 Hz and IID and spectral information is important for higher frequencies. Chapter 3 builds on this foundation and discusses various methods for artificially spatialising sound sources. HRTFs are introduced, along with the mathematical techniques used for applying them to a mono sound to give the perceptual illusion of a spatialised sound.

Chapters 4, 5 and 6 provide an extensive review of the literature on improving the intelligibility of a target speech sound in a harsh auditory environment. Chapter 4 discusses the factors that influence auditory masking. The amount of masking imposed by an interfering sound on a target sound is affected by the duration and spectral content of the masking sounds. The type of signal is also important, as masking exists in two forms, energetic and informational, both of which are significant for speech target sounds. Another factor which influences intelligibility, directly relevant to this research, is the relative spatial locations of the target and interfering sounds. In general, an interferer will cause more disruption to a target sound the smaller their angular separation.

Chapter 5 considers the factors that influence the continuity of a sound when it is interrupted by interfering sounds. The auditory continuity illusion is explained with reference to a variety of target sounds, ranging from simple tones to

complex speech signals. It is shown that replacing segments of a sound with silence is easily detected, even for very short interruptions. If these temporal gaps are filled with a suitable sound then continuity of the target sound is more likely to occur. It is shown that such sounds induce the spectral activity in the hearing system that would have occurred if the target sound had not stopped. It follows that a wideband sound is more likely to induce continuity than a narrowband sound. Hence a broadband inducing sound, such as white noise, is often used to patch spectro-temporal holes in a signal and can help to perceptually recover some of the lost information. Excessive masking of informational components of the target can occur if the interfering sound is too loud and this has a detrimental effect on intelligibility. On the other hand, spectral holes in a speech signal can also disrupt intelligibility. Therefore, a balance must be struck between using too much and too little signal to induce continuity.

The influence of multiple interfering sound sources on a target speech sound is covered in Chapter 6. In particular, the overlap of spectral components between the sounds is considered. This leads to the use of "glimpses" of target speech during spectro-temporal regions of low interferer activity. This is referred to as the "peek" theory and allows the human hearing system to cope with understanding a target speech sound in very challenging acoustic environments. In general, intelligibility improves when the angular separation increases between interfering sound sources and the target sound. Known as binaural unmasking, this is another key focus of the research in this thesis.

The technical development of a signal processing algorithm is discussed in Chapter 7. Here, the challenge was to process a binaurally spatialised mixture of target speech and interferer sounds so as to increase the intelligibility of the target. Initially, the conventional spatialisation of these sound sources using HRTFs is described. As expected, informal listening tests displayed a clear improvement in intelligibility of the target speech as its angular separation from the interfering sounds was increased. Sounds spatialised with this method served to provide a baseline for subsequent algorithm development and analysis. Careful consideration was given to how this conventional approach to

spatialisation might be modified to increase artificially the spatial separation between target and interferer. Respatialisation of the interferer using HRTFs, however, ideally requires prior knowledge of the individual mono sound sources and their directions. Since, in a practical situation, this information is not directly available, some means of estimating them is necessary.

The direction of individual sound sources in a binaural mix can be estimated by creating a cross-correlogram and estimating dominant time lags between the left and right channels. This was simulated and shown to be capable of estimating the directions of a single source and of two concurrent sources. The dominance method was introduced as a way of respatialising a sound source in a binaural mix that only requires knowledge of the dominant sound source at each time-frequency component. Since access to complete *a priori* knowledge of the individual sources contributing to the binaural mix is out of the question, it was hoped that the simpler task of estimating the energetically dominant source components might lead to a good compromise between estimation requirements and intelligibility performance. The respatialisation of only the components where the interferer sound is dominant proved to be very effective. However, practical problems remain. It is questionable whether the cross-correlogram method provides a sufficiently accurate estimation of sound source direction for respatialisation. Furthermore, it is unsuitable for providing accurate direction information on the scale of each time-frequency component. Ratio estimations for determining signal dominance introduce additional complexity for which no promising solutions could be identified.

As a result, the investigation took a new route and began to consider how binaurally encoded sound sources can be moved without the need for accurate direction and amplitude information. The new algorithm is based on the observations that ITD and IID are the dominant cues for localisation below and above approximately 1500 Hz, respectively, and that intelligibility improvements are achievable without the need for individualised spatialisation of the sound sources. It was conjectured that carefully applied amplification of the binaural difference cues could not only give the perceptual illusion of greater sound source separation, but could also provide the sought-after increase in

intelligibility. This simple approach is primarily suited to relocating sound sources in the horizontal plane, which is the plane in which speech sources most commonly reside. Specifically, the target sound source is assumed to lie directly in front of the listener at 0° azimuth and interfering sounds are assumed to lie at a non-zero azimuth no closer to the target than some small angle. All sounds having non-zero interaural differences above this threshold will be relocated more laterally by the algorithm. We coin the term auditory lens for this form of processing due to its resemblance in some respects to an optical lens.

To gain a further improvement in target speech intelligibility, beyond what is possible by increasing spatial separation alone, interfering sounds are also partially attenuated. This, it is argued, beneficially reduces masking of the target. It might seem obvious to attempt to completely remove the interferer components. However, it is shown that this raises a number of issues. Firstly, a social consequence of the complete removal of interferer sounds is that it prevents the listener from being able to switch attention to an alternative sound source, should they wish to do so. At the signal processing level, the interferer components relocated by the lens algorithm still contain energy from the target sound source. Therefore, total removal of those components, when an interferer is particularly dominant, may result in substantial spectral holes appearing in the remaining target sound. The review of auditory continuity illustrated the detrimental impact on intelligibility which occurs when spectral or temporal holes are introduced in a speech signal. It would therefore be beneficial to fill the holes with an appropriate continuity-inducing signal and the interferers, which have indirectly led to the holes in the first place, are an obvious way to fill them. Therefore, it was concluded that there is an optimum attenuation to apply to interfering sounds and maximise intelligibility of the target speech. This attenuation both maintains auditory continuity, whilst preserving the auditory scene, and increases exposure of the target.

The ability of the auditory lens to improve intelligibility of speech in interference was evaluated through the listening tests reported in Chapter 8. This chapter covers the design of the experiment, the processing methods tested and the procedure. The results show a clear benefit to intelligibility when the auditory

lens is used to increase the angular separation between target speech and interference at 5° azimuth. For speech sentences presented with unidirectional multi-talker speech-based noise there was an increase in correct word scores from 25% to 85% for high predictive sentences and from 34% to 82% for low predictive sentences. This stood up well against the baseline condition in which the interferer was spatialised directly to 80° azimuth, where the correct word scores for high and low predictive sentences rose to 93% and 90%, respectively. The listening tests displayed a small learning effect during the experiment for all listening conditions. However, for the subjects listening through the auditory lens the learning effect had reached a maximum within 25 sentences. This may be interpreted as indicating that listeners were quickly able to exploit the binaural advantage it provided and were comfortable with the signal manipulations that had been applied.

This thesis set out to discuss a novel method for improving the intelligibility of a target speech sound when presented simultaneously with multiple interfering sounds. This has been successfully achieved for a unilateral interfering source. Suggestions are discussed for ways in which the auditory lens algorithm might be adapted to accommodate bilateral interferers and other adverse acoustic conditions.

# Appendix A Instructions to subjects

The following instructions were given to subjects as part of the listening experiment discuss in Chapter 8.

*The purpose of this experiment is to determine whether different speech processing techniques improve the intelligibility of speech in the presence of background noise.*

*Please enter your name, age, gender, which hand you use to write and whether you have any known hearing impairment. If you have a hearing impairment please enter additional details in the box provided.*

*All sentences will be presented in English. It is important that English is your primary language. If it is not, please do not continue the experiment.*

*The training phase will allow you to become familiar with the target speech. You are to listen to the sentence being spoken and then enter the words you hear into the box marked "sentence". For the training phase and the main experiment, you will hear each sentence one only. Press the "play" button to hear a sentence. Press the "submit" button once you have typed in the words you have heard. If you are unable to hear any of the words, you can leave the box empty and just press "submit". It is possible that none or all of the sentences can be understood, this is normal and there is no need to be concerned. This is due to the sentences that have been assigned to you. Please enter your response as accurately as possible. For example, "The cat sat on the mat". Alternatively, if you are unsure of one or more words, you mark this using "?", for example, "The ? sat ? mat".*

*It is important that the headphones are placed on your head correctly. The cable should lead to the headphone on your left ear.*

*If you are interrupted, for example by a loud external sound, or would like to make additional comments regarding the sentence, please enter the details into the "comment" box. For example "Dog barking outside during last sentence".*

*You will hear 100 sentences in total. The experiment will last approximately 20 minutes.*

*Thank you for taking part in the experiment.*

# Appendix B Table of results

The following table lists the results for the listening experiment. The percentage correct word scores are shown for 4 blocks of sentences, to determine whether intelligibility improved with time. Also, scores for high and low predictive sentences are given, with a total score for all sentences.

| | | Sentence index | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Subject** | **Processing** | **1 to 25** | **26 to 50** | **51 to 75** | **76 to 100** | **High predictive correct** | **Low predictive correct** | **Total correct** | **Hearing loss** |
| NH | S5 | 14.29 | 30.52 | 22.98 | 32.21 | 18.35 | 32.55 | 25.12 | |
| FW | S5 | 36.99 | 28.13 | 27.63 | 26.11 | 26.91 | 33.22 | 29.92 | |
| PH | S5 | 28.06 | 29.68 | 31.71 | 21.71 | 19.88 | 37.25 | 28.16 | |
| MW | S5 | 30.71 | 19.75 | 29.61 | 20.86 | 26.91 | 23.83 | 25.44 | |
| SK | S5 | 15.79 | 22.01 | 9.43 | 27.33 | 14.37 | 25.50 | 19.68 | |
| CT2 | S5 | 24.46 | 25.16 | 27.33 | 25.32 | 19.27 | 32.55 | 25.60 | Moderate loss in left ear |
| SP | S5 | 32.88 | 27.85 | 21.62 | 30.63 | 23.55 | 33.22 | 28.16 | |
| AC | S5 | 44.60 | 50.93 | 56.33 | 59.48 | 50.46 | 55.70 | 52.96 | |
| JS3 | S5 | 17.14 | 20.78 | 12.35 | 21.02 | 15.29 | 21.14 | 18.08 | |
| KD | S5 | 16.89 | 25.48 | 32.88 | 26.42 | 19.57 | 32.55 | 25.76 | |
| PT | S5 | 25.90 | 27.95 | 22.15 | 35.29 | 23.24 | 31.54 | 27.20 | |
| JW2 | S5 | 30.82 | 34.81 | 44.87 | 41.83 | 35.17 | 41.95 | 38.40 | |
| ST | S5 | 17.57 | 43.42 | 41.77 | 50.97 | 32.11 | 44.63 | 38.08 | |
| JM | S5 | 29.08 | 27.10 | 35.71 | 32.10 | 28.13 | 35.23 | 31.52 | |
| IW | S5 | 20.55 | 15.85 | 22.67 | 30.92 | 18.04 | 27.18 | 22.40 | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | S5 Average | 25.71 | 28.63 | 29.27 | 32.15 | 24.75 | 33.87 | 29.10 | |
| | | | | | | | | | |
| DH | S80 | 94.56 | 96.77 | 88.82 | 97.33 | 94.50 | 93.96 | 94.24 | |
| MC | S80 | 90.14 | 92.21 | 94.97 | 93.67 | 95.72 | 89.93 | 92.96 | |
| RP | S80 | 89.13 | 95.48 | 92.55 | 91.77 | 94.50 | 90.27 | 92.48 | |
| PK | S80 | 77.62 | 91.50 | 92.26 | 92.55 | 87.46 | 89.93 | 88.64 | |
| K | S80 | 72.22 | 66.05 | 83.44 | 79.49 | 75.23 | 75.84 | 75.52 | |
| M | S80 | 91.78 | 91.03 | 98.10 | 96.05 | 94.80 | 93.29 | 94.08 | |
| DT | S80 | 94.08 | 96.62 | 96.89 | 98.69 | 97.25 | 95.97 | 96.64 | |
| PC | S80 | 92.31 | 92.90 | 93.59 | 94.30 | 96.94 | 89.60 | 93.44 | |
| DH2 | S80 | 94.44 | 95.63 | 96.69 | 97.45 | 98.17 | 93.96 | 96.16 | |
| TG | S80 | 93.20 | 90.60 | 96.86 | 95.51 | 94.19 | 92.28 | 93.28 | |
| MS | S80 | 90.85 | 96.05 | 99.40 | 94.74 | 98.47 | 92.28 | 95.52 | |
| MH | S80 | 85.62 | 85.90 | 90.51 | 90.79 | 89.91 | 86.24 | 88.16 | |
| CC | S80 | 90.14 | 89.74 | 97.32 | 95.18 | 94.50 | 91.95 | 93.28 | |
| PM | S80 | 84.83 | 97.47 | 84.57 | 94.63 | 91.44 | 89.60 | 90.56 | |
| EB | S80 | 89.71 | 83.23 | 94.58 | 91.03 | 90.83 | 88.93 | 89.92 | |
| | S80 Average | 88.71 | 90.75 | 93.37 | 93.55 | 92.93 | 90.27 | 91.66 | |
| | | | | | | | | | |
| WS | D | 86.86 | 77.33 | 82.21 | 87.58 | 81.04 | 84.90 | 82.88 | |
| JW | D | 71.81 | 78.71 | 79.62 | 77.92 | 74.92 | 80.20 | 77.44 | |
| GA | D | 82.64 | 88.82 | 81.53 | 86.16 | 84.71 | 85.23 | 84.96 | |
| JB | D | 78.42 | 80.13 | 71.78 | 80.65 | 78.90 | 76.85 | 77.92 | 4 kHz notch in left ear |
| NA | D | 66.90 | 80.65 | 77.56 | 77.85 | 76.76 | 75.84 | 76.32 | |
| JR | D | 64.29 | 71.78 | 69.48 | 77.07 | 67.89 | 74.50 | 71.04 | |
| BC | D | 64.86 | 76.10 | 77.56 | 79.73 | 72.48 | 76.17 | 74.24 | |
| TW | D | 71.33 | 82.78 | 84.52 | 82.00 | 79.51 | 81.88 | 80.64 | |
| JT | D | 81.76 | 87.42 | 91.03 | 91.22 | 87.77 | 87.58 | 87.68 | |
| CS | D | 82.76 | 86.25 | 87.33 | 90.45 | 87.77 | 86.24 | 87.04 | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| DM | D | 73.57 | 71.52 | 81.53 | 73.89 | 73.09 | 78.52 | 75.68 | |
| SS | D | 74.82 | 78.21 | 80.63 | 83.87 | 79.51 | 78.52 | 79.04 | |
| SC | D | 80.14 | 90.07 | 88.82 | 82.72 | 86.85 | 83.89 | 85.44 | Tinnitus & Dyslexic |
| GH | D | 65.03 | 71.24 | 81.76 | 85.99 | 77.06 | 76.17 | 76.64 | |
| RI | D | 71.33 | 70.78 | 83.33 | 80.38 | 76.45 | 76.51 | 76.48 | |
| | D Average | 74.43 | 79.45 | 81.25 | 82.50 | 78.98 | 80.20 | 79.56 | |
| | | | | | | | | | |
| PS | L | 66.90 | 86.09 | 87.26 | 78.62 | 82.26 | 77.85 | 80.16 | |
| RC | L | 77.40 | 89.17 | 94.19 | 90.20 | 92.05 | 80.87 | 86.72 | |
| CT | L | 61.38 | 74.68 | 73.25 | 71.71 | 68.81 | 73.49 | 71.04 | |
| GS | L | 82.98 | 94.44 | 87.16 | 89.44 | 89.60 | 87.92 | 88.80 | |
| EC | L | 78.87 | 90.00 | 89.17 | 91.03 | 90.21 | 84.90 | 87.68 | |
| PR | L | 81.76 | 83.66 | 76.40 | 89.26 | 81.04 | 84.90 | 82.88 | |
| NS2 | L | 60.96 | 87.01 | 86.75 | 85.63 | 83.49 | 77.85 | 80.80 | |
| PW | L | 85.31 | 83.87 | 86.09 | 83.44 | 87.46 | 82.21 | 84.96 | |
| NC | L | 76.87 | 72.73 | 89.81 | 85.06 | 84.10 | 78.52 | 81.44 | |
| RG | L | 75.18 | 93.21 | 77.70 | 95.03 | 90.21 | 81.54 | 86.08 | |
| MH2 | L | 82.39 | 91.25 | 87.90 | 88.46 | 86.85 | 88.93 | 87.84 | |
| LF | L | 71.23 | 72.33 | 86.39 | 84.38 | 81.65 | 75.84 | 78.88 | |
| JU | L | 78.68 | 84.97 | 84.42 | 83.83 | 85.02 | 79.87 | 82.56 | |
| JS2 | L | 87.41 | 90.45 | 90.57 | 77.12 | 86.24 | 87.25 | 86.72 | |
| JS | L | 87.50 | 82.58 | 83.97 | 81.94 | 85.32 | 82.89 | 84.16 | |
| | L Average | 76.99 | 85.10 | 85.40 | 85.01 | 84.95 | 81.66 | 83.38 | |

**Table B-1: Table of percentage correct word scores for participants in the listening experiment.**

# Appendix C Sentences for listening experiment

The list of sentences in Table C-1 were used for the target speech for the listening experiment discussed in Chapter 8. 'H' denotes a high predictive sentence and 'L' a low predictive sentence. These are taken from Kalikow *et al*. (1977), Form 2.1 and Form 2.2.

| Predictive | Sentence |
|---|---|
| H | The watchdog gave a warning growl |
| H | She made the bed with clean sheets |
| L | The old man discussed the dive |
| L | Bob heard Paul called about the strips |
| L | I should have considered the map |
| H | The old train was powered by steam |
| H | He caught the fish in his net |
| L | Miss brown shouldn't discuss the sand |
| H | Close the window to stop the draft |
| H | My TV has a twelve inch screen |
| L | They might have considered the hive |
| L | David has discussed the dent |
| H | The sandal has a broken strap |
| H | The boat sailed along the coast |
| H | Crocodiles live in muddy swamps |
| L | He can't consider the crib |
| H | The farmer harvested his crop |
| H | All the flowers were in bloom |
| L | I am thinking about the knife |
| L | David does not discuss the hug |
| H | She wore a feather in her cap |
| L | We've been discussing the crates |
| L | Miss black knew about the doll |
| H | The admiral commands the fleet |
| L | She couldn't discuss the pine |
| L | Miss black thought about the lap |
| H | The beer drinkers raised their mugs |
| H | He was hit by a poisoned dart |
| H | The bread was made from whole wheat |
| L | Mr black knew about the pad |
| L | You heard Jane called about the van |
| H | I made the phone call from a booth |
| L | Tom wants to know about the cake |
| L | She's spoken about the bomb |
| H | The cut on his knee formed a scab |
| L | We hear you called about the lock |
| L | The old man discussed the yell |
| H | His boss made him work like a slave |

| | |
|---|---|
| H | The farmer baled the hay |
| L | They're glad we heard about the track |
| H | A termite looks like an ant |
| H | Airmail requires a special stamp |
| H | Football is a dangerous sport |
| L | Sue was interested in the bruise |
| L | Ruth will consider the herd |
| H | We saw a flock of wild geese |
| L | The girl talked about the gin |
| L | Paul can't discuss the wax |
| H | Drop the coin through the slot |
| L | I hope paul asked about the mate |
| L | You're glad they heard about the slave |
| L | The girl knows about the swamps |
| H | Hold the baby on your lap |
| H | For your birthday I baked a cake |
| H | The railroad train ran off the track |
| L | They did not discuss the screen |
| L | They were interested in the strap |
| H | Tear off some paper from the pad |
| L | I had a problem with the bloom |
| L | Peter should speak about the mugs |
| H | The fruit was shipped in wooden crates |
| H | The rancher rounded up his herd |
| L | She wants to speak about the ant |
| L | We're discussing the sheets |
| L | The boy would discuss the scab |
| H | The lonely bird searched for its mate |
| L | Tom could have thought about the sport |
| L | You'd been considering the geese |
| H | They drank a whole bottle of gin |
| H | On the beach we play in the sand |
| L | Mr Black considered the fleet |
| H | The airplane went into a dive |
| H | We're lost so let’s look at the map |
| L | I want to know about the crop |
| H | Household goods are moved in a van |
| H | The honey bees swarmed round the hive |
| L | Betty has talked about the draft |
| L | Tom discussed the hay |
| L | Jane was interested in the stamp |
| H | The airplane dropped a bomb |
| H | Cut the bacon into strips |
| L | I had not thought about the growl |
| H | The drowning man let out a yell |
| H | I gave her a kiss and a hug |
| L | Paul should know about the net |
| H | I cut my finger with a knife |
| H | The candle flame melted the wax |
| L | Tom heard Jane called about the booth |
| L | We can't consider the wheat |

| | |
|---|---|
| H | This key won't fit in the lock |
| L | We have not discussed the steam |
| L | Miss Brown might consider the coast |
| L | Mr Brown can't discuss the slot |
| H | The little girl cuddled her doll |
| H | Tom fell down and got a bad bruise |
| L | He hasn't considered the dart |
| H | The furniture was made of pine |
| H | How did your car get that dent |
| L | Mr Smith thinks about the cap |
| H | The baby slept in his crib |

**Table C-1: List of sentences used for the target speech sound source.**

# Appendix D Contents of accompanying CD

This thesis has an accompanying CD that contains all of the audio, Matlab® source code and listening experiment results that have been discussed. There are three folders which have the contents as described in Table D-1.

| Folder name | Description |
|---|---|
| Audio | The unprocessed recordings of the target speech sentences.<br>The sentences processed using each of the four spatialisation methods, S5, S80, D and L, as discussed in Section 8.2.6.<br>The speech-based noise audio files. |
| Matlab | All of the Matlab code used for processing the audio files, including the algorithms for the auditory lens. |
| Results | All of the results from the listening experiment. |

**Table D-1: Description of the folders on the accompanying CD.**

# Appendix E Glossary of mathematical terms

A summary of the mathematical terms used in this thesis are given in Table E-1.

| | |
|---|---|
| $\delta(p, k)$ | Phase difference $\delta$ at frequency bin index $k$ and time window index $p$. |
| $\delta'(p, k)$ | Augmented phase difference $\delta'$ at frequency bin index $k$ and time window index $p$. |
| $\Delta(p, k)$ | Magnitude difference $\Delta$ at frequency bin index $k$ and time window index $p$. |
| $\Delta'(p, k)$ | Augmented magnitude difference $\Delta'$ at frequency bin index $k$ and time window index $p$. |
| $F_D[...]$ | Discrete Fourier Transform. |
| $F_D^{-1}[...]$ | Inverse Discrete Fourier Transform. |
| $F_M(p, k)$ | Spectral mask for TFU at frequency bin $k$ and time window index $p$. |
| HRIR($n$, $\theta$, $\phi$) | Time domain head related impulse response of length N, at index $n$, azimuth $\theta$ and elevation $\phi$. |
| HRTF($k$, $\theta$, $\phi$) | Frequency domain head related transfer function, at frequency bin index $k$, azimuth $\theta$ and elevation $\phi$. |
| HRTF$^{-1}$($k$, $\theta$, $\phi$) | Inverse head related transfer function. |
| $i$ | Time domain index, typically for a non-windowed signal |
| IID | Interaural intensity difference |
| $IID_{FR}$ | IID focal range |
| $IID_{PK}$ | Peak IID offset |
| $IID_{PR}$ | IID processing range |
| ITD | Interaural time difference |
| $ITD_{FR}$ | ITD focal range |

| | |
|---|---|
| $ITD_{PK}$ | Peak ITD offset |
| $ITD_{PR}$ | ITD processing range |
| $n$ | Time domain index, typically for a windowed signal |
| $s_A(i)$, $S_A(k)$ | Mono signal from source A (time, frequency) |
| $s_B(i)$, $S_B(k)$ | Mono signal from source B (time, frequency) |
| $x(i)$ | Time domain signal $x$ at time sample $i$, typically the input to an algorithm. |
| $\hat{x}(i)$ | Estimated despatialised signal $\hat{x}$ at time sample $i$. |
| $X(k)$ | Frequency domain signal $X$ at frequency bin index $k$. |
| $X(p, k)$ | Frequency domain signal $X$ at frequency bin index $k$ and time window index $p$. Referred to as a time-frequency unit, TFU. |
| $w(n)$ | Windowing function of length $N$. |
| $y(i)$ | Time domain signal $y$ at time sample $i$, typically the output of an algorithm. |

**Table E-1: Summary of mathematical terms.**

# References


Advanced Micro Devices Inc. (1997), "System and method for interactive approximation of a head transfer function", 813454(6,181,800)

Algazi, V. R., Avendano, C. and Thompson, D. (1999), "Dependence of subject and measurement position in binaural signal acquisition" *Journal of the Audio Engineering Society,* Vol. 47, No. 11, pp. 937-947

Algazi, V. R., Avendano, C. and Duda, R. O. (2001), "Elevation localization and head-related transfer function analysis at low frequencies", *Journal of the Acoustical Society of America,* Vol. 109, No. 3, pp. 1110-1122

Allred, D.J., (2006), "Evaluation and comparison of beamforming algorithms for microphone array speech processing" *MSc. Thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology*

Asano, F., Suzuki, Y. and Sone, T. (1990), "Role of spectral cues in median plane localization" *Journal of the Acoustical Society of America,* Vol. 88, No. 1, pp. 159-168

Avendano, C., Algazi, V. R. and Duda, R. O. (1999a), "A head-and-torso model for low-frequency binaural elevation effects", *1999 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics,* pp. 179-182

Avendano, C., Duda, R. O. and Algazi, V. R. (1999b), "Modeling the contralateral HRTF", *Presented at the AES 16th international conference on spatial sound reproduction,* pp. 313-318

Barry, D., Fitzgerald, D., Coyle, E. and Lawlor, B. (2005), "Single channel source separation using short-time independent component analysis", *Presented at the AES 119th Convention, New York, October 7-10, 2005*.

Bashford, J. A. and Warren, R. M. (1979), “Perceptual synthesis of deleted phonemes”, *Speech communication papers, edited by J. J. Wolf and D. H. Klatt, New York, Acoustical Society of America,* pp. 423-426

Bashford, J. A. and Warren, R. M. (1987), “Multiple phonemic restorations follow the rules for auditory induction”, *Perception and Psychophysics*, Vol. 42, No. 2, pp. 114-121

Bashford, J. A., Meyers, M. D., Brubaker, B. S. and Warren, R. M. (1988), “Illusory continuity of interrupted speech: speech rate determines durational limits”, *Journal of the Acoustical Society of America*, Vol. 84, No. 5, pp. 1635-1638

Bashford, J. A., Reiner, K. R. and Warren, R. M. (1992), “Increasing the intelligibility of speech through multiple phonemic restorations”, *Perception and Psychophysics,* Vol. 51, No. 3, pp. 211-217

Bashford, J. A., Warren, R. M. and Brown, C. A. (1996), “Use of speech-modulated noise adds strong ‘bottom-up’ cues for phonemic restoration”, *Perception and Psychophysics,* Vol. 58, No. 3, pp. 342-350

Batteau, D. W. (1967), "The role of the pinna in human localization", *Proc. Royal Society London 168 (series B)*, pp. 158-180

Begault, D. R., Wenzel, E. M. and Anderson, M. R. (2000), “Direct comparison of the impact of head tracking, reverberation and individualized head related transfer functions on the spatial perception of a virtual speech source”, *Journal of the Audio Engineering Society,* Vol. 49, No. 10, pp. 904-916.

Berkhout, A. J. (1988), “A holographic approach to acoustic control”, *Journal of the Audio Engineering Society,* Vol. 36, No. 12, pp. 977-995.

Best, V., Gallun, F. J., Ihlefeld, A. and Shinn-Cunningham, B. G. (2006) "The influence of spatial separation on divided listening", *Journal of the Acoustical Society of America,* Vol. 120, No. 3, pp. 1506-1516.

Best, V., Gallun, F. J., Carlile, S. and Shinn-Cunningham, B. G. (2007) "Binaural interference and auditory grouping", *Journal of the Acoustical Society of America,* Vol. 121, No. 2, pp. 1070-1076.

Binns, C. and Culling, J. F. (2007) "The role of the fundamental frequency contours in the perception of speech against interfering speech". *Journal of the Acoustical Society of America,* Vol. 122, No. 3, pp. 1765 - 1776

Blommer, M. A. and Wakefield G. H. (1992), "Investigation of phase distortion in the synthesis of head-related transfer functions", *Journal of the Acoustical Society of America,* Vol. 92, No. 4 pt. 2, pp. 2297-2297

Blommer, M. A. and Wakefield G. H. (1994), "On the design of pole-zero approximations for head-related transfer functions using a logarithmic error measure", *IEEE Transactions on Signal Processing,* Vol. 42, No. 11, pp. 3245-3248

Blommer, M. A. and Wakefield G. H. (1997), "Pole-zero approximations for head-related transfer functions using a logarithmic error criterion", *IEEE Transactions on speech and audio processing,* Vol. 5, No. 3, pp. 278-287

Bronkhorst, A.W. (1995), "Localisation of real and virtual sound sources", *Journal of the Acoustical Society of America,* Vol. 98, No. 5, pp. 2542-2553

Brungart, D. (1998), "Control of perceived distance in virtual audio displays", *Proceedings of the 20th Annual International conference of the IEEE Engineering in Medicine and Biology Society,* Vol. 20, No. 3, pp. 1101-1104

Brungart, D. (1999a), "Auditory parallax effects in the HRTF for nearby sources", *Proceedings 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* pp. 171-174

Brungart, D. S. and Rabinowitz, W. M. (1999b), "Auditory localization of nearby sources. Head-related transfer functions", *Journal of the Acoustical Society of America,* Vol. 106, No. 3, pp. 1465-1479

Brungart, D. S., Durlach, D. I. and Rabinowitz, W. M. (1999c), "Auditory localization of nearby sources. II. Localization of a broadband source", *Journal of the Acoustical Society of America,* Vol. 106, No. 4 pt 1, pp. 1956-1968

Brungart, D. S. (1999d), "Auditory localization of nearby sources. III. Stimulus effects", *Journal of the Acoustical Society of America,* Vol. 106, No. 6, pp. 3589-3602

Brungart, D. (2001a), "Informational and energetic masking effects in the perception of two simultaneous talkers", *Journal of the Acoustical Society of America*, Vol. 109, No. 3, pp. 1101-1109

Brungart, D., Simpson, B. D., Ericson, M. A. and Scott, K. R. (2001b), "Informational and energetic masking effects in the perception of multiple simultaneous talkers", *Journal of the Acoustical Society of America,* Vol. 110, No. 5 Pt 1, pp. 2527-2538

Brungart, D. and Simpson, (2007) "Effect of target-masker similarity on across-ear interference in a dichotic cocktail-party listening task", *Journal of the Acoustical Society of America,* Vol. 122, No. 3, pp. 1724-1734

Brutti, A., Omologo, M. and Svaizer, P. (2008), "Comparison between different sound source localization techniques based on a real data collection", *Hands-free Speech Communication and Microphone Arrays 2008,* pp. 69-72

Butler, R. A. and Humanski, R. A. (1992), "Localization of sound in the vertical plane with and without high-frequency spectral cues", *Perception and Psychophysics,* Vol. 51, No. 2, pp. 182-186

Butler, R. A. and Musicant, A. D. (1993), "Binaural localization: influence of stimulus frequency and the linkage to covert peak areas", *Hearing Research,* Vol. 67, No. 1-2, pp. 220-229

Carlile, S. and Pralong, D. (1994), "The location-dependent nature of perceptually salient features of the human head-related transfer function", *Journal of the Acoustical Society of America,* Vol. 95, No. 6, pp. 3445-3459

Chang, P. R. and Tan, T.-H. (1998), "Fuzzy neural systems for controlling sound localisation in stereophonic reproduction", *Control theory and applications, IEE Proceedings,* Vol. 145, No. 4, pp. 393-401

Chen, J. S., Vanveen, B. D. and Hecox, K. E. (1995), "A spatial feature extraction and regularization model for the head-related transfer function", *Journal of the Acoustical Society of America,* Vol. 97, No. 1, pp. 439-452

Cheng, C. I. and Wakefield, G. H. (1999), "Spatial frequency response surfaces: An alternative visualization tool for head-related transfer functions (HRTF's)", *Acoustics, Speech and Signal Processing, 1999, Proceedings, IEEE International Conference on,* Vol. 2, pp. 961-964

Cheng, C. I. and Wakefield, G. H. (2001), "Introduction to head-related transfer functions (HRTFs): representations of HRTFs in time, frequency and

space", *Journal of the Audio Engineering Society,* Vol. 49, No.4, pp. 231-249

Cherry, E. C. (1953), "Some experiments on the recognition of speech with one and two ears", *Journal of the Acoustical Society of America* Vol. 25, No. 5, pp. 975-979

Cherry, E. C. and Wiley, R. (1967), "Speech communications in very noisy environments", *Nature*, Vol. 214, p. 1164

Cheung, N., Trautmann, S. and Horner, A (1998a), "Head-related transfer function modeling in 3-D sound systems with genetic algorithms", *Journal of the Audio Engineering Society,* Vol. 46. No. 6, pp. 531-539

Cheung, N., Trautmann, S. and Horner, A. (1998b), "Head related transfer function modeling in 3-D sound systems with genetic algorithms", *Acoustics, Speech and Signal Processing, 1998, Proceedings of the 1998 IEEE International Conference on,* Vol. 6, pp. 3529-3532

Christensen, F., Jensen, C. B. and Moller, H. (2000), "The design of VALDEMAR — an artificial head for binaural recording purposes", *Presented at the AES 109th Convention, Los Angeles, September 22-25, 2000*.

CIPIC Interface Laboratory (1998), "Documentation for the UCD HRIR files", *University of California at Davis*

Coleman, P. D. (1968), "Dual role of frequency spectrum in determination of auditory distance", *Journal of the Acoustical Society of America* Vol. 44, No. 2, pp. 631-632

Cooney, R., Cahill, N. and Lawlor, R., (2006), "An enhanced implementation of the ADRess (Azimuth Discrimination and Resynthesis) music source

separation algorithm", *Presented at the 121st AES Convention, San Francisco, October 5-8, 2006*

Culling, J. F. and Colburn, H. S. (2000), "Binaural sluggishness in the perception of tone sequences and speech in noise", *Journal of the Acoustical Society of America,* Vol. 107, No. 1, pp. 517-527

Daniel, J., Nicol, R. and Moreau, S. (2003), "Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging" *Presented at the AES 114th Convention, Amsterdam, March 22-25 2003*

Darwin, C. J., Akeroyd, M. A. and Hukin, R. W. (2002), "Binaural factors in auditory continuity", *Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan,* pp. 1-4

Das, N., Routray, A. and Dash, P. K., (2007), "ICA Methods for blind source separation of instantaneous mixtures: A Case Study", *Neural Information Processing – Letters and Reviews,* Vol. 11, No. 11, pp. 225-246

Davies, M., Jafari, M., Abdallah, S., Vincent, E. and Plumbley, M. (2007), "Blind source separation using space-time independent component analysis", *Blind Speech Separation,* Makino, S., Lee, T-W, and Sawada, H. Eds., pp. 79-99, Springer, 2007

DeSimio, M. P., Anderson, T. R. and Westerkamp, J. (1996), "Phoneme recognition with a model of binaural hearing", *IEEE Transactions on Speech and Audio Processing,* Vol. 4, No. 3, pp. 157-166

Douglas, S. C. and Gupta, M., (2007), "Convolutive blind source separation for audio signals", *Blind Speech Separation,* Makino, S., Lee, T-W, and Sawada, H. Eds., pp. 3-45, Springer, 2007

Drake, C. and McAdams, S. (1999), "The auditory continuity phenomenon: Role of temporal sequence structure", *Journal of the Acoustical Society of America,* Vol. 106, No. 6, pp. 3529-3538

Drullman, R. and Bronkhurst, A. W. (2000), "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation" *Journal of the Acoustical Society of America,* Vol. 107, No. 4, pp. 2224-2235

Duda, R. O. and Martens, W. L. (1998), "Range-dependence of the response of a spherical head model", *Journal of the Acoustical Society of America,* Vol. 104, No. 5, pp 3048 – 3058

Duda, R. O. Avendano, C., Algazi, V. R. (1999), "An adaptable ellipsoidal head model for the interaural time difference", *Acoustics, Speech and Signal Processing, 1999, Proceedings, IEEE International Conference on,* pp. 965 968

Durlach, N. I. (1964), "Note on binaural masking level differences at high frequencies", *Journal of the Acoustical Society of America,* Vol. 36, No. 3, pp. 576-581

Edmonds, B. A. and Culling, J. F. (2005) "The spatial unmasking of speech: evidence for within-channel processing of interaural time delay". *Journal of the Acoustical Society of America,* Vol. 117, No. 5, pp. 3069-3078

Edmonds, B. A. and Culling, J. F. (2006) "The spatial unmasking of speech: Evidence for better ear listening.", *Journal of the Acoustical Society of America,* Vol. 120, No. 3, pp. 1539-1545

Elfner, L. F. and Homick, J. L. (1967), "Continuity effects with alternately sounding tones under dichotic presentation", *Perception and Psychophysics,* Vol. 2, No. 1, pp. 34-36

Elfner, L. F. (1971), "Continuity in alternately sounded tonal signals in a free field", *Journal of the Acoustical Society of America,* Vol. 49, No.2 pt. 2, pp. 447-449

Evans, M. J., Angus, J. A. S. and Tew, A. I. (1998), "Analyzing head-related transfer functions using surface spherical harmonics", *Journal of the Acoustical Society of America,* Vol. 104, No. 4, pp. 2400-2411

Faller, C and Baumgarte, F. (2001). "Efficient representation of spatial audio using perceptual parameterization", *presented at IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* Mohonk Mountain House, New Paltz, NY

Feuerstein, J. F. (1992), "Monaural versus binaural hearing: ease of listening, word recognition, and attentional effort", *Ear Hear*, Vol. 13, No. 2, pp. 80-86

Fletcher, (1940) "Auditory patterns", *Review of Modern Physics*, Vol. 12, No.1, pp. 47-65

Freyman, R. L., Helfer, K. S., McCall, D. D. and Clifton, R. K. (1999), "The role of perceived spatial separation in the unmasking of speech", *Journal of the Acoustical Society of America,* Vol. 106, No. 6, pp. 3578-3588

Freyman, R. L., Balakrishnan, U. and Helfer, K. S. (2001), "Spatial release from informational masking in speech recognition" , *Journal of the Acoustical Society of America*, Vol. 109, No. 5, pp 2112-2122.

Gallun, F. J., Mason, C. R. and Kidd, G., Jr. (2007) "The ability to listen with independent ears", *Journal of the Acoustical Society of America*, Vol. 122, No. 5, pp 2814-2825

Gimsing, S. (2008), "Use of hearing aids five years after issue", *Ugeskrift for Laeger,* Vol. 170, No. 43, pp. 3407-3411.

Hall, J.W., Haggard, M.P. and Fernandes, M.A. (1984), "Detection in noise by spectro-temporal pattern analysis", *Journal of the Acoustical Society of America,* Vol. 76, No. 1, pp. 50-56.

Hall, J.W., Buss, E. and Grose, J.H. (2007), "Spectral integration and wideband analysis in gap detection and overshoot paradigms", *Journal of the Acoustical Society of America,* Vol. 122, No. 6, pp. 3598-3608.

Hammershoi, D., Møller, H. and Sorensen, M. F. (1992), "Head related transfer functions: measurement on 24 human subjects", *Presented at the 92nd Convention of the AES, Vienna, 1992*

Hartmann, W. M. (1999), "How we localise sound", *Physics Today,* November, pp. 24-29

Hawley, M. L., Litovsky, R. and Colburn, H. S. (1999), "Speech intelligibility and localization in a multi-source environment", *Journal of the Acoustical Society of America,* Vol. 105, No. 6, pp. 3436-3448

Healy, E. W. and Bacon, S. P. (2007) "Effect of spectral frequency range and separation on the perception of asynchronous speech", *Journal of the Acoustical Society of America*, Vol. 121, No. 3, pp. 1691-1700.

Henning, G. B. (1974), "Detectability of interaural delay in high-frequency complex waveforms", *Journal of the Acoustical Society of America,* Vol. 55, No. 1, pp. 84-90

Hetherington, C. and Tew, A. I. (2003), "Parameterizing human pinna shape for the estimation of head-related transfer functions", *Presented at the 114th AES convention, Amsterdam, March 22-25, 2003*

Houtgast, T. (1972), “Psychophysical evidence for lateral inhibition in hearing”, *Journal of the Acoustical Society of America,* Vol. 51, No. 6, pp. 1885-1894

Humanski, R. A. and Butler, R. A. (1988), “The contribution of the near and far ear toward localization of sound in the sagittal plane”, *Journal of the Acoustical Society of America,* Vol. 83, No. 6, pp. 2300-2310

Huopaniemi, J. and Riederer, K. A. J. (1998), “Measuring and modeling the effect of source distance in head-related transfer functions”, *Proc. Joint meeting of the Acoustical Society of America and the International Congress on Acoustics (ICA/ASA'98),* pp. 2083-2084

Iyer, N., Brungart, D. S. and Simpson, B. D. (2007) “Effects of periodic masker interruption on the intelligibility of interrupted speech”, *Journal of the Acoustical Society of America,* Vol. 122, No. 3, pp 1693-1701

Jenison, R. L. (1995), “A spherical basis function neural network for pole zero modelling of head-related transfer functions”, *Applications of signal processing to audio and acoustics, 1995, IEEE ASSP Workshop on,* pp. 92-95

Johnstone, P. M., Litovsky, R. (2006) “Effect of masker type and age on speech intelligibility and spatial release from masking in children and adults”, *Journal of the Acoustical Society of America,* Vol. 120, No. 4, pp. 2177-2189

Kahana, Y., Nelson, P. A., Petyt, M. and Choi, S. (1999), “Numerical modelling of the transfer functions of a dummy-head and of the external ear”, *Presented at the AES 16$^{th}$ international conference on spatial sound reproduction*

Kalikow, D. N., Stevens, K. N. and Elliott, L. L. (1977), "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability", *Journal of the Acoustical Society of America,* Vol. 61, No. 5, pp. 1337-1351

Kashino, M. and Warren, R. M. (1996), "Binaural release from temporal induction", *Perception and Psychophysics,* Vol. 58, No. 6, pp. 899-905

Katz, B. F. G. (2001a), "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation", *Journal of the Acoustical Society of America,* Vol. 110, No. 5 pt. 1, pp. 2440-2448

Katz, B. F. G. (2001b), "Boundary element method calculation of individual head-related transfer function. II. Impedance effects and comparisons to real measurements", *Journal of the Acoustical Society of America,* Vol. 110, No. 5 pt. 1, pp. 2449-2455

Kelly, M. C., and Tew, A. I. (2002), "The continuity illusion in virtual auditory space" *Presented at the AES 112$^{th}$ Convention, Munich, May 10-13 2002*

Kewley-Port, D., Burkle,T. Z. and Lee, J. H. (2007) "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners", *Journal of the Acoustical Society of America,* Vol. 122, No. 4, pp. 2365-2375

Kidd, G. Jr., Mason, C. R. and Rohtla, T. L. (1995), "Binaural advantage for sound pattern identification." *Journal of the Acoustical Society of America*, Vol. 98, No. 4, pp. 1977-1986

Kidd, G. Jr., Mason, C. R., Rohtla, T. L. and Deliwala, P. S. (1998), "Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns.", *Journal of the Acoustical Society of America,* Vol. 104, No. 1, pp. 422-431

Kistler, D. J. and Wightman, F. L. (1992), "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction", *Journal of the Acoustical Society of America,* Vol. 91, No. 3, pp. 1637-1647

Knapp, C and Carter, G. (1976), "The generalized correlation method for estimation of time delay", *IEEE Transactions on Acoustics, Speech and Signal Processing,* Vol. 24, No. 4, pp. 320-327

Köbler, S., Rosenhall, U. and Hansson, H. (2001), "Bilateral hearing aids – effects and consequences from a user perspective", *Scandinavian Audiology,* Vol. 30, No. 4, pp. 223-235

Kochkin, S. (2000), "MarkeTrak V: "why my hearing aids are in the drawer": the consumers' perspective", *The Hearing Journal*, Vol. 53, No. 2, pp. 34–42

Kohlrausch, A. (1990), "Binaural masking experiments using noise maskers with frequency-dependent interaural phase differences. I: Influence of signal and masker duration", *Journal of the Acoustical Society of America,* Vol. 88, No. 4, pp. 1737-1748

Kuhn, G. F. (1977), "Model for the interaural time differences in the azimuthal plane", *Journal of the Acoustical Society of America,* Vol. 62, No. 1, pp. 157-167

Kulkarni, A. and Colburn, H. S. (1995), "Infinite impulse response models of the head-related transfer function", *Journal of the Acoustical Society of America,* Vol. 97, p. 3278

Landone, C. and Sandler, M. B. (1998a), "Digital filtering for 3D binaural sound", *Digital filters: An enabling technology, IEE Colloquium on,* pp. 9/1-9/8

Landone, C. and Sandler, M. B. (1998b), "3-D sound systems: a computationally efficient binaural processor", *Audio and music technology: the challenge of creative DSP, IEE Colloquium on,* pp. 6/1-6/8

Li and Loizou, (2007) "Factors influencing glimpsing of speech in noise", *Journal of the Acoustical Society of America,* Vol. 122, No. 2, pp 1165-1172

Lindemann, W. (1986), "Extension of a binaural cross-correlation model by contralateral inhibition. I simulation of lateralisation of stationary signals", *Journal of the Acoustical Society of America*, Vol. 80, No. 6, pp. 1608-1622

Little, A. D., Mershon, D. H. and Cox, P. H. (1992), "Spectral content as a cue to perceived auditory distance", *Perception,* Vol. 21, No. 3, pp. 405-416

Lopez, J. J. and Gonzalez, A. (2001), "Experimental evaluation of cross-talk cancellation regarding loudspeakers' angle of listening", *IEEE Signal Processing Letters,* Vol. 8, No. 1, pp. 13-15

Macpherson, E. A. and Middlebrooks, J. C. (2002), "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited", *Journal of the Acoustical Society of America,* Vol. 111, No. 5, pp. 2219-2236

McAdams, S., Botte, M. C. and Drake, C. (1998), "Auditory continuity and loudness computation", *Journal of the Acoustical Society of America,* Vol. 103, No. 3, pp. 1580-1591

McFadden, D. (1966), "Masking-level differences with continuous and with burst masking noise", *Journal of the Acoustical Society of America,* Vol. 40, No. 6, pp.1414-1419

McFadden, D., Russell, W. E., Pulliam, K.A., (1972), "Monaural and binaural masking patterns for a low-frequency tone", *Journal of the Acoustical Society of America,* Vol. 51, No. 2 pt 2, pp. 534-543.

McIlwain, P. (2001), "Spatialised sound: the listener's perspective" *Proceedings of the 2001 ACMA conference, University of Western Sydney*, pp. 57-66

McNaughton, D., Fallon, K., Tod, J., Weiner, F. and Neisworth, J. (1994) "Effect of repeated listening experiences on the intelligibility of synthesized speech", *Augmentative and Alternative Communication*, Vol. 10, No. 3, pp.161-168

Middlebrooks, J. C. and Green, D. M. (1990), "Directional dependence of interaural envelope delays", *Journal of the Acoustical Society of America,* Vol. 87, No. 5, pp. 2149-2162

Middlebrooks, J. C., Makous, J. C. and Green, D. M. (1989), "Directional sensitivity of sound pressure levels in the human ear canal", *Journal of the Acoustical Society of America,* Vol. 86, No. 1, pp. 89-108

Middlebrooks, J. C. (1999), "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency", *Journal of the Acoustical Society of America,* Vol. 106, No. 3, pp. 1493-1510

Miller, G. A. and Licklider, J. C. R. (1950), "Intelligibility of interrupted speech", *Journal of the Acoustical Society of America*, Vol. 22, No. 2 , pp. 167-173

Minnaar, P., Oleson, S. K., Christiansen, F. and Moller, H. (2001a), "Localization with binaural recordings from artificial and human heads", *Journal of the Audio Engineering Society,* Vol. 49, No. 5, pp. 323-336

Møller, H. (1992), "Binaural auralization: Head-related transfer functions measured on human subjects" *Presented at the 93rd AES convention, October 1992*

Møller, H., Sørensen, M. F., Hammershøi, D. and Jensen C. B. (1995a), "Head-related transfer functions of human subjects" *Journal of the Audio Engineering Society,* Vol. 43, No. 5, pp. 300-321

Møller, H., Jensen, C. B., Hammershøi, D. and Sørenson, M. F. (1995b), "Design criteria for headphones", *Journal of the Audio Engineering Society,* Vol. 43, No. 4, pp. 218-232

Møller, H., Hammershøi, D., Jensen, C. B. and Sørensen, M.F. (1995c), "Transfer characteristics of headphones measured on human ears", *Journal of the Audio Engineering Society,* Vol. 43, No. 4, pp. 203-217

Møller, H., Sørensen, M. F., Jensen, C. B. and Hammershøi, D. (1996), "Binaural technique: do we need individual recordings?", *Journal of the Audio Engineering Society,* Vol. 44, No. 6, pp. 451-468

Møller, H., Hammershøi, D., Jensen, C. B. and Sørenson, M. F. (1999), "Evaluation of artificial heads in listening tests", *Journal of the Audio Engineering Society*, Vol. 47, No. 3, pp. 83-100

Moore, B. C. J. (1993), "Characterization of simultaneous, forward and backward masking", *AES 12th International Conference: The perception of reproduced sound,* pp. 22-33

Moore, B. C. J. and Glasberg, B. R. (2007), "Modelling binaural loudness", *Journal of the Acoustical Society of America,* Vol. 121, No. 3, pp. 1604-1612

Moore, G. E. (1965), "Cramming more components onto integrated circuits", *Electronics,* Vol. 38, No. 8, pp. 22-33

Musicant, A. and Butler, R. A. (1984), "The influence of pinnae-based spectral cues on sound localization", *Journal of the Acoustical Society of America,* Vol. 75, No. 4, pp. 1195-2000

Noble, W., Byrne, D. and Ter-Horst, K. (1997), "Auditory localisation, detection of spatial separateness, and speech hearing in noise by hearing impaired listeners", *Journal of the Acoustical Society of America,* Vol. 102, No. 4, pp. 2343-2352

Norris, J. W. (1998), "An adaptive filtering method to calculate HRTF", *Multimedia Signal Processing, 1998, IEEE Second Workshop on,* pp. 149-154

Oticon (2008a), "The audiology in Epoq – a whitepaper", *Oticon whitepaper*

Oticon (2008b), "Epoq product information", *Oticon datasheet*

Palomaki, K. J., Brown, G. J. and Wang, D. L. (2001) "A binaural auditory model for missing data recognition of speech in noisy and reverberant conditions", *In Proc. Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis (CRAC)*

Peissig, J. and Kollmeier, B. (1997), "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners", *Journal of the Acoustical Society of America,* Vol. 101, No. 3, pp. 1660-1670

Plack, C. J. and White, L. J. (2000), "Perceived continuity and pitch perception", *Journal of the Acoustical Society of America,* Vol. 108, No. 3, pp. 1162-1169

Pralong, D. and Carlile, S. (1994), “Measuring the human head related transfer functions: A novel method for construction and calibration of a miniature 'in-ear' recording system” *Journal of the Acoustical Society of America,* Vol. 95, No. 6, pp. 3435-3444

Pralong, D. and Carlile, S. (1996), “The role of individualized headphone calibration for the generation of high fidelity virtual auditory space”, *Journal of the Acoustical Society of America,* Vol. 100, No. 6, pp. 3785-3793

Raykar, V. C., Duraiswami, R. and Yegnanarayana, B. (2005), “Extracting the frequencies of the pinna spectral notches in measured head related impulse responses.”, *Journal of the Acoustical Society of America,* Vol. 118, No. 1, pp. 364-374

Recio A., Rhode, W.S., (2000), “Basilar membrane responses to broadband stimuli”, *Journal of the Acoustical Society of America,* Vol. 108, No. 5 pt. 1, pp. 2281 – 2298.

Rickard, S., (2007), “The DUET blind source separation algorithm”, *Blind Speech Separation,* Makino, S., Lee, T-W, and Sawada, H. Eds., pp. 217-241, Springer, 2007

RNID (2005), “Annual survey report 2005”, *The Royal National Institute for Deaf People, Registered charity numbers 207720 (England and Wales) and SC038926 (Scotland)*

RNID (2007), “Annual survey report 2007”, *The Royal National Institute for Deaf People, Registered charity numbers 207720 (England and Wales) and SC038926 (Scotland)*

Roman, N., Wang, D. L. and Brown, G. J. (2003) “Speech segregation based on sound location”, *Journal of the Acoustical Society of America,* Vol. 114, No. 4 pt. 1, pp. 2236 – 2252

Roy, O. and Vetterli, M. (2007) "Distributed spatial audio coding in wireless hearing aids", *JIEEE workshop on applications of signal processing to audio and acoustics (WASPAA),* pp. 227 – 230

Saberi, K., Dostal, L., Sadralobadai, T., Bull, V. and Perrott, D.R. (1991), "Free-field release from masking", *Journal of the Acoustical Society of America,* Vol. 90, No. 3, pp. 1355-1370

Samuel, A. G. (1981), "The role of bottom-up confirmation in the phonemic restoration illusion", *Journal of Experimental Psychology: Human Perception and Performance,* Vol. 7, No. 5, pp. 1124-1131

Shinn-Cunningham, B. G. (1998), "Applications of Virtual Auditory Displays", *Proceedings of the 20th international conference of the IEEE engineering in biology and medicine society,* Vol. 20, No. 3, pp. 1105-1108

Shinn-Cunningham, B. G., Santarelli, S. and Kopco, N. (2000a), "Tori of confusion: binaural localization cues for sources within reach of a listener", *Journal of the Acoustical Society of America,* Vol. 107, No. 3, pp. 1627-1636

Shinn-Cunningham, B. G. (2000b), "Distance cues for virtual auditory space", *Invited paper, special session on virtual auditory space, Proceedings of the first IEEE Pacific-rim conference on multimedia, 13-15 December 2000, Sydney, Australia,* pp. 227-230

Shinn-Cunningham, B. G., Schickler, J., Kopco, N. and Litovsky, R. (2001), "Spatial unmasking of nearby speech sources in a simulated anechoic environment", *Journal of the Acoustical Society of America,* Vol. 110, No. 2, pp. 1118-1129

Shinn-Cunngham, B. G. and Wang, D. L. (2008) "Influences of auditory object formation on phonemic restoration", *Journal of the Acoustical Society of America,* Vol. 123, No. 1, pp. 295-301

Shirley, B. G. and Kendrick, P. (2008) "Performance of independent component analysis when used to separate competing acoustic sources in anechoic and reverberant conditions", *Presented at the AES 124th Convention, Amsterdam, May 17-20, 2008*

Shoji, S., (2007) "Efficient individualisation of binaural audio signals", *PhD. thesis, Department of Electronics, University of York*

Sivonen, V. P. (2007), "Directional loudness and binaural summation for wideband and reverberant sounds", *Journal of the Acoustical Society of America,* Vol. 121, No. 5, pp. 2852-2861

Slaney, M. (1993), "An efficient implementation of the Patterson-Holdsworth auditory filter bank", *Apple Computer Technical Report #35*

Smith, M. W. and Faulkner, A. (2006), "Perceptual adaptation by normally hearing listeners to a simulated 'hole' in hearing", *Journal of the Acoustical Society of America,* Vol. 120, No. 6, pp. 4019-4030

Sorri, M., Luotonen, M. and Laitakari, K. (1984), "Use and nonuse of hearing aids". *British Journal of Audiology*, Vol. 18, pp. 169-172.

Spittle, G. A. (2008) "The applications and challenges of processing audio over Bluetooth", *Audio Engineering Society 23rd UK Conference 2008*, pp 17-1 - 17-4.

Srinivasan, S. and Wang, D. L. (2005), "A schema-based model for phonemic restoration", *Speech Communication,* Vol. 45, pp.63-87

Stone, M. A. and Moore, B. C. J. (1999), "Tolerable hearing aid delays I. Estimation of limits imposed by the auditory path alone using simulated hearing losses", *Ear Hear.,* Vol. 20, No.3, pp. 182-192

Talantzis, F., Ward, D. B. and Naylor, P.A. (2006), "Performance analysis of dynamic acoustic source separation in reverberant rooms", *IEEE Transactions on Audio, Speech and Language Processing,* Vol. 14, No.4, pp. 1378-1390

Tan, C.-J. and Gan, W.-S. (2000), "Direct concha excitation for the introduction of individualized hearing cues", *Journal of the Audio Engineering Society,* Vol. 48, No.7/8, pp. 642-653

Tao, Y., Tew, A. I. and Porter, A. J. (2003), "A study on head shape simplification using spherical harmonics for HRTF computation at low frequencies", *Presented at the 114$^{th}$ AES convention, Amsterdam, March 22-25, 2003*

Teranishi, R., and Shaw, E. A. G. (1968), "External acoustic models with simple geometry", *Journal of the Acoustical Society of America,* Vol. 44, No. 1, pp. 257-263

Toyama, M., Uchiyama, M. and Nomura, H. (1999), "Head related transfer function representation of directional sound for spatial acoustic events modeling", *Multimedia Signal Processing, 1999, IEEE 3rd Workshop on,* pp. 221-226

van de Par, S. and Kohlrausch, A. (1997), "A new approach to comparing binaural masking level differences at low and high frequencies", *Journal of the Acoustical Society of America,* Vol. 101, No. 3, pp. 1671-1680

van de Par, S. and Kohlrausch, A. (1998), "Comparison of monaural (CMR) and binaural (BMLD) masking release", *Journal of the Acoustical Society of America,* Vol. 103, No. 3, pp. 1573-1579

Wang, D. L. and Brown, G. J. (1999), "Separation of speech from interfering sounds based on oscillatory correlation", *IEEE Transactions on Neural Networks*, Vol. 10, No. 3, pp. 684-697

Warren, R. M. (1970), "Perceptual restoration of missing speech sounds", *Science,* Vol. 167, pp. 392-393

Warren, R. M. and Obusek, C. J. (1971), "Speech perception and phonemic restorations", *Perception and Psychophysics,* Vol. 9, pp. 358-362

Warren, R. M., Obusek, C. J. and Ackroff, J. M. (1972), "Auditory induction: Perceptual synthesis of absent sounds", *Science,* Vol. 176, pp. 1149-1151

Warren, R. M., and Bashford, J. A. (1976), "Auditory contralateral induction: An early stage in binaural processing", *Perception and Psychophysics,* Vol. 20, pp 380-386.

Warren, R. M. (1984), "Perceptual restoration of obliterated sounds", *Psychological Bulletin,* Vol. 96, No. 2, pp. 371-383

Warren, R. M., Bashford, J. A., Healy, E. W. and Brubaker, B. S. (1994), "Auditory induction: Reciprocal changes in alternating sounds", *Perception and Psychophysics*, Vol. 55, No. 3, pp. 313-322

Warren, R. M., Reiner, K. R., Bashford, J. A. and Brubaker, B. S. (1995), "Spectral redundancy - intelligibility of sentences heard through narrow spectral slits", *Perception and Psychophysics,* Vol. 57, No. 2, pp. 175-182

Warren, R. M., Hainsworth, K. R., Brubaker, B. S., Bashford, J. A. and Healy, E. W. (1997), "Spectral restoration of speech: intelligibility is increased by inserting noise in spectral gaps", *Perception and Psychophysics,* Vol. 59, No. 2, pp. 275-283

Wenzel, E. M., Arruda, M., Kistler, D. J. and Wightman F. L. (1993), "Localization using nonindividualised head-related transfer functions", *Journal of the Acoustical Society of America,* Vol. 94, No. 1, pp. 111-123

Wightman, F. L. and Kistler, D. J. (1989a), "Headphone simulation of free field listening. I Stimulus synthesis", *Journal of the Acoustical Society of America,* Vol. 85, No. 2, pp. 858-867

Wightman, F. L. and Kistler, D. J. (1989b), "Headphone simulation of free-field listening. II: Psychophysical validation", *Journal of the Acoustical Society of America,* Vol. 85, No. 2, pp. 868-878

Wightman, F. L. and Kistler, D. J. (1991), "Localization of virtual sound sources synthesised from model HRTFs", *Applications of signal processing to audio and acoustics, 1991, Final program and paper summaries, IEEE ASSP Workshop on,* pp.0_51-0_52

Wightman, F. L. and Kistler, D. J. (1992), "The dominant role of low-frequency interaural time differences in sound localization", *Journal of the Acoustical Society of America,* Vol. 91, No. 3, pp. 1648-1661

Woodworth, R. S. and Schlosberg, G. (1962) "Experimental Psychology", *Holt, Rinehard and Winston, New York*, pp. 349-361

Wright, D., Hebrank, J. H. and Wilson, B., (1974) "Pinna reflections as cues for localization", *Journal of the Acoustical Society of America,* Vol. 56, No. 3, pp. 957-962

Zotkin, D. N., Duraiswami, R., Grassi, E. and Gumerov, N. A. (2006) “Fast head related transfer function measurement via reciprocity”, *Journal of the Acoustical Society of America,* Vol. 120, No. 4, pp. 2202-2215.

Zurek, P.M. and Durlach, N. I. (1987), “Masker bandwidth dependence in homophasic and antiphasic tone detection”, *Journal of the Acoustical Society of America,* Vol. 81, No. 2, pp. 459-464.

Zurek, P. M. (1993), “A note on onset effects in binaural hearing”, *Journal of the Acoustical Society of America,* Vol. 93, No. 2, pp. 1200-1201

# Bibliography


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

Bourgeois, J. and Minker, W. (2009), "Time-domain beamforming and blind source separation. Series: Lecture Notes in Electrical Enginering, Vol. 3." *Springer*

Bregman, A. S. (1990), "Auditory scene analysis." *Cambridge, MA: MIT Press*

Hyvärinen, A., Karhunen, J. and Oja, E., (2001), "Independent Component Analysis", *New York: Wiley*

Ifeachor, E. C. and Jervis, B. W. (2002), "Digital signal processing a practical approach", *Harlow: Pearson Education Limited, Addison-Wesley (second edition)*

Moore, B. C. J. (1997), "An introduction to the psychology of hearing", *London: Academic Press (fourth edition)*

Oxford Dictionary, (1995), "The concise Oxford dictionary" *Oxford: Clarendon Press (ninth edition)*

Rabiner, L. R. and Gold, B. (1975), "Theory and application of digital signal processing" *Prentice Hall*

Rossing, T. D. (1990) "The science of sound", *Addison-Wesley (second edition)*

Warren, R. M. (2008), "Auditory perception", *Cambridge: Cambridge University Press (third edition)*