The Use of Optimal Cue Mapping to Improve the Intelligibility and Quality of Speech in Complex Binaural Sound Mixtures

Jingbo Gao

Ph.D

University of York
Electronics

January 2016

Abstract

A person with normal hearing has the ability to follow a particular conversation of interest in a noisy and reverberant environment, whilst simultaneously ignoring the interfering sounds. This task often becomes more challenging for individuals with a hearing impairment. Attending selectively to a sound source is difficult to replicate in machines, including devices such as hearing aids. A correctly set up hearing aid will work well in quiet conditions, but its performance may deteriorate seriously in the presence of competing sounds. To be of help in these more challenging situations the hearing aid should be able to segregate the desired sound source from any other, unwanted sounds.

This thesis explores a novel approach to speech segregation based on optimal cue mapping (OCM). OCM is a signal processing method for segregating a sound source based on spatial and other cues extracted from the binaural mixture of sounds arriving at a listener's ears. The spectral energy fraction of the target speech source in the mixture is estimated frame-by-frame using artificial neural networks (ANNs). The resulting target speech magnitude estimates for the left and right channels are combined with the corresponding original phase spectra to produce the final binaural output signal. The performance improvements delivered by the OCM algorithm are evaluated using the STOI and PESQ metrics for speech intelligibility and quality, respectively. A variety of increasingly challenging binaural mixtures are synthesised involving up to five spatially separate sound sources in both anechoic and reverberant environments. The segregated speech consistently exhibits gains in intelligibility and quality and compares favourably with a leading, somewhat more complex approach. The OCM method allows the selection and integration of multiple cues to be optimised and provides scalable performance benefits to suit the available computational resources. The ability to determine the varying relative importance of each cue in different acoustic conditions is expected to facilitate computationally efficient solutions suitable for use in a hearing aid, allowing the aid to operate effectively in a range of typical acoustic environments. Further developments are proposed to achieve this overall goal.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor Tony Tew for his guidance and support throughout my Ph.D journey. We first became acquainted in August 2011. To be honest, he is not only my supervisor, but also my friend. I respect him and I feel he has been like a parent to me in the UK. I am grateful that he accepted me to be his student and taught me how to be a researcher. I am so lucky to have also had the opportunity to get to know his family. He has played a very important role in my life in the UK and I will always remember him and his kindness.

I would like to thank Professor DeLiang Wang and Dr. Yi Jiang for kindly providing the Deep Neural Network binaural classification speech segregation algorithm (Jiang et al., 2014). They helped me greatly when I was reproducing their algorithm.

I would like to give special thanks to Dr. John Szymanski for his guidance and support in my TAP meetings. I would like to thank my flatmate, Dimitrios Zantalis, for the many helpful discussions and encouragement at home. I would like to thank Yunfeng Ma for all his help during my stay in the UK. Thank you too to Yuan Wang for the times we worked together in the University library.

I also would like to thank all my colleagues in the Audio Lab. Special thanks go to Andrew Chadwick for providing me with guidance on how to use all the equipment I needed from the Lab. Thanks to Jiajun Yang for the good times we had in the office.

Finally, my greatest thanks go to my parents. I thank them for their love and support. I would not have been able to study in the UK without their full support.

Declaration

I, Jingbo Gao, declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References. Methods and partial results of the pilot study (Chapter 6) have been published in the following papers:

- Gao, J. & Tew, A. I (2015). The segregation of spatialised speech in interference by optimal mapping of diverse cues. In proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. 19 Apr 2015, Brisbane. p. 2095-2099.

Chapter 1

Introduction

1.1 Background

Speech is a powerful communications bridge between humans, though it can suffer from sources of acoustic interference which increase the effort required to listen to speech and can reduce its intelligibility. The problem becomes more serious for people suffering from a hearing deficit. According to a recent study by Action On Hearing Loss, there were approximately 10 million people in 2015 with a hearing loss in the UK (RNID, 2015), which means that about one person in six has some degree of hearing impairment.

Hearing loss which occurs at birth is known as congenital hearing loss. Acquired hearing loss after birth usually develops gradually, but can be sudden. Hearing loss can be indicated by many signs, most obviously by experiencing difficulty with hearing what people are saying or requiring higher sound levels than others when listening to music or watching television. There are a multitude of reasons why sounds may not be processed successfully and these can lie at any point within the auditory system.

Hearing loss can loosely be divided into two types: conductive and sensorineural. Problems of the outer ear or middle ear result in conductive hearing loss. This is often caused by a blockage (e.g. due to earwax or fluid arising from an ear infection) or a structural problem (such as a disorder of the hearing bones or a perforated eardrum).

Sensorineural hearing loss may be caused by damage to the inner ear or the auditory nerve. It often results from defective outer hair cells or damage to the cochlea resulting from genetic disorders, injury, ageing or the cumulative effect of loud noise. It is possible to suffer from both conductive and sensorineural hearing loss and this is described as a mixed hearing loss. Hearing impairment can vary greatly in its severity, ranging from slight to profound deafness.

For many decades hearing aids have been the primary choice for alleviating the effects of a hearing deficit. A person with sensorineural hearing impairment may have less ability to hear a weak sound and less tolerance of an intense sound (a phenomenon known as recruitment) than a normal-hearing person. Hearing aids with electronic compression can compensate for this deficit. In general, in order to communicate effectively, suffers from sensorineural hearing impairment require a higher signal-to-noise ratio than normal-hearing people.

In the 1930s, wearable electric hearing aids began to be developed. These were very large and were worn somewhere on the body and so were referred to as body aids. The advent of transistor technology in the 1960s saw a significant reduction in their size and led to the development of behind-the-ear (BTE) devices. The size of hearing aids continued to reduce, leading to the appearance of completely-in-the-canal (CIC) hearing aids in the 1980s (Hearing Aids, 2015).

The small size of hearing aids is not the only challenge associated with them. One of the biggest challenges and opportunities arrived with the introduction of digital technology in hearing aids in the 1980s (e.g. Graupe et al. (1986)). Some operations, such as block processing to represent a signal with fine resolution in the frequency domain, cannot realistically be implemented in analogue aids.

The switch to digital signal processing (DSP) in hearing aid technology has had a major impact. Digital hearing aids have the ability to make decisions about how to process sounds depending on the acoustic environment (Dillon, 2001). DSP algorithms can selectively reduce the amplitude of interfering sounds and improve the ability of the user to understand a conversation. An algorithm by Tellakula (2007), has been used to improve signal-to-noise ratio in high-end hearing aids (Widex Inc., 2015). In the last decade, hearing aids have largely completed the transition from analogue to digital technology. Great strides have been made, not only in terms of more and more sophisticated signal processing algorithms, but also with greater miniaturisation and power efficiency. The latter issue in particular has continually held back digital hearing aid development as computational complexity and power consumption go hand-in-hand.

Digital hearing aids can be categorised as monaural or bilateral. Monaural hearing aids are worn in only one ear, whereas bilateral aids require users to wear a pair of monaural hearing aids, one in each ear. Conventionally, each monaural hearing aid in a bilateral pair processes the sound individually using its own independent algorithm, without sharing any information between the two. Modern binaural hearing aids are also fitted bilaterally, but they are able to communicate with each other and share binaural information. This gives them the ability to process sounds in a more sophisticated and effective manner.

Although modern hearing aids can improve the intelligibility of speech and preserve sound quality, they tend to work well only in relatively quiet environments. Noisy or reverberant environments can seriously affect their performance (Kochkin, 2000; RNID, 2007) to the extent that 13 % of people interviewed by Gimsing (2008) had stopped using their hearing aids after five years of having them fitted, predominantly due to unpleasant sound quality or a lack of benefit. Two years later, Kochkin (2010) reported that the number of “satisfied” or “very satisfied” customers had not improved, with lack of benefit and poor sound quality remaining the key reasons cited for this.

It is clear that a more advanced hearing aid is needed that is able to provide users with a better experience in terms of greater speech intelligibility and higher speech quality in adverse, varying conditions. Similar issues exist in the field of automatic speech recognition (ASR), despite the advances in such systems in the last few years. Once again, however, recognition rates are high for speech in a relatively quiet environment with a microphone close to the talker, but performance markedly deteriorates in the presence of interfering sounds, especially competing speech (Narayanan and Wang, 2014).

1.2 Motivation

In everyday life, humans with normal hearing display a remarkable ability to attend selectively to a single sound source in the presence of competing concurrent sources, background noise and reverberation. This phenomenon can be demonstrated, for example, in the case of holding a conversation with someone in noisy surroundings, where it is often possible to pay attention to the target speech of interest without being affected by other nearby concurrent conversations or sounds. This selective attention ability of the human hearing system has been referred to as the cocktail party effect (Cherry, 1953). The effect applies widely in daily life, because the sound waves reaching our ears generally do not emanate from a single sound source but, rather, come from a combination of sound sources, many of which can be considered to be unwanted. In addition, listeners are able to adapt to and tolerate moderate reverberation in an acoustic environment (Nábělek and Robinson, 1982). Humans also have the ability to localise the direction of a speech source in a mixture of competing sound sources by virtue of having two ears and by using head movements (Rayleigh, 1907). By utilising both ears, human listeners can attend to a desired sound signal under extremely adverse conditions and binaural hearing compared to monaural leads to significant improvements in speech intelligibility in noisy environments (Moore, 2012).

Binaural hearing makes it possible for a person to locate a directional sound source with greater ease than when using a single ear (Harris, 1965) and it increases speech intelligibility in noisy environments. These observations inspired the development of binaural digital hearing aids. Recently, Thakur et al. (2015) demonstrated that it is possible to implement in hardware a sound segregation algorithm in real time. He developed a framework for the algorithm that emulates the human ability of selective attention to a single sound source using a field programmable gate array (FPGA). With todays technology it is also feasible to share information between and with two bilateral devices via a wireless link, as demonstrated by the integration of Bluetooth technology (Bluetooth Technology Website, 2015) in hearing aids (e.g. Oticon (2008)). Such technologies are fundamental to the construction of binaural hearing aids.

One form of signal processing applicable to a digital hearing aid is sound segregation, in which elements of a sound source (such as speech) are extracted from an acoustic mixture to form an auditory stream (Wang and Brown, 2006). Segregation can, for example, be performed on the basis that the direction of the wanted sound source is different from the directions of the interfering sources. Based on the premise that extracting a clean version of the wanted speech from a binaural mixture improves both the intelligibility and the quality of the wanted speech, an ideal goal for a hearing aid is to achieve perfect source segregation. The focus of the work in this thesis is on developing a binaural digital hearing aid algorithm which exploits segregation cues to advance a significant distance closer to this goal.

Machine learning is a branch of artificial intelligence which focuses on the study and construction of algorithms that can learn from data and make predictions. In recent years, there has been a surge of interest in researching machine learning and applying it to speech processing. Artificial neural networks (ANNs) are a popular example of a machine learning model inspired by the biological neural networks found in the central nervous systems of animals. In the training stage of the ANN, when both the clean speech and the binaural mixture are accessible, speech segregation can be considered as a supervised learning problem. A time-frequency mask is commonly used in the segregation process. The computational goal of machine learning is to estimate the ideal mask, such that the features of the target speech may be extracted from the noisy speech mixture. The mask commonly has two forms: a binary mask (IBM) or a ratio mask (IRM). In this application, the speech segregation problem is formulated as a classification problem when estimating the IBM and a mapping problem when estimating the IRM.

1.3 Objectives

Typical auditory environments in which hearing aids are used include situations with multiple concurrent sound sources. The sound from these sources may be reflected and diffracted by surfaces both near and far and the location of each source may be constantly changing. In practice, the interfering sounds themselves will exhibit a range of characteristics. For instance, they may be stochastic or periodic in nature or a mixture of the two. In reverberation, the reflected sound may be highly correlated with the original source, whereas the target speech and interfering sources may be relatively uncorrelated, such as in the case of multiple talkers. These variations may create conflicting requirements when attempting to segregate the target sound source.

Most efforts at segregation of a speech source involve systems that try to tackle the cocktail party problem. To be useful in a hearing aid, such a system must be capable of segregating sound sources in real-time with sufficiently low delay in the acoustic output from the device. Visual cues contribute to speech intelligibility in noise (Sumby and Pollack, 1954) and for this to be effective a latency of under 160 ms in the acoustic signal reaching the listener has been shown to be adequate (Grant and Greenberg, 2001). However, the delay should generally be under 10 ms for hearing aids, to avoid the user hearing a delayed version of their own voice (Stone and Moore, 1999). For this reason, and because of power consumption constraints, which are due to battery size limitations, the algorithm must be relatively computationally simple.

Not only does the proposed algorithm have the potential to help someone with a hearing impairment to focus on wanted speech better in noise, the same can be expected to apply to users with normal hearing. There are other potential applications besides hearing aids for the proposed work, including improving the robustness of portable speech recognition devices and sound-based human-computer interaction systems, and hearing for robots.

The research presented in this thesis is aimed principally at improving the performance of digital hearing aids. The work is motivated principally by the desire to raise the intelligibility of target speech by exploiting some of the attributes of human binaural hearing when interfering sounds are present. The primary mechanism by which this will be achieved is speech segregation. Based on this, we develop and evaluate a scheme for reducing the level of competing sounds in a binaural mixture of sound signals using machine learning principles, thereby reducing the degree of masking being imposed on the wanted sound source.

1.4 Hypothesis

It is hypothesised that:

The intelligibility and quality of a target speech source in a binaural mixture with spatially distinct competing concurrent interfering sounds may be increased using a machine learning algorithm which is suitable for implementation in a hearing aid.

Associated with the primary hypothesis, there are several supplementary research questions:

1) *Does optimal mapping of an increasing number of diverse cues improve the segregation of one sound source in a binaural mixture in terms of intelligibility and quality?*

2) *Does the use of a ratio mask estimated by a neural network, compared with the equivalent binary mask, improve segregation of one sound source in a binaural mixture in terms of intelligibility and quality?*

3) *Is it possible to determine the varying relative contribution of diverse cues for estimating a mask in a range of simulated multiple-source and reverberant acoustic conditions?*

4) *Is it possible to allow maximum benefit to be drawn from limited computational resources by configuring the optimal mapping of cues?*

All these questions are addressed in this thesis in order to validate the hypothesis. Section 9.2, within the conclusion at the end of the thesis, returns to these questions and reviews the extent to which they have been answered.

1.5 Thesis structure

The thesis is organised into nine chapters. Chapter 2 presents a brief overview of the human hearing system. In particular, it introduces the peripheral auditory system and discusses the psychology of the perception of sound, including masking and auditory grouping. Chapter 3 contains an overview of binaural hearing, beginning with consideration of sound localisation for a single source, the effect on localisation of multiple sources and the advantages afforded by binaural hearing. Head-related transfer functions, sound spatialisation and binaural synthesis are introduced as fundamental prerequisites for artificially spatialising sound sources. Chapter 4 introduces speech intelligibility and quality assessment metrics. It briefly reviews existing techniques for improving speech intelligibility and speech quality in noise.

In Chapter 5, a new binaural processing method for improving speech intelligibility and quality in interference is described. Firstly, the concept of a time-frequency mask is defined and means for extracting a variety of spatial cues from a multi-source binaural mixture are described. This leads to the description of a novel algorithm called optimal cue mapping (OCM). OCM utilises extracted cues to estimate a spectral ratio mask which is applied frame-by-frame to the binaural mixture to segregate the wanted speech. A pilot study of OCM is described in Chapter 6. It starts with a simple fixed-direction, three-source anechoic configuration. The contribution of each type of cue is investigated and the performance of each mask is evaluated. With the aim of applying OCM in increasingly realistic situations, a more complicated problem space with variable-direction interferers in both anechoic and reverberant conditions is established in Chapter 7. The OCM algorithm is compared in Chapter 8 with another state-of-the-art binaural segregation system for a variety of configurations.

The results of the research are summarised in Chapter 9, and suggestions for further work are presented.

Chapter 2

Human Hearing and Sound Perception

In order to design a successful hearing aid algorithm, an appreciation of the human hearing system is needed. This chapter begins by briefly describing the outer, middle and inner ears, which are the major physiological components of the peripheral auditory system. Attention is then focused on masking, which is of special relevance when listening to speech in a complex binaural soundfield. The results of applying a mask (the spectral energy fraction of the target source in the mixture) in the segregation process generally leads to a fragmented version of the original target speech and so the final part of this chapter considers some higher functions of hearing using the powerful conceptual framework of auditory scene analysis to understand how these fragments are given meaning.

2.1 The peripheral auditory system

The human auditory system can be split into three parts: the outer ear, middle ear and inner ear, as shown in figure 2.1. The outer ear is itself principally

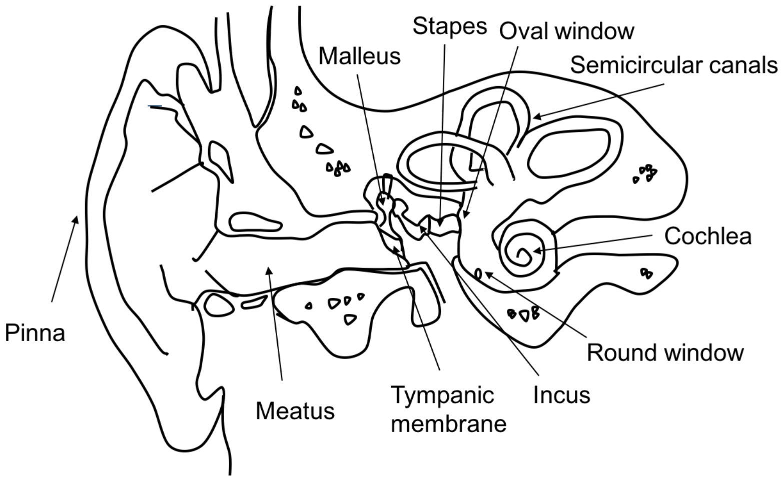


Figure 2.1: The peripheral auditory system. Redrawn from Moore (2012).

composed of three further parts. The first of these is the pinna, also known as the auricle, which is an irregularly shaped organ that comprises complex folds of cartilage. The pinna acts as an acoustic filter whose spectral characteristics for sounds entering the ear canal vary depending on the location of the sound source. The set of cues forms a unique fingerprint for each possible direction, which provides information to help the hearing system localise a sound and attend to it in the presence of other interfering sounds. This will be considered further in section 3.1.4. Furthermore, the size of the pinna and its detailed shape differ between individuals, which means that the filtering effects it produces are unique to each listener.

The second part of the outer ear is the meatus or ear canal. It is a tube which directs the sound from the pinna towards the tympanic membrane (eardrum) which is the third major part of the outer ear. The ear canal has an average length of 25 mm and a diameter of between 7 mm and 10 mm (Moore, 2012). The eardrum sits at an angle of approximately 55° to the ear canal. It is elliptical, with horizontal and vertical diameters of 8-9 mm and 9-10 mm, respectively. Its average thickness is approximately 0.074 mm.

Sound travels down the auditory canal and causes the eardrum to vibrate. The vibrations are transmitted through the bones of the middle ear (the

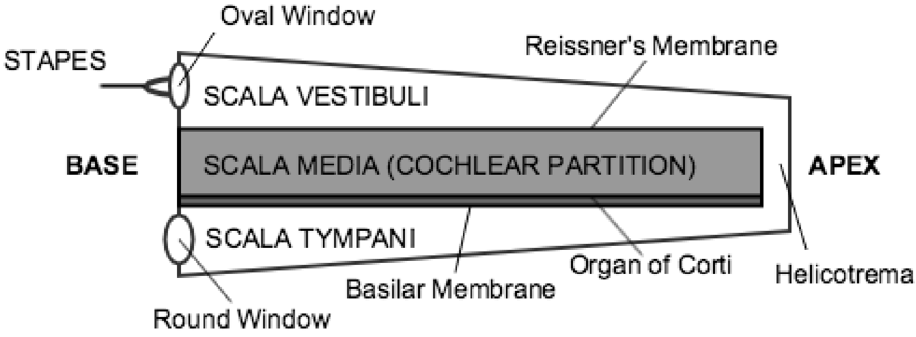


Figure 2.2: Schematic view of unrolled cochlea. Adapted from Howard and Angus (2009).

malleus, incus and stapes) to the oval window. Collectively known as the ossicles, the most important function of the middle ear is to convert air vibrations in the ear canal to fluid movement inside the cochlea. The shape and arrangement of the bones serve to match the acoustic impedance of the air medium to the fluid medium. The stapes presses onto the oval window and its movements are transmitted to the fluid in the cochlea. A second membrane-covered opening exists, known as the round window. When the oval window moves inwards, the round window moves outwards, and vice versa. The transfer of acoustic energy depends on the difference between the sound pressure applied on the oval window and that applied on the round window.

The inner ear contains the cochlea and auditory nerve. The cochlea is a fluid-filled spiral which consists of two to three turns. It has two chambers, the scala vestibuli and scala tympani, running along its length of 35 mm. The two chambers are separated by Reissners membrane and the basilar membrane (BM), and they are connected through the helicotrema at the apex (or apical) end of the cochlea. The organ of corti runs along the basilar membrane and is coated with around 12,000 outer hair cells and 3,500 inner hair cells. The base end of the cochlea is where the two previously mentioned windows, the oval window and the round window, are located. A schematic view of an unrolled cochlea is shown in figure 2.2.

The properties of the basilar membrane vary considerably from base to apex, and its response is affected differently by sounds of different frequencies. At the base, the basilar membrane is stiff and narrow, while it is wider and less stiff at the apex. As a result, high-frequency sounds cause maximum displacement of the basilar membrane near the base and low-frequency sounds have maximum effect near the apex.

When the stapes presses against the oval window, a pressure difference is applied across the basilar membrane. This is transmitted along the length of the basilar membrane, causing the round window to move outwards. The wave motion that occurs on the basilar membrane in response to a sound stimulus is referred to as a travelling wave. The wave begins at the oval window, rises to a peak where the basilar membrane is most sensitive to the frequency of the sound stimulus, and fades away as the energy is absorbed towards the helicotrema. Each point on the basilar membrane moves up and down sinusoidally. The inner hair cells convert the physical vibration of the basilar membrane into neural activity, which is then carried to the brain via the auditory nerve.

Two popular theories exist to explain the pitch perception of sound (see, for example, Howard and Angus (2009)), namely the place theory and the temporal theory. The place theory states that pitch perception depends on which regions of the basilar membrane are vibrating. The sensation of different frequency pitches derives exclusively from the motion of particular groups of hair cells. According to the temporal theory, pitch is coded by the precise timings of nerve cell firing, which are synchronised with the phase of an incoming periodic sound. Neither theory fully or properly explains pitch perception and each is applicable under different conditions.

2.2 Critical bands and masking

The human hearing system responds to sound frequencies ranging approximately from 20 Hz to 20 kHz and it can detect pressure variations between 20 μPa and 20 Pa. The term "absolute threshold of hearing" is applied to indicate the minimum audible sound level of a pure tone for an average ear with normal hearing. Therefore, to describe the pressure, the lowest perceptible sound pressure level (SPL), p_0, is used as a reference. The sound pressure level in dBs, L_{SPL}, is then given by the equation:

$$L_{SPL} = 20log_{10}\frac{p}{p_0} \qquad (\mathrm{dB_{SPL}}) \tag{2.1}$$

where p denotes the sound pressure of the stimulus and p_0 refers to the sound level of 20 μPa or $2 \times 10^{-5}\,\mathrm{N/m^2}$.

The sensitivity of human hearing to different frequencies is not uniform, which means the minimum threshold of hearing is a function of frequency and it also varies from person-to-person. It is affected by age, gender and social factors (Davis, 1995). There are two different methods for measuring the threshold of hearing (or absolute sensitivity). The first is the minimum audible pressure (MAP) (Killion, 1978). MAP is measured by delivering a sound to a listener's ear, usually over a headphone, and then asking them to indicate the minimum audible level. The absolute intensity is then obtained by placing a probe microphone inside the ear canal. The second method is the minimum audible field (MAF) (Robinson and Dadson, 1956). In this measurement, the sound is delivered through a loudspeaker rather than a headphone. It requires a free field environment, such as an anechoic chamber, which is free from reflections and diffractions caused by obstacles. For reference purposes, the sound pressure level is also measured without the listener, at the position corresponding to the centre of their head.

Examples of the results obtained using each method are shown in figure 2.3. It can be seen that they yield different results. The difference above

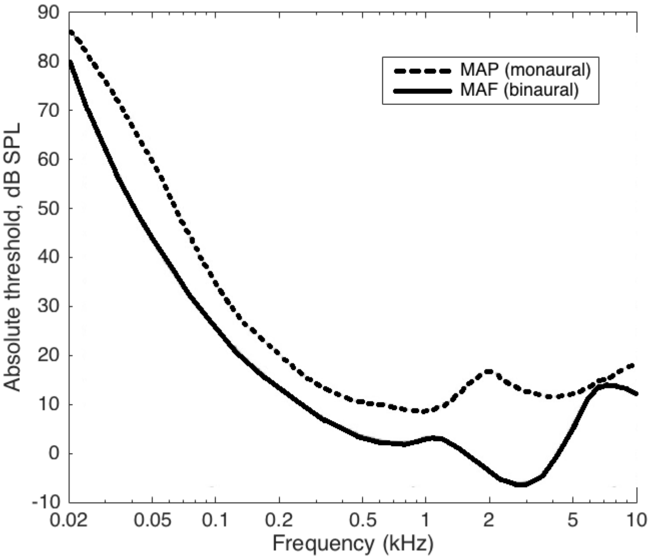


Figure 2.3: Minimum thresholds of hearing. Redrawn from Moore (2012).

1 kHz is attributed by Moore (2012) to physiological noise caused by the vascular system when wearing the headphone. Most significantly, the figure shows that the human hearing system is most sensitive at frequencies between 3 and 4 kHz.

2.2.1 Critical bands

Both MAP and MAF are measured in a quiet environment. The absolute threshold of human hearing can, however, be raised in the presence of other sounds or noise (Greenwood, 1961). For example, a listener may fail to detect a small amplitude tone when noise with relatively greater energy is presented simultaneously. This phenomenon is known as masking. In addition, a sound source is said to be masked when components of it cannot be heard due to the presence of another sound. The sound source that is masked is called the maskee, and the masking sound is known as the masker.

To investigate masking, Fletcher (1940) carried out an experiment to measure the threshold for detecting a sinusoidal signal as a function of the bandwidth of a band-limited noise. The noise had constant power density and its centre frequency was set equal to the sinusoidal signal frequency. The results of the experiment indicate that the hearing threshold for a tone increases as the noise bandwidth increases until it reaches a certain relatively stable value. Increasing the bandwidth of the noise further does not affect the threshold appreciably.

This behaviour can be accounted for in the peripheral auditory system if the cochlea acts like a bank of bandpass filters with overlapping passbands. These filters are termed critical bands or auditory filters. When a listener is presented with a tone and a band-limited noise, the bandpass filter with matching centre frequency passes the tone and the noise within that critical band and this creates the masking effect. Any noise outside the critical bandwidth plays little part in the masking process, creating the approximately constant threshold value observed when the noise bandwidth exceeds the bandwidth of the auditory filter.

The notched-noise method described by Patterson (1976) may be used to infer a shape for the auditory filters. In this method, a tone of frequency f is chosen, and masker noise with a stopband bandwidth $2\Delta f$ and centred at f is applied. This is illustrated in figure 2.4. The noise passing through the auditory filter is represented by the shaded regions. As the stop-band of the noise gets wider, the amount of noise decreases and this results in a lowering of the threshold of perception. Based on this, the shape of the auditory filter can be estimated and appears as a convex curve in figure 2.5. The critical bands are spread continuously across the spectrum. There is no evidence to indicate any discontinuities between them.

The critical bandwidth of auditory filters does not vary uniformly with frequency. The bandwidth is narrower at low frequencies and becomes wider as the frequency is increased. Data from subjective measurements of critical bandwidth and centre frequency by Zwicker and Terhardt (1980) are shown in table 2.1. The lowest frequency range here is from 0 Hz to 100 Hz, which

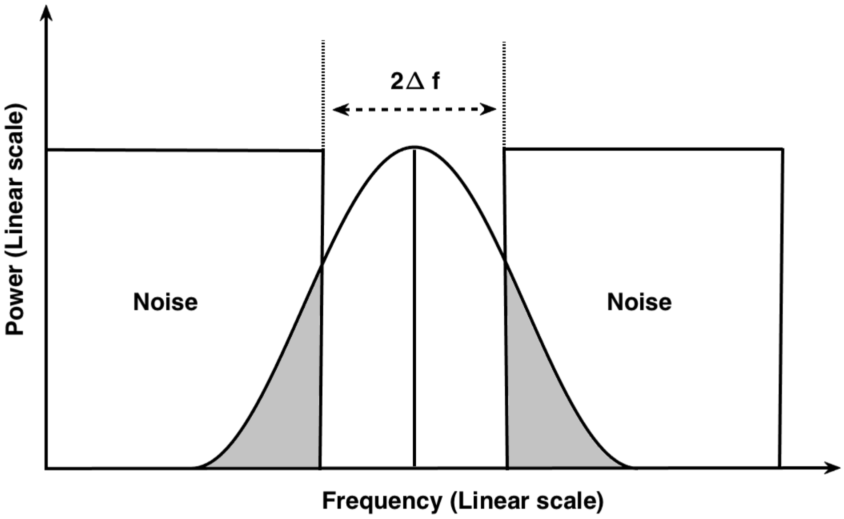


Figure 2.4: Schematic illustration of the notched-noise method. Redrawn from Patterson (1976).

includes the inaudible frequencies between 0 Hz to 20 Hz.

The shape of the auditory filters is complex and so they are often described in terms of their equivalent rectangular bandwidths (ERBs). The ERB is the bandwidth of a brickwall bandpass filter required to pass the same noise energy as the auditory filter it represents. The ERB of an auditory filter is intended to represent its bandwidth but not its shape.

A formula that estimates the ERB in Hz as a function of centre frequency has been developed by Glasberg and Moore (1990) and is given by:

$$ERB = 24.7(4.37fc + 1) \tag{2.2}$$

where fc is the centre frequency in kHz.

Table 2.1: The values of critical band rate and critical bandwidth, adapted from Zwicker and Terhardt (1980).

Range (lower Hz - upper Hz)	Critical bandwidth (Hz)	Centre frequency(Hz)
0-100	100	50
100-200	100	150
200-300	100	250
300-400	100	350
400-510	110	450
510-630	120	570
630-770	140	700
770-920	150	840
920-1080	160	1000
1080-1270	190	1170
1270-1480	210	1370
1480-1720	240	1600
1720-2000	280	1850
2000-2320	320	2150
2320-2700	380	2500
2700-3150	450	2900
3150-3700	550	3400
3700-4400	700	4000
4400-5300	900	4800
5300-6400	1100	5800
6400-7700	1300	7000
7700-9500	1800	8500
9500-12000	2500	10500
12000-15500	3500	13500

2.2.2 Masking

In the previous section, the concept of masking was introduced and was shown to occur when the perception of one sound is affected by the presence of noise which is spectrally close to it. This type of masking in the frequency domain is known as simultaneous masking, since maskee and masker are presented at the same time. Masking also occurs when maskee and masker do not coincide temporally, which is known as non-simultaneous masking. In the following sections, both types of masking are discussed.

2.2.2.1 Simultaneous masking

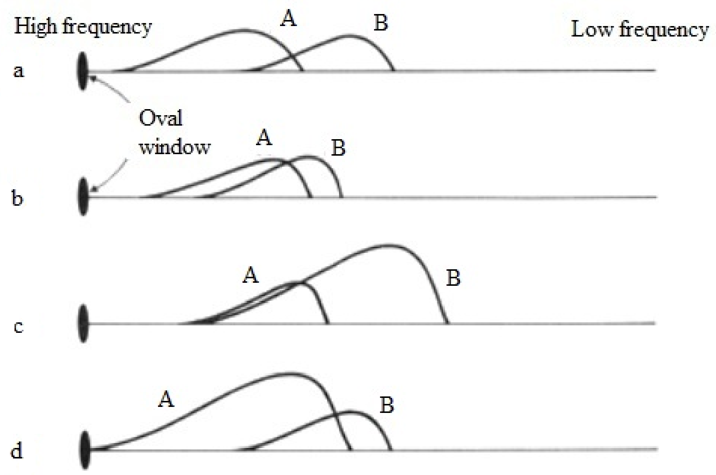


Figure 2.5: Basilar membrane displacement for tone A and B, tone B has a lower frequency than tone A. (a) two tones are barely overlapped, (b) two tones are masked by each other, (c) tone A is almost fully masked by tone B, (d) tone A partially masks tone B. Redrawn from Rossing (1990).

When a sound arrives at the cochlea, it vibrates and excites the hair cells along the basilar membrane, as described in section 2.1. For a sine wave, there is a position of maximum displacement on the basilar membrane and this position depends on the frequency of the tone. Neighbouring regions of the basilar membrane are also displaced, but to a lesser degree. If another tone with a slightly different frequency is active simultaneously, it is only perceptible if its energy is sufficient to displace the basilar membrane with a greater amplitude than the original tone does at the position of peak response for this new frequency. Hence, a low energy signal can be hidden under the region of displacement caused by the original sine wave. This is illustrated in figure 2.5.

The masking threshold (that is the boundary between perceptible and imperceptible) can be measured by presenting a pure tone masker signal at a certain frequency with a fixed intensity and a maskee with a variable frequency and intensity. The graphical representation of masked thresholds as a function of frequency is known as a masking pattern or masked audiogram.

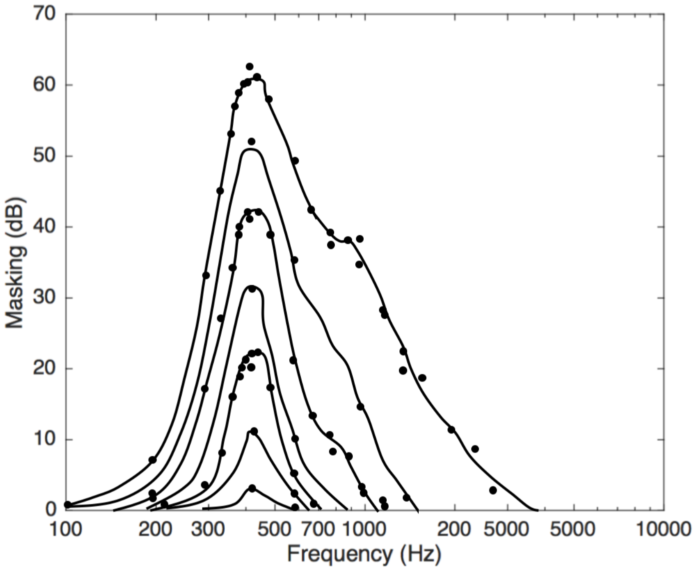


Figure 2.6: Masking pattern for a narrow band of noise with centre frequency of 410 Hz. The overall noise level is indicated above each curve which shows the elevation in threshold of a pure tone signal as a function of frequency. Redrawn from Moore (2012), original from Egan and Hake (1950).

Figure 2.6 shows an example of a masking pattern for narrowband noise with a centre frequency of 410 Hz. For a low level masker (20 and 30 dB SPL), the threshold exhibits a symmetrical shape. As the masker level increases, however, the masking curve becomes wider and the shape becomes asymmetric. As can be seen in figure 2.5, higher frequencies are more effectively masked than lower frequencies. This is referred to as the upward spread of masking. It happens because a low frequency sound has to propagate a longer distance along the basilar membrane compared to a high frequency one. This causes displacement of higher frequency points on the basilar membrane while the low frequency sound is propagating past them.

2.2.2.2 Non-simultaneous masking

Masking not only occurs when two signals are presented simultaneously, but also when they are presented in succession, as shown in figure 2.7. The louder sound causes the other sound before or after it to become imperceptible. This type of masking is known as non-simultaneous masking or temporal masking. It can be subdivided into two types: forward masking, which describes the case where masking occurs just after the presence of the masker, and backward masking, which occurs when the maskee precedes the masker.

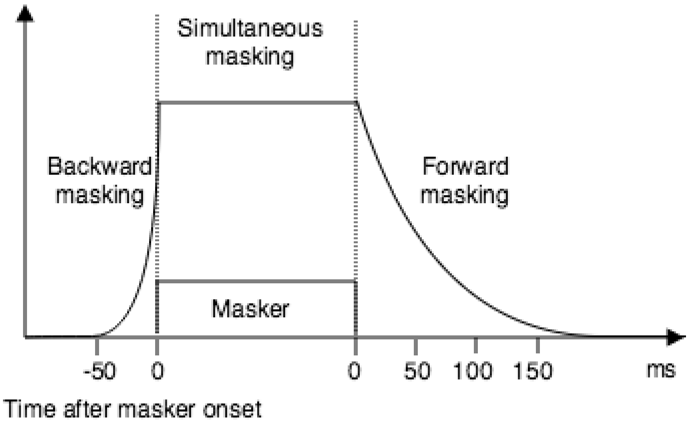


Figure 2.7: Schematic drawing of the region of non-simultaneous masking. Redrawn from Fastl and Zwicker (2007).

The masking threshold reduces to 0 dB 100-200 ms after the end of the masker. The increase in forward masking is not straightforwardly related to the masker level. When the masker level increases by 10 dB, for example, there may be only a 3 dB increase in the masking threshold. Furthermore, the amount of masking increases as the duration of the masker is increased at least up to 50 ms (Moore, 2012). Forward masking is also influenced by the relationship between the frequencies of maskee and masker. The occurrence of forward masking is thought to be due to the ringing effect, which is the continued displacement of the basilar membrane after the end of a loud masking signal; a quieter signal cannot be heard if it overlaps with the ringing. The response to a lower level signal is also reduced by the short term fatigue in the auditory nerve produced by the masker. The neural activity excited by the masker persists at a higher level in the auditory system after the signal has ended.

Backward masking is a more minor effect than forward masking. Significant backward masking tends to last only about 1-2 ms and can usually be ignored. The amount of backward masking detected mainly depends on how much training the subject has received. Trained subjects often display little or even no backward masking. It is thought that this masking may in part be caused by the fact that louder sounds propagate faster through the hearing system and overtake a preceding lower level signal, causing the latter to become masked (Rossing, 1990).

2.2.2.3 Binaural masking level differences

When two identical signals are presented to both ears, the presentation is termed diotic. On the other hand, when the signal arriving at the left ear is different from the one at the right ear, the presentation is said to be dichotic. When listening to spatially separated sounds, the human auditory system exploits phase, level and spectral differences in the signals arriving at the two ears to assist in distinguishing them. Therefore, in certain conditions, if the sounds are presented dichotically, the masking threshold falls. For example, in the last case in table 2.2, when a sine wave maskee with frequency below 1500 Hz and possessing an interaural phase difference of π radians is presented simultaneously with a diotically presented white noise masker, the masking threshold is 15 dB lower than when the sine wave is presented diotically. This shift in masking threshold is known as the binaural masking level difference (BMLD) and represents the difference in the masking threshold between diotic and dichotic presentations of a signal.

Table 2.2: The masking level differences in different conditions, adapted from Moore (2012).

Interaural condition	MLD in dB
$\mathrm{N}_u\mathrm{S}_\pi$	3
$\mathrm{N}_u\mathrm{S}_0$	4
$\mathrm{N}_\pi\mathrm{S}_m$	6
$\mathrm{N}_0\mathrm{S}_m$	9
$\mathrm{N}_\pi\mathrm{S}_0$	13
$\mathrm{N}_0\mathrm{S}_\pi$	15

Table 2.2 shows the masking level difference (MLD) for several masker and maskee conditions. Labels N and S denote a noise masker and a sinusoidal wave (tone) maskee, respectively. The subscript ‘0’ means that the phase of the signal is the same at both ears, also known as a homophasic signal. The subscript ‘π’ means that the phase of the signals at the left and right ears differs by π radians, also known as an antiphasic signal. The second subscript ‘m’ denotes the target signal is presented monaurally. The last subscript ‘u’ means that the signals at each ear are uncorrelated. As shown in the table, the introduction of phase difference between signals arriving at the left and right ears affects the masking threshold. The results indicate that it is easier for the human hearing system to detect the maskee in noise when the signals are spatially separated. The maximum MLD occurs for the case $\mathrm{N}_0\mathrm{S}_\pi$ when the target signal is located to one side of the listener and the noise is placed in the median plane.

2.3 Auditory scene analysis

Unlike the visual system, where the spatial location of individual visual objects are directly mapped to specific locations on the retina, there is no direct mapping in the auditory system between the spatial location of a sound source and the location of physical displacements along the basilar membrane. Despite this, the human auditory system is still able to localise and segregate multiple concurrent sound sources in a complex mixture of sounds by using alternative perceptual mechanisms. The area of research that investigates this basis of auditory perception is referred to as auditory scene analysis (ASA). Bregman made major contributions to the early development of ASA, including a seminal text on the subject (Bregman, 1990).

It is relatively easy for a listener to attend to a solitary sound source because all the auditory information decoded by the brain contributes to the signal stream from that source. However, attention to an individual sound source becomes more complicated when one or more additional sound sources are introduced. This is because the sound source contents overlap spectrally and temporally. The goal of ASA is to recover separate descriptions of "each separate thing" in such an environment (Bregman, 1990). In fact, "each separate thing" may consists of many sounds (e.g. footsteps).

Conceptually, ASA is concerned with two main processes: segmentation, which describes the decomposition of the acoustic input into a set of time-frequency fragments and grouping, which is the combining of fragments that are likely to have originated from the same sound source. Bregman (1990) makes a distinction between two types of perceptual grouping: primitive grouping and schema-driven grouping. In the following sections, segmentation and grouping are discussed. Furthermore, how different acoustic features, such as onset, harmonicity, amplitude and frequency modulation, and spatial cues, contribute to perceptual grouping is reviewed.

2.3.1 Segmentation

As described in section 2.2, the auditory system continually performs a frequency decomposition of all sounds presented to the ears, resulting in temporal-spectral fragments. These broken down time-frequency regions are the fundamental building blocks that form a stream (see section 2.3.2). They are localised in the sense that they belong to a particular time and frequency region. In addition, there are many properties which can be used to describe spectral fragments, such as amplitude modulation, frequency modulation, fundamental frequency and binaural cues (ITD and ILD). Segmentation takes place between auditory peripheral processing and perceptual grouping.

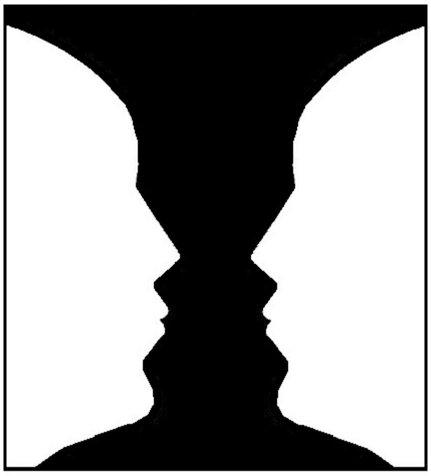


Figure 2.8: Illustration of exclusive allocation in which either two faces or a vase can be seen. Redrawn from Bregman (1990).

Humans have the ability to assign particular spectral fragments to different streams at different times. According to Gestalt principles, the spectral fragments can only be assigned to a single stream at any one time (Bregman, 1990). This phenomenon is referred to as exclusive allocation and is demonstrated visually in figure 2.8. However there are exceptions to this rule (Bregman, 1990).

2.3.2 Primitive grouping

The role of auditory perception is to group spectral fragments and attribute them to the individual sound source that created them, a process known as grouping. The spectral fragments for a sound source can be linked together in time, resulting in an auditory stream. Grouping results in the perception of a single sound source rather than a set of separate sounds. In this section, we introduce the primitive grouping which is also known as bottom-up grouping. Primitive grouping includes the integration of sequential and simultaneous segments.

Many primitive grouping principles can be considered as the Gestalt principles of perceptual organisation. A number of laws developed by the Gestalt psychologists (e.g. Köhler (1970) and Koffka (2013)) aim to explain perceptual organisation, or the manner employed by the brain in which small elements form mental patterns. These principles are applicable to vision and audition. The laws state that elements having common attributes can be grouped together. Next we explain these attributes in greater detail.

2.3.2.1 Sequential integration

Sequential integration of segments explains how the spectral fragments from the same source occurring at different instances are grouped together. Many cues are used by the auditory system to inform the grouping process. Two sound sources should likely be grouped together if they are sufficiently similar in respect of a low-level relationship, such as time or frequency. This property of the auditory system is referred to as the Gestalt principle of proximity.

Spectral and temporal relations

Spectral and temporal characteristics which affect grouping were demonstrated in experiments by Van Noorden (1975). These experiments employed

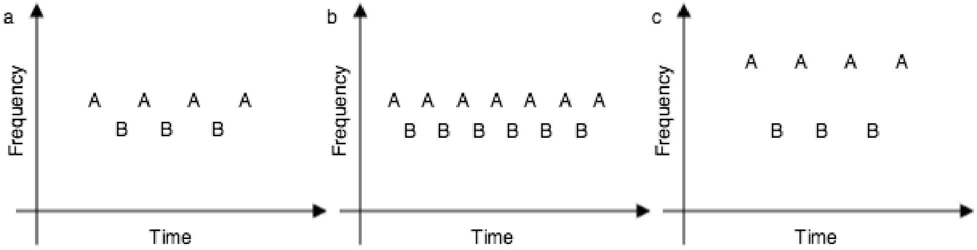


Figure 2.9: Sequential grouping of alternating pure tones, adapted from Wang and Brown (2006).

sequences of separate 40 ms tones with onset-to-onset times ranging between 60 and 150 ms. The sequences of tones consisted of fixed frequency tones, denoted by B, and a variable frequency tone, denoted by A. The tones were organised in the pattern $ABA-ABA-...$, where the hyphen in the sequences indicates a silence. The fixed frequency tone B was set to 1 kHz, and the frequency of the variable frequency tone A was slowly lowered from a frequency much higher than B to a frequency much lower than B and then raised back up again. The duration of the frequency sweeping was 80 s.

A simple example of sequential grouping is shown in figure 2.9. When the tones are presented to listeners with a frequency difference of less than about four semitones. Listeners' perceived organisation of the tones depends on the rate of the two tones. In figure 2.9 (a), tone A and B with low rate (the time between onsets is 150 ms) are grouped together and one stream of alternating tones is heard. However, it becomes more difficult to hear a single stream when the rate of A and B is increased, as shown in figure 2.9 (b). Similarly, when tone A and B are presented at the same low rate but with a larger frequency difference, more than 12 semitones, they are segregated into two streams (see figure 2.9 (c)). For intermediate frequency differences, listeners could switch at will between perceiving the tones as integrated or segregated. Therefore, the separation was measured under two conditions in experiments by Van Noorden (1975). Firstly, listeners were asked to try to perceive all the tones as part of a single sequence and to report when they

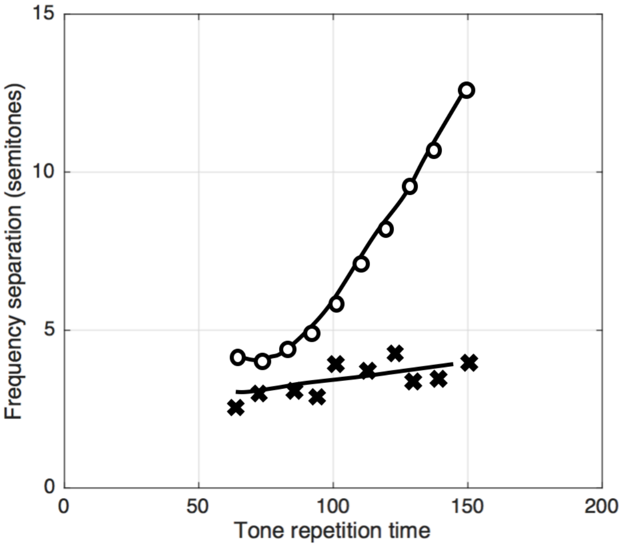


Figure 2.10: The influence of tone repetition rate and tone frequency separation on streaming: Temporal coherence boundary (O) and fission boundary (X). Redrawn from Bregman (1990), original from Van Noorden (1975).

could not. Secondly, they were asked to try to perceive the tone sequences as separate and again report when they could not. These two conditions resulted in two thresholds: the temporal coherence boundary (TCB) and the fission boundary (FB) for the first and second conditions, respectively. TCB measures the point at which the auditory system is forced to segregate the two tones by primitive processes, and the FB is a measure of the limits of the attention-based processes in grouping a stream through selective attention (Bregman, 1990).

The two boundaries are plotted in figure 2.10 as a function of frequency separation and tone rate, where tone rate is the onset-to-onset time. The graph indicates that listeners distinguish the two tones as separate streams for smaller frequency separations when they are presented at higher rates. It also shows the limits of attention-based processes that form a stream. Above the TCB curve, two tones are always perceived as part of two separate sequences, while below the FB curve, the two tones are always perceived as part of the same stream. In the region in between the curves, exclusive allocation occurs. Listeners could hear either integrated or segregated tones at will, but never both possibilities at once.

Frequency properties of a complex tone which contribute to grouping include fundamental frequency, pitch and spectral balance (Bregman, 1990). It is noteworthy that pitch is perceived at the fundamental frequency of a complex tone, whether or not the fundamental frequency component is present (Licklider, 1951). The relative levels of the harmonics of a complex tone form the spectral balance. Bregman (1990) stated that all these frequency properties influence sequential grouping. This type of grouping follows the Gestalt principle of proximity.

Spatial location

Another grouping principle for sequential integration of sound is spatial location. Sounds that originate from the same spatial location or segments having common interaural time or level differences tend to be grouped. However, the grouping effect is not as strong as fundamental frequency. This is demonstrated in the experiment conducted by Deutsch (1975). In the experiment, it is shown that a tune becomes unrecognisable when the notes of the tune are played to the left and right ears alternately. However, the tune becomes recognisable again when each note is presented with a drone (i.e. a lower constant frequency tone) to the opposite ear. This result indicates that fundamental frequency cues can dominate location cues in grouping when the location cues become weaker in the presence of another simultaneous tone. Darwin (1997) confirmed this and went on to demonstrate that spatial location influences the grouping more as a stimulus continues and it may become the dominant cue at longer durations.

2.3.2.2 Simultaneous integration

The other type of primitive grouping is simultaneous integration, in which components from the same source that occur at the same time are grouped together. An example of simultaneous integration is when we hear separate instruments playing a chord (Darwin, 1997). It can be easy to distinguish the individual contributions of less skilled players who are out of tune or time, or who play with different consonant time intervals. A potent Gestalt principle in this context is known as common fate, which states that segments tend to be grouped together when they change in the same way at the same time. Many simultaneous grouping principles can be phrased in terms of common fate.

Harmonicity

In natural speech, there are many starts, stops and pauses. Hence, for two competing voices, temporal relations, such as common onset, can be grouped to help separate two speech signals. Listening to one voice during pauses in the other help to separate two speech streams as well. When there are no pauses in the speech, some other grouping principle may contribute to the separation. Harmonically related frequency components tend to be grouped into the same stream. The human auditory system is able to identify multiple harmonic series when they have different fundamental frequencies, and can infer a missing fundamental. It becomes difficult to distinguish a target voice when two voices are present which have the same fundamental (Bird and Darwin, 1998). This is similar to the example given above where two or more instruments played simultaneously may be heard as a chord. In such cases, spectral density can be another important grouping principle. The likelihood of grouping increases with increasing spectral density or when a set of partials have similar intensities.

Spatial location

Two indispensable attributes for grouping are time and frequency. That is, two simultaneous sounds will be perceived as separate if they differ in frequency. Similarly, two sounds with the same frequency but which occur at different times will be perceived as separate. Compared with time and frequency, Kubovy (1981) states that spatial location is not an indispensable attribute. Kubovy uses the example that two identical sounds from different spatial locations will be perceived to be fused into a single sound coming from an intermediate direction.

In an informal experiment, Bregman (1990) discovered this is not always the case, however. In some conditions, segregation of two identical complex tones which only differ in spatial location may occur. The first complex tone consisted of frequency components at 200, 400, 600 and 800 Hz, and the second one consisted of frequency components at 300, 600, 900 and 1200 Hz. All the components had equal intensity. These complex tones were simulated in the horizontal plane with azimuths of 45° and $-45°$ respectively. They were presented at irregular intervals (different onset and offset times) but overlapped for substantial durations. At the time when the 600 Hz component was active in both tones and were of identical intensity and phase in each ear, then, according to Kubovy (1981), this 600 Hz component would be fused and perceived in front of the listener. This was not the case, however. Instead, listeners found that the 600 Hz component always remained spatialised in the directions of both complex tones. Bregman (1990) explained that this is because the 600 Hz component is grouped with other frequency components at the instants of onset and offset.

Amplitude and frequency modulation

It is very unlikely that different sound sources will change in the same way and at the same time instants (Bregman, 1990). For this reason, frequency components that have the same temporal modulation have a tendency to be grouped into the same stream. This principle applies to both amplitude and frequency modulation and it is an example of common fate. Amplitude modulation typically has two forms: onset/offset synchrony and changes in amplitude. Onset/offset synchrony between two pure tones causes them to be grouped (Bregman and Pinker, 1978). A common modulation pattern for changes in amplitude is periodic. In terms of frequency modulation, it may involve gliding changes and micromodulation. A relatively slow and gradual shift in partials can be referred to as a gliding change. Micromodulation, on the hand, refers to smaller and faster changes in frequency. All these factors contribute to simultaneous grouping (Bregman, 1990).

2.3.3 Schema-driven grouping

Compare with primitive grouping described in section 2.3.2, schema-based (top-down) grouping, on the other hand, is a high-level process which is aided by learning and experience. An example of schema-based streaming is the human ability to separate simultaneous speech sources based on semantic structure and understanding. This type of grouping is a hypothesis-driven process, in that listeners are able to pick up streams from a mixture by using stored knowledge of familiar patterns or schemas. This stored patterns can be particular types of sounds, such as music, speech and environment sound. An example to occur with speech sounds is given here. The two different synthesised vowels, "ee" and "ah", which have the same onset and offset times, the pitch at any time and the same loudness contour. Regarding to primitive grouping principle, they are unlikely to be separated. While listeners are able to hear the two individual vowels.

2.4 Summary

In this chapter, the human hearing system and how sounds are perceived have been briefly described. These psychophysical mechanisms provide one approach for estimating the ideal binary masks and ideal ratio masks in the algorithm described in Chapter 5.

The structure of the peripheral auditory system includes the outer ear, middle ear and inner ear. A sound enters the ear canal after interacting with the pinna, and propagates to the eardrum, which passes the sound vibrations to the middle ear. The middle ear transmits the vibrations of the eardrum to the cochlea through the ossicles. The cochlea in the inner ear performs a frequency decomposition along the basilar membrane. The inner hair cells pass neural responses from the basilar membrane to the brain via the auditory nerve.

In simplistic terms, the cochlea can be thought of as a bank of bandpass filters. The bandwidths of the critical bands are not uniform, becoming wider as the frequency increases. The notched-noise method can be used to estimate the shape of the auditory filters, which is found to be an asymmetrical upward convex curve.

When listeners are presented with multiple sounds, masking may occur. Simultaneous masking occurs when the masker and maskee are active at the same time and are spectrally close to each other. Nonsimultaneous masking describes the masking which occurs when the maskee exists just before or just after the masker. The masking level difference reveals the difference in masking threshold which arises between the dichotic and diotic presentations of a sound. It demonstrates that there is an advantage to listening with two ears in the presence of a mixture of two spatially separate sound sources.

The perceptual mechanisms underlying the ability of the human auditory system to segregate multiple sound sources are explored in auditory scene analysis (ASA). ASA can account for, among many other things, for the cocktail party effect. Bregman (1990) breaks the segregation process into two overall stages: segmentation and grouping. In the first stage, the acoustic signal is decomposed into time-frequency segments which are described by many properties, such as amplitude modulation, frequency modulation, fundamental frequency, interaural time or level difference. The second stage groups the segments which originate from a single source into a perceptual stream. Primitive grouping consists of simultaneous and sequential grouping. This follows the Gestalt principles of proximity and common fate. Segments can be similar in terms of temporal relations, frequency, spatial location, harmonicity, amplitude modulation or frequency modulation. Schema-based grouping, on the other hand, is a high-level process which is based on listeners learning and experience. Simulating and exploiting the perceptual mechanisms of ASA have given rise to the research field of computational auditory scene analysis (CASA), which will be considered further in Chapter 4.

The human hearing system has the ability to localise sound. Spatial location contributes to both simultaneous and sequential grouping, using properties such as common interaural time or level difference. Hence, an understanding of how the hearing system uses localisation cues, particularly binaural ones, is required. In the next chapter, we review human binaural hearing which is also essential in the development of a binaural hearing aid algorithm.

Chapter 3

Binaural Hearing

In the previous chapter, an overview was given of the human hearing system and the perception of sound. One aim of the current research is to exploit some of the attributes of binaural hearing in an algorithm for improving speech intelligibility. Therefore, it is helpful at this stage to review the ability of the hearing system to localise sound and how it uses localisation cues, particularly binaural ones. This chapter begins with a description of the binaural spatial hearing system. It goes on to consider sound localisation for a single source. Head-related transfer functions, sound spatialisation and binaural synthesis are introduced as fundamental prerequisites for artificially spatialising sound sources.

3.1 Sound localisation

In preparation for subsequent sections, a head-related coordinate system is introduced to describe the direction of a sound source (see figure 3.1). With respect to the interaural axis, the frontal plane bisects the front and back halves of the head, and the horizontal plane bisects the upper and lower halves of the head. The median plane bisects the left and right halves. The direction of a sound source can be referenced as (θ, ϕ), where θ and ϕ denote the azimuth and the elevation angles of the source relative to the head, respectively. For example, a direction of (0°, 0°) refers to the direction 0° azimuth and 0° elevation, which is directly in front of the head in the horizontal plane. The ear closest to a sound source is referred to as the ipsilateral ear and the more distant one is the contralateral ear.

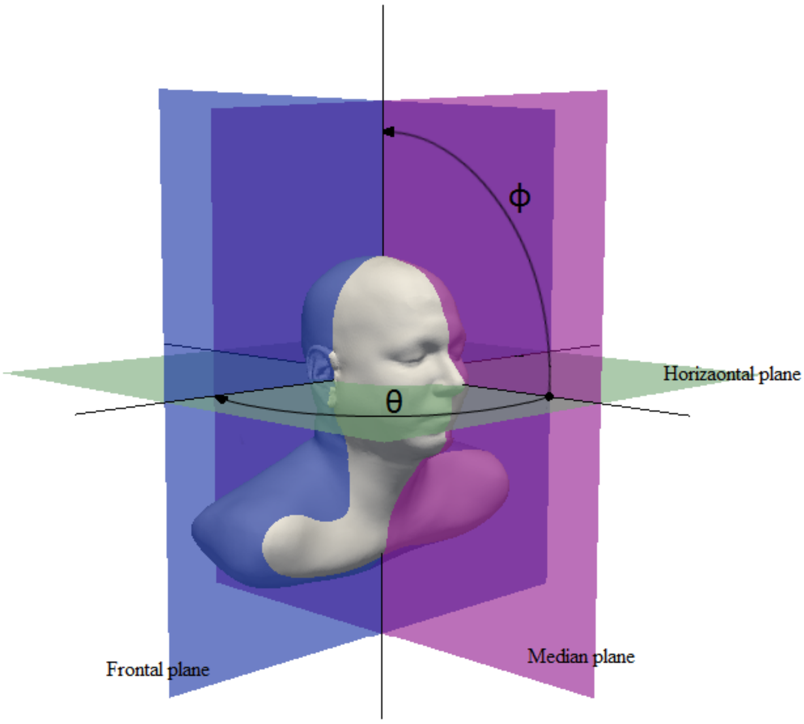


Figure 3.1: Head-related coordinate system. θ and ϕ represent azimuth and elevation, respectively. The arrows on θ and ϕ are pointing in the positive directions, with the origin straight ahead.

In previous sections, the physiology and psychoacoustics of the human hearing system are discussed. These aspects of the hearing system are not directly responsible for localising or segregating a sound source, but they generate some of the information which assists in the process. The two principal localisation cues for determining the direction of a sound source are the interaural time difference (ITD) and the interaural level difference (ILD). Rayleigh (1907) proposed the duplex theory which suggests that localisation cues are based on ITD at low frequencies and ILD at high frequencies. These and other cues which are able to provide localisation information are described next.

3.1.1 Interaural time difference

As mentioned above, an important cue for localisation by the human auditory system is interaural time difference (ITD). ITD describes the difference in the arrival time between the signals received at the left and right ears. Woodworth's formula (Woodworth and Schlosberg, 1962) provides a simple way to estimate ITD. In this formula, the sound source is assumed to lie at infinity (the far field condition) and the head is modelled as a horizontal section through a simple sphere (illustrated in figure 3.2).

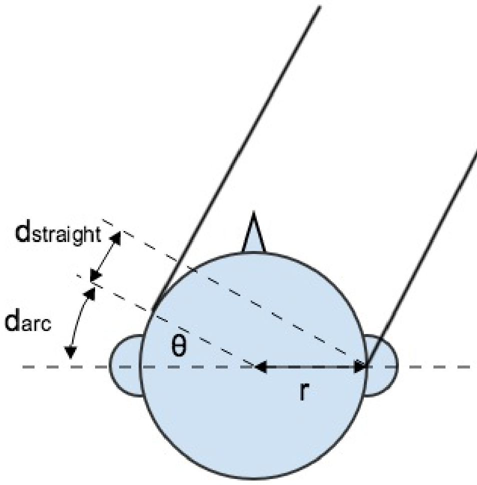


Figure 3.2: The path difference between left and right ear for a distant source.

The radius of the head is r and θ is the azimuth direction of the sound source in radians. The total path difference d_{total} that is shown in the figure consists of the arc d_{arc} and the line $d_{straight}$. d_{arc} and $d_{straight}$ can be expressed as:

$$d_{arc} = r\theta \tag{3.1}$$

$$d_{straight} = r \sin\theta \tag{3.2}$$

Thus the total path difference is:

$$d_{total} = d_{arc} + d_{straight} = r(\theta + \sin\theta) \tag{3.3}$$

If c is the speed of sound in the air, the expression for ITD is given by

$$ITD = \frac{r(\theta + \sin\theta)}{c} \tag{3.4}$$

For example, when θ is 90° (i.e. the source is located to the right of the head, representing the maximum path difference in this simple model), using 340 m/s for the speed of sound and assuming the radius of the head to be 90 mm, the path difference and ITD$_{max}$ are obtained by:

$$d_{total} = 0.09(\frac{\pi}{2} + \sin\frac{\pi}{2}) = 231.4\,\mathrm{mm} \tag{3.5}$$

$$ITD_{max} \approx 680\,\mu\mathrm{s} \tag{3.6}$$

Figure 5.8 in chapter 5 shows plots of some ITDs as a function of azimuth angles as calculated using equation 3.4.

When a sinusoidal signal is considered, ITD can alternatively be expressed as an interaural phase difference (IPD) between the left and right ears. Due to phase wrapping, this equivalence is only unambiguous for a sinusoid whose wavelength λ is greater than twice the interaural path difference. In other words, IPD is able to provide unambiguous information about the location of a sound source only for low frequency sinusoids below approximately 1.5 kHz. For example, for a 5 kHz sinusoid (i.e. with a period of 200 μs), an ITD of

400 μs would result in two complete cycles of IPD, incorrectly implying a source direction of 0° azimuth. There is evidence to suggest, however, that the hearing system uses the temporal envelope of the signal to estimate interaural time difference between the two ears at higher frequencies (Henning, 1974), and this is discussed further in section 3.1.2.

For a given ITD, if the head is kept stationary, there is insufficient information to fully localise a sound source and there is a cone-shaped surface over which the ITD is constant. This is an example of a cone of confusion (Mills and Tobias, 1972) and is illustrated in figure 3.3. To resolve this ambiguity the hearing system needs additional information which is described in subsequent sections.

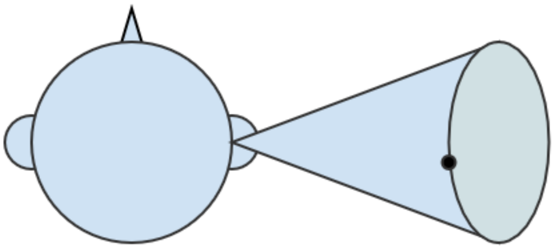


Figure 3.3: An example of a cone of confusion for a spherical head. All points on the cone's surface for this simple model have the same ITD. Points on circular cross sections of the cone have approximately the same ILD.

Kuhn (1977) conducted an experiment to reveal how ITD varies with frequency. He determined that ITD is influenced by varying interactions between the sound wave and the head, with diffraction occurring at low frequencies and creeping waves forming around the head at high frequencies. Kuhn defined a nondimensional parameter $\prod$ to describe the resulting variation in ITD:

$$ITD = \prod(\frac{r}{c}\sin\theta_{inc}) \tag{3.7}$$

where r is the radius of the head, c is the ambient speed of sound, and

θ_{inc} is the angle of incidence. His results show that for frequencies below 500 Hz $\prod$ equals 3 and ITD is frequency independent. ITD then decreases to a minimum between 1.4 kHz and 1.6 kHz. For frequencies above 3 kHz, $\prod$ equals 2. For example, assuming the radius of the head is 90 mm, the angle of incidence is 45° and the speed of sound in air is 340 m/s, then according to equation 3.7 the ITD is approximately 560 μs for frequencies below 500 Hz and falls to 370 μs above 3 kHz.

3.1.2 Interaural envelope difference

As mentioned in the previous section, the human hearing system is able to detect interaural time differences for signals with frequencies above 1.5 kHz, despite ambiguities arising in the interaural phase difference. When signals are analysed in the cochlea, the resulting bandpass-filtered signals essentially contain two kinds of information: the temporal fine structure, which forms the notional output waveform of each band, and the temporal amplitude envelope of the signals. For a signal with frequencies above 1.5 kHz, the human auditory system tends to use the interaural envelope difference (IED) rather than the interaural time delay of the fine structure (Henning, 1974). Because the envelope of a high frequency sound is modulated at a low frequency, the ambiguity in phase of the high frequency sound is resolved. In Henning's experiments, a 300 Hz modulation of a 3.9 kHz carrier was presented to each ear. In order to ensure that no ILD was present, the mean level and the modulation depth were made the same in both ears. The results indicated that the detection of interaural delay for these tones was as good as for the 300 Hz tone. Therefore, the envelope modulation of high frequency sounds provided localisation using ITDs alone.

The extent of the lateralisation achieved by an interaural delay is, however, small when IED carriers greater than approximately 1.6 kHz are presented over headphones (Bernstein and Trahiotis, 1985). Moreover, Middlebrooks and Green (1990) pointed out that the average modulation depth was too small, or the modulation frequency too high, to permit utilisation of IED in a typical free-field listening scenario.

3.1.3 Interaural level difference

Another cue that is used to detect the direction of a sound is the interaural level difference (ILD). At high frequencies, sounds are shadowed by the head, resulting in level differences between the left and right ears. This is shown in a simplified form in figure 3.4. The head shadowing effect becomes stronger as the wavelength gets shorter, as simulated in figure 3.4 (a). In plot (b), low

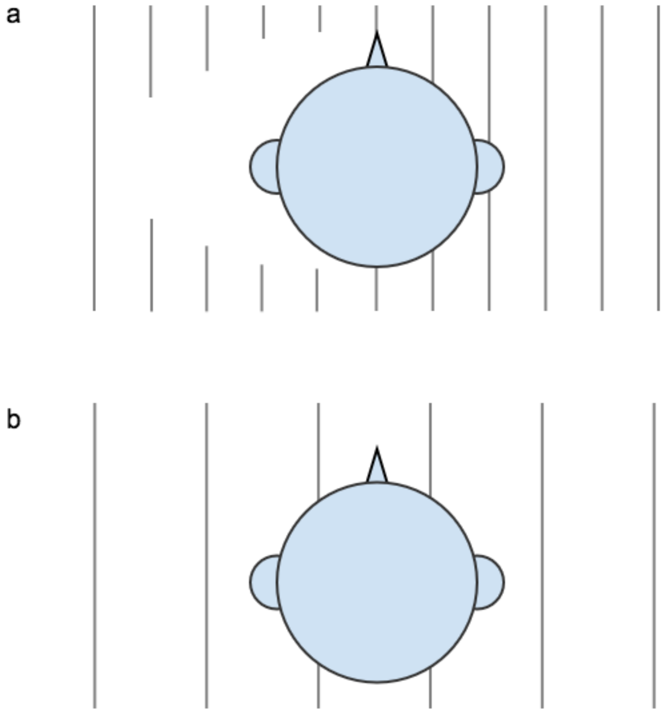


Figure 3.4: Simplified diagram showing (a) head shadowing at high frequencies and (b) diffraction at low frequencies. The source is coming from the right side of the head.

frequency sounds with wavelengths longer than the diameter of the head are shown tending to diffract around the head, reducing the ILD.

Using 340 m/s for the speed of sound c and setting the wavelength of the sinusoidal sound equal to a head diameter of 180 mm, equation 3.8 can be used to evaluate the frequency f above which head shadowing becomes significant.

$$f = \frac{c}{\lambda} = \frac{340}{0.18} \approx 1.9\,\text{kHz} \tag{3.8}$$

Thus, sounds above approximately 1.9 kHz are affected by head shadowing effects. Figure 3.5 demonstrates the frequency-dependent nature of the ILD cue due to decreasing diffraction as the frequency increases. For a sinusoidal sound that is distant from the listener, ILDs are very small below approximately 500 Hz, while they can be as large as 20 dB at high frequencies.

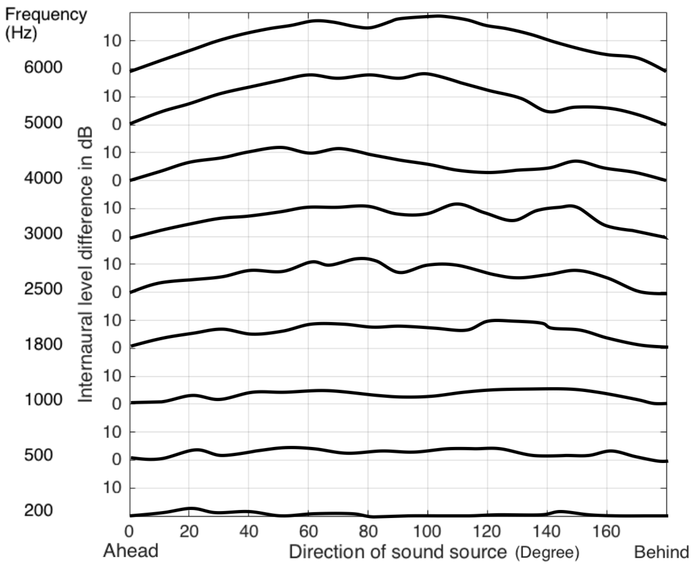


Figure 3.5: Interaural level difference as a function of frequency and direction. Redrawn from Feddersen et al. (1957).

Middlebrooks et al. (1989) observed that ILD varies with both azimuth and elevation of the sound source. ILDs at some frequencies display lateral asymmetry, reflecting the morphological differences which typically occur between the two ears. As a consequence, substantial ILDs can arise even for sounds located in the median plane.

For sound sources very close to the head, the situation is different. Considerable ILDs can occur even at lower frequencies in this situation (Brungart and Rabinowitz, 1999). Because of this behaviour, ILD also provides a distance cue for sounds close to the head and so helps to resolve the cone of confusion to a single circle by reducing the possible sound locations on the cone's surface. The role of ILD as a distance cue will be discussed in more detail in section 3.1.5.

3.1.4 Spectral cues

In addition to ITD and ILD cues, the auditory system also receives detailed spectral information that aids sound source localisation. When a sound propagates from a source to each eardrum, it is influenced and filtered by the torso, shoulders, head and pinnae, resulting in changes to the original signal spectrum. This affects both the monaural and the interaural signal spectra, furnishing additional cues for localising the sound source. These spectral changes are embodied in an individual's head-related transfer functions (HRTFs). Since HRTFs fully describe the magnitude and phase changes experienced by an incoming direct sound between its source and the eardrum, they necessarily also contain the ITD/IPD and ILD cues. The HRTF is the frequency domain description of this process; the equivalent description in the time domain is known as the head-related impulse response (HRIR). The HRTF and HRIR form a Fourier transform pair. Due to variations in the morphology of the torso, head and pinnae, HRTFs vary from person-to-person (Møller et al., 1995).

Figure 3.6 illustrates the frequency range predominantly influenced by relevant parts of the body. The cavum conchae lies at the entrance to the ear canal and is the largest resonant region of the pinna. The cavum conchae, the ear canal and eardrum also modify the spectrum of arriving sounds, but the changes do not contain directional information (Algazi et al., 1999).

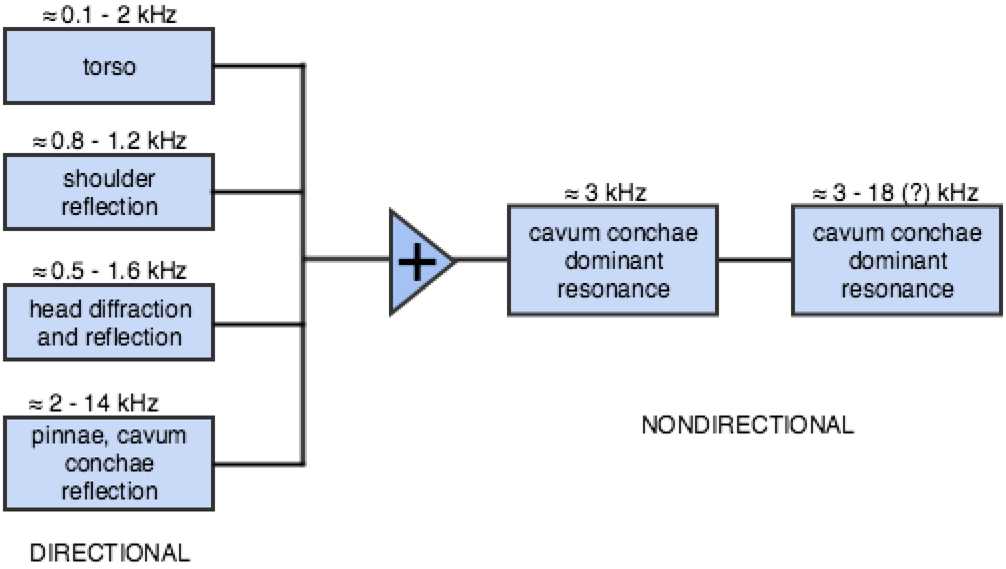

≈0.1 - 2 kHz
torso
≈0.8 - 1.2 kHz
shoulder reflection
≈0.5 - 1.6 kHz
head diffraction and reflection
≈2 - 14 kHz
pinnae, cavum conchae reflection
DIRECTIONAL
+
≈ 3 kHz
cavum conchae dominant resonance
≈ 3 - 18 (?) kHz
cavum conchae dominant resonance
NONDIRECTIONAL

Figure 3.6: Factors affecting the propagation of a sound between the source and the eardrum. The range of frequencies most likely affected by each factor is indicated. The upper ranges of the ear canal resonance must usually be approximated by models. Redrawn from Begault et al. (1994).

According to Begault et al. (1994), the key spectral changes for localisation are due to the pinna. Because of the shape and size of the folds and cavities within the pinna, their greatest impact is on the higher frequency components in signals (Begault et al., 1994; Moore, 2012). Musicant and Butler (1984) confirmed that the ability of listeners to localise 4 kHz high-passed noise was significantly superior to their localisation acuity when the pinna cues were partially removed by occluding the external ears. They found that there was no further performance degradation for 4 kHz low-passed filtered noise when the pinna cues were removed.

It is widely accepted that spectral cues from the pinna contribute most strongly to sound localisation in elevation. For example, early research undertaken by Humanski and Butler (1988) shows that the performance of localising a sound in the vertical plane is similar when using only the ipsilateral ear or when either using both ears. This supports the notion that it is the monaural pinna spectral cues which contribute most to localisation judgments in the vertical plane. In monaural listening, further experiments have shown that listeners perform better for high-passed noise than low-passed noise in terms of sound localisation in the vertical plane (Butler and Humanski, 1992). Butler and Humanski found that listeners were able to localise the low-passed noise binaurally rather than monaurally in lateral vertical planes, suggesting that only binaural temporal and level difference cues are used at these frequencies. In addition, they noticed that the influence of the pinna diminishes for source elevations over 45°. They suggested that time and intensity difference and pinna cues must be available for sound to be fully localised in the vertical plane.

As mentioned above, it is not only the pinna that contributes to the formation of HRTFs. Other parts of the body are also obstacles in the sound propagation path to the eardrum, and they too modify the spectra of incoming sounds. They can create low-frequency cues and help to explain why some binaural recordings produced using a dummy head, with and without a torso, create sound images with different elevations. Motivated by the observation that listeners are able to accurately estimate the elevation of sources positioned outside the median plane, Algazi et al. (2001) analysed HRTFs and attributed this ability to low frequency elevation-dependent head diffraction and torso reflections.

Spectral cues also help disambiguate frontal sources from rear sources and supply the elevation information which helps to resolve further the cone of confusion (Begault et al., 1994). For a nearby source, ITD and ILD cues can reduce the possible locations on the cone surface to a single circle (see section 3.1.3). The extra front-back and elevation information further reduce the circle of possible locations down to a small solid angle approximating a single point.

3.1.5 Distance cues

Shinn-Cunningham (2000) investigated the role of overall signal level in determining the distance of a sound source. For a distant source, the overall sound level follows the inverse square law. For example, there is 6 dB reduction in the received energy for each doubling of the source distance. However, for a sound source less than one meter from the head, the inverse square law does not apply. In fact, for a nearby source at a particular distance, the level depends also on source direction. The level of a nearby sound changes relatively slowly with distance in the median plane and changes faster as a function of distance on the interaural axis. Based on these results, Shinn-Cunningham concluded that overall signal level can only provide relative distance information unless the listener has prior knowledge about the source level.

Changes in spectral shape of a sound source with distance form another distance cue (Coleman, 1963). Coleman pointed out that high frequencies and low frequencies are attenuated differently as they propagate through air. As the source moves further away, high frequencies diminish more rapidly in amplitude than low frequencies. Later work carried out by Little et al. (1992) shows that this loss of high frequency energy results in high frequency sounds tending to be heard closer to the listener than low frequency sounds. Thus, a relative distance cue is provided by nonuniform spectral attenuation though an absolute distance cue requires prior knowledge of the source spectrum.

ILD not only provides direction information, as described in section 3.1.3, but also distance information. For nearby sources, ILD varies strongly with source distance (Brungart and Rabiowitz, 1996). ILD increases substantially at all frequencies for a lateral source when its distance decreases below 1 m (Duda and Martens, 1998; Brungart et al., 1999). Brungart (1998) showed that, based on distance-dependent changes in the ILD, an absolute distance cue can be added into a virtual auditory display for close sources.

In reverberant conditions, a sound arriving at the ears consists of the direct sound, early reflections and late reverberation. The ratio of direct-to-reverberant sound pressure at the ipsilateral ear varies with source distance and hence also contributes to distance perception. Shinn-Cunningham (2000) found that the pressure ratio varies nearly linearly with distance for a sound source located at 90° azimuth between 1 m and 0.15 m. This pressure ratio also varies with source distance when the source is placed at 0°, where ILD is approximately zero and therefore where ILD changes cannot contribute to distance perception.

3.2 Resolution of spatial hearing

In section 3.1, a general overview was given of cues that contribute to sound localisation. As described in the cocktail party problem, the human hearing system provides us with the ability to localise and selectively access a particular source. In order to segregate one target source in the presence of other competing sources based on localisation, it is helpful to know how accurately we are able to localise a signal source. In this section, the resolution (or acuity) of the human hearing system for distinguishing spatially separated sound sources is described.

3.2.1 Minimum audible angle

The minimum angular separation between spatial locations of two tone pulses which can be detected by the hearing system is referred to as the minimum audible angle (MAA), first described by Mills (1958). Its measurement in the horizontal plane is obtained in static conditions by presenting a tone pulse at a fixed spatial location and then presenting it again, slightly shifted left or right in azimuth. The smallest angular change between two locations for which there is a just noticeable difference is recorded. MAAs vary in magnitude as a function of azimuth and elevation (Perrott and Saberi, 1990). The MAA thresholds are dependent on the frequency of the tone (Mills, 1958). The results from Mills (1958) show that the smallest MAAs are about 1° in azimuth for sounds between 250 Hz and 1 kHz which come from directly in front of the listener. MAA rises sharply for frequencies in the range 1 to

1.5 kHz. Perrott and Saberi (1990) measured the MAA threshold for changes in elevation and reported MAA values of about 4°.

3.2.2 Minimum audible movement angle

Another measure of the resolution of spatial hearing is the minimum audible movement angle (MAMA), which was first measured by Harris and Sergeant (1971). The MAMA is defined as the minimum angle that a sound source must move through before the change of location is detected. It can be obtained by presenting a sound source through a loudspeaker on a rotating arm. The listener must detect whether the sound source is moving or stationary. Using tones with frequencies between 800 Hz and 6.4 kHz, Harris and Sergeant (1971) measured MAMAs for a very slow moving sound source at a rate of 2.8°/s. They found that MAMAs lie in the range 2° to 4°. Perrott and Musicant (1977) used a tone at 500 Hz as a stimulus. They found that MAMAs increase as the velocity of the sound source is increased, with MAMAs of 8.3°, 12.9° and 21.2° for a sound source with rates of 90°/s, 180°/s and 360°/s, respectively. Grantham (1986) provided more evidence that MAMAs vary with sound source velocity. Chandler and Grantham (1992) concluded that MAMAs are larger than the corresponding MAAs in most cases. MAA can be considered as a special case of MAMA, where the rate of source movement is zero. However, this is an over simplification and Grantham (1986) points out that MAMAs are affected by sound source duration as well. Specifically, the MAMA increases sharply from 5° to 20° or more when the duration of the stimulus is decreased below 100 to 150 ms. Moreover, their experiments showed that the human hearing system is most sensitive to sound source movement directly in front of the head and least sensitive at large lateral angles.

3.3 HRTF measurement

A binaural pair of HRTFs describes the spectral filtering that occurs between a sound source for a given direction and each of the listener's eardrums. HRTFs vary as functions of azimuth, elevation and distance. Of particular relevance to this work is the use of HRTFs to spatialise a sound source.

The measurement of an HRTF involves playing a stimulus signal through a loudspeaker at the required position. A microphone is placed at or close to the eardrum to capture the output from the loudspeaker after it has been modified by the HRTF. The analogue signal at the microphone is pre-amplified and digitised and compared with the original stimulus. The transfer function of the full signal path is then defined as the total system transfer function (TSFT). Apart from the HRTF in the path, the digital signal has also been affected by the unwanted transfer function of the amplifier, loudspeaker, microphone, cables and room. These unwanted transfer functions make up the system transfer function (STF). The STF can be measured in a similar way by repeating the measurement for the full path between the loudspeaker and the microphone without the presence of the listener. The desired HRTF $H(\omega)$ is then given by:

$$\begin{aligned} H(\omega) &= \frac{TSTF}{STF} \\ &= \frac{H_{DA}(\omega)H_{AMP}(\omega)H_{SPKR}(\omega)H(\omega)H_{MIC}(\omega)H_{PREAMP}(\omega)H_{AD}(\omega)}{H_{DA}(\omega)H_{AMP}(\omega)H_{SPKR}(\omega)H_{MIC}(\omega)H_{PREAMP}(\omega)H_{AD}(\omega)} \end{aligned} \tag{3.9}$$

where $H(\omega)$ is a single HRTF and can be transferred to the time domain via the inverse Fourier transform, to produce the corresponding HRIR, and $H_{DA}(\omega)$, $H_{AMP}(\omega)$, H_{SPKR}, (ω), $H(\omega)$, $H_{MIC}(\omega)$, $H_{PREAMP}(\omega)$ and $H_{AD}(\omega)$ are the transfer functions for analogue-to-digital converter, amplifier, loudspeaker, microphone, preamplifier and digital-to-analogue converter, respectively. This measurement process is repeated to obtain the HRTFs (one left, one right) for the desired set of directions.

Ideally, to measure the HRTF accurately for a particular direction, the microphone should be placed directly at the eardrum and the loudspeaker should represent a point source. In reality the microphone is usually placed at the entrance to the occluded ear canal and a single-driver loudspeaker is used with a wide frequency response (Butler and Musicant, 1993; Pralong and Carlile, 1994; Algazi et al., 1999).

It is far from practical to place a microphone at the eardrum, because it is a clinical procedure and can potentially result in damage to the eardrum (Wightman and Kistler, 1989). It has been shown that the ear canal provides very little directionally dependent filtering over the audio frequency range (Algazi et al., 1999) (see section 3.1.4). The perceptually salient directional detail is still preserved when the microphone is placed at the entrance to the occluded ear canal. Generally, HRTFs are measured at a distance of at least 1 m, to minimise the parallax effect (Brungart, 1999), which creates a difference between the angles of incidence at each ear.

The acoustic measurement of a full set of several hundred HRTFs is a time-consuming process. A large research effort continues to go into investigating alternative approaches for synthesising HRTFs. This topic is beyond the scope of the work described in this thesis. Approaches can be summarised as those using the boundary element method (Brebbia and Dominguez, 1996), differential pressure synthesis (Tao et al., 2003), the statistical relationship between morphology and HRTFs (Brown and Duda, 1997), physical and functional models of HRTFs and functional (Brown and Duda, 1998). This topic is beyond the scope of the work described in this thesis and will not be considered further.

HRTFs are generally measured in an anechoic chamber or in an environment where room reflections can easily be removed. Sometimes, however, it is required to capture the room reverberation as well so that a sound can be spatialised in realistic conditions. The resulting measurement is known as a binaural room impulse response (BRIR). A BRIR includes the HRTF, but also the early reflections and reverberant tail due to the room. A BRIR is affected by many factors, such as the position of the sound source in the room, the position of the listener and the directions that the sound source and listener are facing.

When listening in an enclosed space, a complex fusion of sounds arrives at the ears. The sounds include signals directly from the sound source and also indirectly from reflecting surfaces, such as the ceiling, floor and walls in a room.

As mentioned above, a BRIR is characterised by three parts; the direct sound, early reflections and the reverberant tail. The direct sound travels in a straight line from the sound source to the listener and is the first element to arrive at the ears of the listener. The first early reflection travels indirectly from the source to the listener and therefore arrives after the direct sound. It is reflected by the closest surface in the environment. Further early reflections typically follow soon after, which have been reflected by one, two or more surfaces before reaching the listener's ears. The early reflections are generally lower in amplitude than the direct sound. This is partly due to the greater distances they travel (sound pressure level reduces by about 6 dB for each doubling of the distance). The amplitude is also affected by the size and position of the reflective surfaces, and the material from which the surface is made will reflect some frequencies better than others (Howard and Angus, 2009). Gradually, reflections arriving at the listener reduce in amplitude and become increasingly temporally closely spaced and chaotic; these late reflections form the reverberant tail.

3.4 Conclusion

This chapter introduces the human sound localisation system. An understanding of human localisation cues is essential in the development of a binaural hearing aid algorithm.

Interaural time difference (ITD) and interaural level difference (ILD) are the two main binaural cues used to determine the spatial location of a sound source. ITD is the dominant cue at frequencies below approximately 1.5 kHz while ITD becomes more important for frequencies above 1.9 kHz. In addition, interaural envelope difference (IED) plays a role at high frequency in the scenario where no strong low-frequency ITD cues are available.

When a sound propagates from a source to our eardrums, it is influenced and filtered by our body surfaces. This results in spectral changes to the original signal which form another type of localisation cue. Of the surfaces involved, the spectral modifications caused by the pinna provide the most important direction cues. Furthermore, the ambiguities that are created using ITD and ILD alone can be resolved with the aid of the spectral signature.

The human auditory system is able to detect sound location changes accurately to approximately 1° to 3° azimuth. In elevation, the minimum angle of change that humans can detect is around 4° to 5°.

An understanding of the binaural and spectral cues that humans use to localise sound sources and the accuracy of detection of sound location changes is essential for designing a source separation algorithm which is based on the sound spatial location. This research aims to use spatial cues and other cues extracted from the binaural mixture of sounds arriving at a listener's ears. The integration of these cues forms a core part of the technical work for the proposed algorithm in Chapter 5.

Head-related transfer functions are also introduced in this chapter. They are unique to each individual and can be used to spatialise sound sources in a virtual auditory system, which is the main method we use to create the training and test data sets in Chapters 6, 7 and 8. The training data is used to train the proposed algorithm, which is based on artificial neural networks. The test data is used to evaluate the performance of the systems.

Human listening performance including intelligibility and localisation of multiple sources in adverse conditions will be described in the next chapter. Sound source segregation lies at the heart of our research and the next chapter also reviews a variety of methods for performing this task.

Chapter 4

Segregation of Multiple Sound Sources by Listeners and Machines

In the previous chapter we discussed characteristics of the binaural spatial hearing system for a single sound source. This chapter begins by considering the performance of human hearing in adverse conditions. It covers the advantages afforded by binaural hearing and the effect on intelligibility and localisation of multiple simultaneous sound sources.

This research is concerned with the intelligibility of a target speech source in the presence of competing talkers in both anechoic and reverberant conditions. Many source separation algorithms have successfully separated the target speech from such a mixture and demonstrated an improvement in intelligibility. Consequently, we review some of these algorithms and discuss their performance and limitations. Algorithms considered include blind source separation, which typically utilises more than two microphones, model-based separation, which usually utilises only one microphone, and binaural computational auditory scene analysis (CASA).

4.1 Human listening performance

Humans have a remarkable ability to attend selectively to a single sound source within a wide variety of complex acoustic environments. Miller (1947) concluded that listeners with normal hearing can achieve adequate intelligibility rates at 0 dB signal-to-noise ratio (SNR) levels, which typically occur when talker and listener are within 0.7 m of each other when everyone is talking at the same level (Plomp, 1977). Furthermore, to achieve 95 % speech intelligibility for short sentences in speech-shaped noise requires that the speech energy arriving at the listener is only 2 dB above the speech reception threshold (SRT) (Steeneken, 1992). This demonstrates that a small gain in SNR could lead to a significant improvement in the intelligibility.

Nábělek and Robinson (1982) investigated human speech recognition performance on a single speech source in both anechoic and reverberant conditions. The results illustrate that intelligibility decreases in the presence of reverberation for subjects of all ages and longer reverberation times reduce the intelligibility further. In a set of typical results, for example, the speech recognition accuracy was 99.7 % in the anechoic condition when listening monaurally, but dropped to 97.0 %, 92.5 % and 87.7 % for reverberation times of 0.4 s. 0.8 s and 1.2 s, respectively. Bolt and MacDonald (1949) and Nábělek et al. (1989) proposed two factors to account for the degradation of speech intelligibility in reverberation. The first is self-masking, which refers to the temporal smearing of frequency changes within each phoneme. The second is overlap-masking, which occurs when the energy of a preceding phoneme masks a subsequent phoneme.

A number of studies have measured the ability of humans to understand speech in adverse conditions and have investigated the factors influencing intelligibility in the presence of competing sources (distracters). Hirsh (1950) suggests that the binaural release from masking can be exploited in binaural hearing aids. In this section, we begin by looking at the influences on intelligibility related to the benefits which result from the use of binaural hearing over monaural hearing (the so-called binaural advantage). These are of particular interest in view of the binaural arrangement of microphones adopted for the optimal cue mapping (OCM) algorithm described in Chapter 5.

4.1.1 Spatial release from masking

Cherry (1953) introduced the notion of the "cocktail party effect" and argued that binaural hearing provides an advantage over monaural hearing in challenging auditory conditions such as these. This advantage can be demonstrated by the binaural intelligibility level difference (BILD), which is related to the binaural masking level difference (BMLD) (see section 2.2.2). The BILD measures the signal-to-noise ratio difference between monaural and binaural representations that deliver the same level of intelligibility. Because there is no need for the background signal to be entirely silent to achieve a certain intelligibility level, the BILD is usually smaller than the corresponding BMLD.

Pollack and Pickett (1958) compared the reception of monosyllabic words in monaural and binaural listening conditions in the presence of 1, 2, 4 and 7 competing talkers. In the monaural listening (control) condition, the target source and competing sources were presented to a single ear. In the binaural listening (stereophonic) condition, two different sets of competing sources were presented to each ear and the target source was presented binaurally. The target and competing sources were controlled by the signal-to-background ratio (S/B ratio or SBR), which is the ratio between the target source power and the total power of all the competing sources. The intelligibility was defined as the percentage of correctly identified words. Pollack and Pickett measured BILD as a function of the number of competing sources at 50 % intelligibility, known as the speech reception threshold (SRT). The BILD is 12 dB when only one competing source is presented. The BILD decreases to 5.5 dB at the SRT when seven competing sources are presented, for both monaural and binaural listening conditions. The results illustrate the advantage of binaural hearing over monaural.

Hawley et al. (1999) also measured the intelligibility of speech in competing sound sources. He concluded that the proximity of the competing talker to the target talker influences intelligibility of the target more than the number of talkers. In addition, he demonstrated a spatial release from masking due to the binaural advantage when the target and competing sources are placed in different directions.

Moncur and Dirks (1967) and Nábělek and Pickett (1974) demonstrated that the binaural release from reverberation provides an improvement in speech intelligibility. When listening binaurally in reverberant conditions, Nábělek and Robinson (1982) found that there is a 5 % intelligibility improvement when the sound source is presented binaurally, compared with the intelligibility rate achieved using monaural listening.

Culling et al. (2003) investigated the effects of reverberation on the perceptual segregation of competing speech sources. In their experiments, the target and masker sources were either both located at 0° or at ±60° azimuth. They found that it is easier to understand the target speech when it is spatially separated from a competing source in an anechoic environment compared with when the environment is reverberant. They concluded that reverberation not only seriously affects a listener's ability to distinguish differences in the fundamental frequencies of competing voices, but also in their spatial locations.

4.1.2 Informational and energetic masking

Plomp (1976) measured the intelligibility of speech in the presence of a masker in anechoic and reverberant conditions, and binaurally and monaurally. A difference between this and previous research was that both noise and speech maskers were employed. His results show that the masked threshold (the signal-to-noise ratio required for intelligible speech) in the presence of a noise masker is approximately 3 dB higher than it is in the presence of a speech masker. That is, spatial release from masking is larger for a speech masker than for a noise masker.

The difference in spatial release in these conditions leads to the concepts of informational masking and energetic masking. If both target and masker signals are similar clearly audible sounds, then informational masking increases the difficulty with which listeners are able to discriminate the target sound from the interfering sound (Durlach et al., 2003). On the other hand, the condition that one or more portions of the target signal are rendered inaudible, where the masker signal overlaps in time and frequency, is called energetic masking.

Experiments conducted by Kidd Jr. et al. (1998) and Ihlefeld and Shinn-Cunningham (2008) further demonstrate that spatial separation facilitates detection of a target signal in the presence of an informational masker compared with a noise masker. Informational masking is caused by failures in either across-time linkage of target segments or in top-down selection of the target (Ihlefeld and Shinn-Cunningham, 2008). In addition, distinct interaural level differences in energetic masking help listeners to locate the target segments, but have little influence on streaming. This led Ihlefeld and Shinn-Cunningham (2008) to suggest that these mechanisms influence informational masking and spatial release between target and masker, improving streaming and target selection. When target and interference sources are presented from the same location, however, an informational masker is able to disguise the target signal more than a noise (energetic) masker. According to Kidd Jr. et al. (1998), the listener does not know which linkage of segments (target streaming or masker streaming) to focus on in the case of informational masking.

Brungart et al. (2006) examine how to isolate energetic masking. An ideal binary mask which retains only the dominant time-frequency regions of the target signal was used to isolate energetic masking in their experiments. Their results suggest that energetic masking influences speech-in-noise masking more than speech-on-speech masking. Additionally, when the target-to-interference ratio was between -12 dB and 0 dB, intelligibility was almost equivalent to that for the target alone, which suggests that spectral overlap may have a relatively small impact in most automatic speech recognition tasks with multiple concurrent talkers.

4.1.3 Auditory glimpsing

Based on sparseness and the redundancy of speech, humans have the ability to use "glimpses" of speech in spectro-temporal regions where it is least affected by background sounds. This section reviews a number of factors which contribute to the appeal of auditory glimpsing and the intelligibility improvement it can provide.

Strange et al. (1983) first evaluated the performance of the human hearing system when provided with partial signal information in the time domain, created by gaps in the temporal waveform. Seven modified /b/-vowel-/b/ (consonant-vowel-consonant) syllables were used in the experiment. For each of the syllables, different parts of the waveform were deleted and the temporal relationships of the remaining parts were manipulated. The results of syllable identification tests showed that untrained listeners were able to identify vowels accurately, even when vowel nuclei were silent, just based on dynamic spectral information contained in the preserved initial and final transitions. Furthermore, when the durational information which indicates intrinsic vowel length was eliminated, dynamic spatial information appeared to be implicated in the glimpsing process.

Kasturi et al. (2002) investigated speech intelligibility with either a single "hole" in various bands or two "holes" in disjoint or adjacent bands in the spectrum. They found the vowel and consonant recognition performance showed a modest decrease when a single hole occurred either in the low- or the high-frequency region of the spectrum. The vowel recognition rate was sensitive to the location of the holes when a second hole was introduced. However, a rate of around 70 % for correct consonant recognition was observed even when the middle- and high-frequency regions of the spectrum were missing. Kasturi et al. (2002) found that different frequencies contributed to vowel recognition unequally, whereas all frequency bands were equally important for consonant identification. Li and Loizou (2007) extended these ideas by further evaluating the dependence of glimpsing on spectral regions, and they also investigated the impact of spectral width and the duration of the glimpses. They confirmed that the frequency location and total duration of the glimpses affected speech recognition significantly. Their results indicate that the majority of the information used to improve speech intelligibility is contained in the low- and mid-frequency bands. They suggest that multiple short glimpse windows generally result in higher speech intelligibility than the presence of a few longer ones.

Cooke (2006) adopted an automatic speech recognition (ASR) system to identify consonants in noise. He argued that the proportion of the spectro-temporal regions glimpsed was a good predictor of intelligibility. In addition, he suggested that in a speech-in-noise segregation algorithm it may be simpler to focus on the regions with advantageous local SNR than to estimate the energy proportion of the speech signal in every time-frequency unit. An understanding of the mechanisms used by listeners to raise speech intelligibility in adverse conditions can, he argues, provide insights which promote progress in ASR (Cooke, 2006), which is a sentiment that is equally applicable to the OCM algorithm discussed later in this thesis.

4.1.4 Sound source localisation

In section 3.2 the resolution of human spatial hearing for a single sound source is discussed. In this section, speech localisation acuity of a target source in the presence of competing sources is considered. It was suggested by Begault et al. (1994) that accurate localisation of multiple sound sources is achieved by the independent analysis of localisation cues within each critical band.

Hawley et al. (1999) investigated localisation acuity in normal-hearing listeners for a target sound source in the presence of three competing sources in the horizontal plane. All the sources were spoken by the same male talker at the same level and were presented to listeners via seven loudspeakers with a regular separation of 30°. Listeners were asked to point to the source of a known sentence in the presence of unknown sentences from unknown locations. The results show that binaural localisation of a clearly audible speech target in the presence of three competing talkers is robust, though mis-localisation of a competing sound may occur.

Best et al. (2005) performed related experiments, measuring the effect of auditory spatial perception with a broadband masker and a broadband target stimulus. In their experiments, the temporal overlap and the total time duration of two sources were varied and the overall target-to-masker ratio was maintained at 0 dB. Their results show that a broadband masker does not affect the localisation of a broadband target, even where there is substantial overlap in time and frequency. This localisation robustness was not affected when the stimuli had simultaneous onsets and offsets. The results also demonstrated a small systematic error away from the direction of the masker in the lateral localisation angle of the target in the presence of a simultaneous noise masker.

4.2 Evaluation methods for speech enhancement systems

Since the aim of this research is to improve the intelligibility and quality of speech in the presence of competing sounds, it is essential to be able to measure the changes in these metrics due to the speech enhancement system.

Speech intelligibility is rated in terms of the percentage of spoken words that are recognised correctly. Speech quality is a function of the realism and naturalness of the signal. Generally, highly intelligible speech also exhibits good speech quality, and good speech quality results in highly intelligible speech, although this is not always the case. Increasing the quality of speech does not necessarily increase its intelligibility (Gold et al., 2011), but Ramírez and Górriz (2011) report that it can reduce listener fatigue. For this reason alone, speech quality is an important consideration in speech processing.

This section describes evaluation methods for speech enhancement systems in terms of the improvements they offer in speech intelligibility and quality. There are two categories of performance evaluation. The first is based on subjectivity and requires human listeners to make judgements during listening experiments. The second is objective and estimates the performance numerically using signal analysis.

4.2.1 Speech intelligibility

For a listener to understand a sentence, it is often not necessary for them to correctly identify every phoneme, or even every word in a sentence. Due to the redundancy in speech, the incorrectly perceived sounds may be replaced subconsciously with the correct ones. This is referred as the perceptual restoration of missing speech sounds (Warren et al., 1970) and it is related to the law of closure which describes human brain's tendency to fill in gaps in information and ignore contradictory information. Thus, the ability to understand speech varies widely, depending on an individual's linguistic competence.

4.2.1.1 Subjective methods

Fletcher and Steinberg (1929) were the first to describe articulation test material when they formulated a list of 66 consonant-vowel-consonant nonsense syllables. The number of correctly heard syllables indicated the phoneme intelligibility. A similar method was proposed by Miller and Nicely (1955), but the number of consonants was reduced to 16. Listeners were trained to identify component phonetic units and may have be confused by phonemes which did not accord with spelling. Therefore, evaluation of speech intelligibility is often based on word-level testing.

When undertaking the word intelligibility test, listeners were limited to responses which were real words. This allowed them to respond using words with defined spellings. Problems could arise, however, due to variations in the extent of listeners' vocabularies, since the test depended on the successful recall of memorised words that they had previously heard. One possible solution to these problems is to create many word lists with balanced difficulty, although this increases the length of the test and may require a listener to attend more than one test session.

Another subjective method for measuring intelligibility is at the sentence level. Kalikow et al. (1977) developed a set of sentences known as the speech perception in noise (SPIN) test. These are phonetically balanced with one key word per sentence at the end of the sentence. Based on the key word, these sentences are classified either as high predictability sentences (e.g. "The boat sailed along the coast") or low predictability sentences (e.g. "Jane was interested in the stamp"). Intelligibility scores are indicated by the proportion of correctly recognised key words in the sentences. Recently, a modified form of SPIN test has been devised which can be used in multiple signal-to-noise ratio (SNR) conditions and which allows the SNR corresponding to the 50% recognition rate to be calculated (Wilson et al., 2012).

4.2.1.2 Objective methods

Most objective methods for estimating intelligibility are based on measurements of how a signal changes across frequency bands. The work of French and Steinberg (1947) led to the definition of the Articulation Index (AI) as a standardised method of objectively evaluating speech intelligibility. There are two successors to AI: the Speech Intelligibility Index (SII) (ANSI, 1997) and the Speech-Transmission Index (STI) (Commission et al., 2003). These approaches are computed from the speech and noise levels in a set of frequency bands by weighting according to the relative contribution of the band to intelligibility. In another words, these are SNR-based methods. These approaches are, however, unable to model the effects of speech enhancement algorithms operating in the time-frequency (T-F) domain (Ephraim and Malah, 1985). For example, coherence SII and a normalised covariance-based STI procedure both show low correlation with speech intelligibility after ideal time frequency segregation processing has been applied (Taal et al., 2009).

More recently, the Short-Time Objective Intelligibility (STOI) metric, based on correlation of the spectral amplitude modulation of the clean and degraded speech, has been developed by Taal et al. (2011). It has been shown that STOI scores are highly correlated with the subjective intelligibility of speech in noise and T-F weighted noisy speech (Gomez et al., 2011; Schwerin and Paliwal, 2014).

Using the STOI metric, both clean and degraded speech are converted from the time domain into the frequency domain using the short-time Fourier transform (STFT) with 50 % overlap and a Hanning window of duration 25.6 ms. The complex-valued STFT coefficients are denoted by $\hat{x}(m, b)$ where m and b are the time frame index and frequency bin, respectively. The spectral coefficients are grouped into one-third octave bands. The norm of the j^{th} one-third octave band for clean speech is given by:

$$X_j(m) = \sqrt{\sum_{b=b1(j)}^{b_2(j)-1} |\hat{x}(m, b)|^2} \tag{4.1}$$

where b_1 and b_2 denote the lower and upper edges of the one-third octave band. Similarly, the T-F representation of the same one-third octave band can be obtained for the processed speech and denoted as $Y_j(m)$. Then the short-time temporal envelope of clean speech is defined as:

$$\mathbf{x}_{j,m} = [X_j(m - N + 1), X_j(m - N + 2), ..., X_j(m)]^T \tag{4.2}$$

where $N = 30$, which corresponds to a time duration of 384 ms (since the duration of each window is 25.6 ms). Similarly, the short-time temporal envelope of the degraded speech is denoted by $\mathbf{y}_{j,m}$. Before comparing the clean and degraded speech, $\mathbf{y}_{j,m}$ is clipped to remove the impact of frames with low speech energy. The clipped $\mathbf{y}$ is denoted as $\bar{y}$ and given by:

$$\bar{\mathbf{y}}_j(m) = \min\left(\frac{||\ \mathbf{x}_{j,m}\ ||}{||\ \mathbf{y}_{j,m}\ ||}\mathbf{y}_{j,m}(n), \lambda\mathbf{x}_{j,m}(n)\right) \qquad n \in 1, ..., N \tag{4.3}$$

where $\mathbf{x}(n)$ indicates the n^{th} element of $\mathbf{x}$, $||\cdot||$ denotes the Euclidean norm, and $\lambda = 6.623$ represents a $-15\,dB$ signal-to-distortion ratio. The intermediate intelligibility is calculated from the correlation between the two vectors:

$$d_{j,m} = \frac{(\mathbf{x}_{j,m} - \mu_{\mathbf{x}_{j,m}})^T(\bar{\mathbf{y}}_{j,m} - \mu_{\bar{\mathbf{y}}_{j,m})}}{||\ (\mathbf{x}_{j,m} - \mu_{\mathbf{x}_{j,m}})\ ||\quad ||\ (\bar{\mathbf{y}}_{j,m} - \mu_{\bar{\mathbf{y}}_{j,m})}\ ||} \tag{4.4}$$

where $\mu(\cdot)$ represents the sample average of the vector. The overall STOI score is then obtained by averaging $d_{j,m}$ over all bands and frames:

$$d = \frac{1}{JM}\sum_{j,m} d_{j,m} \tag{4.5}$$

where M and J denote the total number of frames and one-third octave bands, respectively.

4.2.2 Speech quality

In this section we consider the measurement of speech quality and, as in the previous section about measuring speech intelligibility, we split the assessment of speech quality into two basic approaches; subjective and objective.

4.2.2.1 Subjective methods

There are broadly two types of subjective speech quality measurements. The first is based on human-generated mean opinion scores (MOSs). Listeners are first trained by playing a set of reference speech stimuli indicative of the range of quality levels they will be presented with during the main test. They are required to rate the quality of the speech in each stimulus using a numerical MOS scale. The second approach is the preference test, typically in which speech stimuli are presented in pairs to a listener, who is required to select the stimulus which they perceive to have the higher quality.

4.2.2.2 Objective methods

An alternative way to measure speech quality is to automate the estimation of the MOS by means of an algorithmic comparison between the clean and the corresponding degraded signals. The Perceptual Evaluation of Speech Quality (PESQ) is an international standard for estimating the MOS, which has been agreed by the International Telecommunication Union-Telecommunication Standardisation Sector (Rix et al., 2001).

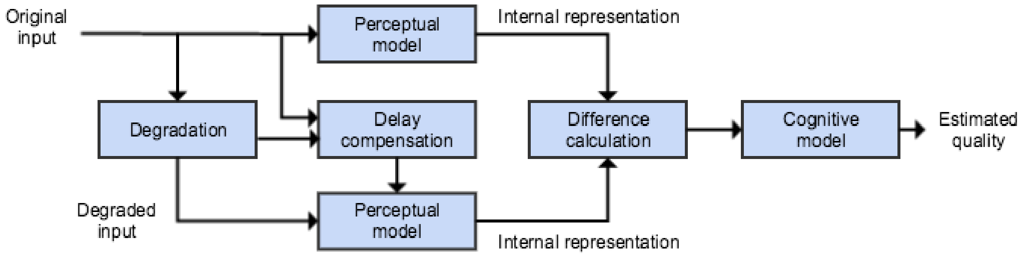


Figure 4.1: Simplified diagram of the PESQ algorithm, redrawn from Kondo (2012).

A simplified diagram of the PESQ algorithm is shown in figure 4.1. Both original and degraded signals are internally represented by means of a perceptual model. A delay compensation process is used to time-align the degraded signal with the original signal. The MOS estimation is then derived using a cognitive model based on the differences between the internal representations of the two signals. Experiments conducted by Kondo (2012) compared the objective quality measurement produced by the PESQ algorithm with the MOS scores produced subjectively by ten listeners. The results indicate that the objective MOS generally agrees with the corresponding subjective score.

4.3 Blind source separation

In this section, we begin by describing the blind source separation (BSS) problem. The aim of BSS is to recover a sound source from a mixture of sounds by means of several observations of the mixture using only information embedded in the mixtures themselves.

In real acoustic conditions, the delays and reflections experienced by a propagating sound need to be taken into account by the BSS model. For Q microphones and M sources, the observed mixture x_q at the qth microphone at time sample n can be described as (Dmour et al., 2011):

$$x_q(n) = \sum_{i=1}^{M} \sum_{p=0}^{P-1} a_{qi}(p) s_i(n-p) \qquad q \in [1, ..., Q] \tag{4.6}$$

where s_i is the ith source signal, a_{qi} denotes the room impulse response of length P, from source i to microphone q.

Recovery of the original sound sources is usually achieved by estimating the matrix of the unmixing filter w_{iq} (Dmour et al., 2011). Therefore, the estimated ith source $\hat{s}_i(n)$ can be written as:

$$\hat{s}_i(n) = \sum_{q=1}^{Q} \sum_{u=1}^{U} w_{iq}(n) x_q(n-u) \qquad i \in [1, ..., M] \tag{4.7}$$

where U denotes the length of the unmixing filters.

4.3.1 Independent component analysis

Independent component analysis (ICA) can be considered as a particular class of solution to address the BSS problem. It is based on two main assumptions. Firstly, it is assumed that the unknown sources have identical probability distributions which are statistically independent of each other, which is known as the independent and identically distributed assumption (the i.i.d assumption). In practice, this is a reasonable assumption to make in audio applications, since two highly correlated sound sources are unlikely to be active simultaneously. When this assumption breaks down, the likely result is that the sources become grouped into a single sound source. Secondly, each source must have a non-Gaussian distribution. If the sources are Gaussian, it is impossible to infer the direction of the columns of the mixing matrix (SID, 2005). ICA is able to separate sources in overdetermined situations, which means that the number of observed mixtures must at least equal the number of sources.

The concept of ICA was first proposed by Comon (1994) and further developed by Bell and Sejnowski (1995), motivated by the signal separation challenge. Bell and Sejnowski focused on solving the instantaneous noiseless mixing problem and they did not consider the source propagation delay. They provided a unifying framework to the BSS problem in their information maximisation approach. ICA can achieve good separation performance in the time domain, once the algorithm converges (Amari et al., 1997; Douglas and Sun, 2003).

An acoustic mixture usually includes delays and convolutions with spatial impulse responses. More recent research implements ICA in the Fourier domain (Smaragdis, 1998; Barry et al., 2005). The main idea is to perform ICA in each frequency band rather than in the time domain, wherein delays and convolutions which are shorter than the window length can be treated as phase modifications. This creates what is known as the permutation problem, in which the different frequency components belonging to each sound source must be identified and grouped in order to reconstruct each signal separately in the time domain. Many methods have been proposed to assign continuity criteria for extracting the correct components. For example, the permutation problem can be addressed based on a combination of direction-of-arrival estimation and correlation of signal envelopes between frequency bands (Sawada et al., 2004), or by using either one of these methods individually (Ikram and Morgan, 2002; Anemüller and Kollmeier, 2000).

A review of the performance of ICA for the separation of competing acoustic sources in anechoic and reverberant conditions is given by Kendrick and Shirley (2008). They conclude that separation performance varies with different microphone topologies and with the type of ICA algorithm. The results further illustrate that performance is degraded by the presence of reverberation. Furthermore, the requirement for overdetermination in ICA limits its application in hearing aids.

4.3.2 Beamforming

While ICA focuses on the statistical properties of signals, beamforming takes advantage of spatial information about them and forms another BSS solution family. Beamforming is a signal processing technique using an array of microphones of known topology. The goal is to preserve the acoustic signals coming from a particular, desired direction and to cancel out sounds coming from all other directions.

The delay-and-sum beamformer is one of the simplest kinds of fixed beamformer. With reference to figure 4.2, a wanted sound from a distant point source arrives at each microphone in turn in the linear array. The delay in the arrival time from one microphone to the next depends on the angle θ_s and is proportional to their separation. Appropriate delays, $T1$, $T2$, ..., are applied to compensate for the different times of arrival, such that the wanted signal from each microphone is brought into time alignment. After summation, this results in the wanted signal $\hat{s}$ being enhanced while signals from other directions tend to be canceled out. Steering the beamformer is straightforwardly achieved by adjusting the time delays. For a narrowband frequency signal, the time delay can be achieved by applying a phase shift. For a wideband signal the time delays can be applied to each frequency band after decomposition.

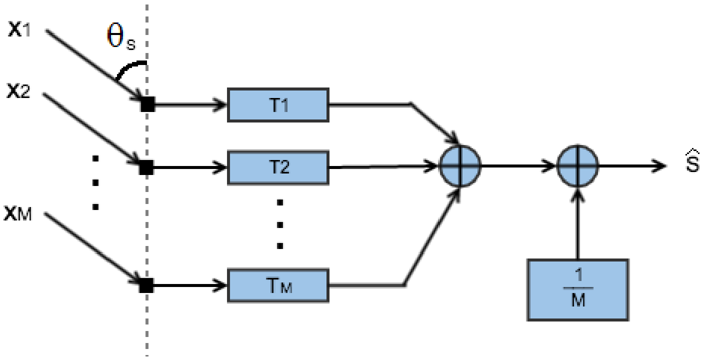


Figure 4.2: Delay-and-sum Beamforming, redrawn from Adel et al. (2012).

The delay-and-sum beamformer has a major limitation, particularly for small arrays. In some unwanted directions, the interference cannot be adequately rejected using a fixed beam pattern (Feng and Jones, 2006). This problem has been addressed using more sophisticated techniques such as adaptive beamforming, of which linearly constrained minimum-variance (LCMV) is one example (Capon, 1969). In general, the goal is to minimise the average energy contributed by the interferers. The coefficients in a LCMV system are adjusted such that signals from the target direction pass unmodified and the amplitude of the interfering signals from other directions is minimised. These coefficients can be optimised for a particular target direction.

For narrowband signals, the optimal filter coefficients vector c_{opt}, derived by Capon, can be written as (Feng and Jones, 2006):

$$c_{opt} = \frac{R^{-1}e}{e^{H}R^{-1}e} \tag{4.8}$$

where R^{-1} is the inverse of the cross-correlation matrix between the microphone signals, e is the steering vector containing the relative phase and amplitude information of the source in the target direction, and H indicates the complex- conjugate transpose. Difficulties with this type of beamformer are that it is computationally expensive and sensitive to numerical errors in the matrix inversion.

For wideband signals, (Frost III, 1972) introduced a wideband adaptive beamformer. He applied an adaptive filter to the signals at each microphone, allowing a frequency-dependent response. Due to the gradient operation in this method, however, its main disadvantage is slow convergence. To improve this, Griffiths and Jim (1982) introduced the generalised sidelobe canceller, which is an unconstrained minimisation technique. This method still converges slowly, but is simpler to implement than Frost's method. The beamformer by Griffiths and Jim has proven very popular (Feng and Jones, 2006).

Beamforming using large arrays has demonstrated substantial improvements in speech intelligibility (Kates and Weiss, 1996; Schum, 2003). However, the use of beamforming in a binaural hearing aid is limited by the physical constraints of the number and location of the microphones. To achieve a relatively good performance the number of microphones in the beamforming system must exceed the number of sound sources (Allred, 2006). Furthermore, the performance of a beamformer degrades in reverberation (Greenberg and Zurek, 1992; Ricketts and Hornsby, 2003). Therefore, applying beamforming techniques in a typical binaural hearing aid using only two microphones creates difficulties in dealing with everyday listening environments where there are often more than two sources and these are often affected by reverberation.

4.4 Model-based separation

Another family of solutions to improve speech intelligibility using signal separation is based on machine learning models for speech. A number of approaches have been developed, including hidden Markov models (HMMs) and non-negative matrix factorisation (NMF).

A hidden Markov model possesses temporal continuity, making it suitable for modelling speech. Varga (1990) modelled both speech and a noise signal using an HMM. In a subsequent development by Roweis (2000), an HMM was trained using a narrowband spectrogram to model each talker. After training, these models were combined to form a factorial HMM. The separation of multiple talkers can be achieved by first inferring an underlying state sequence for the factorial HMM, then a binary mask can be constructed using the model outputs. The approach by Roweis is only suitable, however, when the models are trained using the same talkers, which makes them impractical for hearing aid use.

More recent research has improved the performance of separation algorithms based on factorial HMMs. A new probabilistic model, the factorial scaled hidden Markov model (FS-HMM), has been proposed by Ozerov et al. (2009), who applied FS-HMM to a variety of speech separation problems. In a further enhancement, Mysore et al. (2010) describe the non-negative factorial HMM (N-FHMM) for modelling sound mixtures. Hershey et al. (2010) demonstrated that their factorial HMM-based algorithm could outperform a human listener in monaural separation tasks and produced comparable results in a talker recognition challenge. Specifically, their algorithm achieved an overall recognition rate of 21.6%, compared with a human recognition rate of 22.3%. Hershey et al. pointed out that model-based speech separation still has room for improvement to be viable in real-world applications. Sticking points continued to include poor adaptation to unknown talkers and environments (Hershey et al., 2010). These factors have limited the application of model-based methods in hearing aids, where an ability to generalise and accommodate unknown talkers and environments is highly desirable.

4.5 Computational auditory scene analysis

The term auditory scene analysis (ASA) was introduced in section 2.3. ASA refers to the human ability to organise sounds, particularly in the perception of an acoustic mixture. The related phenomenon of selective attention is often exemplified using the cocktail party effect (Cherry, 1953), in which a listener focuses on one sound source within a mixture. Information which facilitates the segregation of a wanted sound source from a mixture includes the monaural and binaural cues embedded in it.

Computational auditory scene analysis (CASA) can be defined as a machine system that emulates the sound source segregation functions of the human hearing system as described in ASA (Rosenthal and Okuno, 1998; Wang and Brown, 2006). It very often attempts to replicate the perceptual and neural mechanisms in ASA to segregate various kinds of sounds. Specifically, CASA systems address the sound segregation problem by using a variety of acoustic cues, such as periodicity (Parsons, 1976; Okuno et al., 1999; Cooke, 2005), onset and offset detection (Brown and Cooke, 1994; Smith and Fraser, 2004; Hu and Wang, 2008), amplitude modulation extraction (Kollmeier and Koch, 1994; Hu and Wang, 2004), frequency modulation (Kumaresan and Rao, 1999; Cooke and Ellis, 2001) and interaural difference (Lyon, 1983; Harding et al., 2006; Mandel et al., 2010b). In addition, sound stream segregation can be used as a front-end for automatic speech recognition (ASR) systems in real-world environments (Okuno et al., 1996).

Broadly speaking, CASA systems can be divided into monaural (one-microphone) and binaural (two-microphone) approaches. Together, they are capable of extracting complementary sets of acoustic cues. The binaural approach in CASA is particularly relevant to our research. Like optimal cue mapping, which we present in Chapter 5, binaural CASA typically combines various cues extracted from a binaural mixture based on a time-frequency mask. Hence, in this section we present a review of the binaural approach within CASA.

4.5.1 Binaural CASA

As described in section 4.1, binaural hearing contributes to sound localisation and auditory robustness in reverberation. This has led a number of researchers to develop binaural CASA systems which are based on two spatially separated microphones. By definition, binaural CASA systems extract binaural cues by comparing the signals arriving at the left and right ears.

In early research carried out by Lyon (1983), a binaural cocktail party processor was proposed. The system localised and separated sound sources based on interaural time difference (ITD) which were estimated by cross-correlating the outputs from a model of a left and right pair of binaural cochlear filterbanks. Bodden (1993) proposed a similar approach which additionally incorporated interaural level difference (ILD). The target source was separated from the mixture by deriving and applying a soft mask comprising the energy ratios between target speech and the mixture in each critical band.

A related anechoic binaural speech separation algorithm described by Roman et al. (2003) uses a skeleton cross-correlogram, in which the local peaks are identified and replaced with a narrower Gaussian function to estimate the sound source location. The skeleton representation helps to avoid the broad peaks in the correlogram which occur at low frequencies. Roman observed that interaural time difference (ITD) and interaural level difference (ILD) distributions for the target source in a mixture display an azimuth-dependent characteristic. In particular, changes in the relative strengths of a target and an interferer source lead to a systematic shift in the estimated ITD and ILD. The target source is segregated by using a mask estimator which has been pre-trained to exploit the ITD and ILD features jointly.

Harding et al. (2006) proposed a soft mask estimator for ASR based on the statistics of binaural interaction. This approach is similar to that of Roman et al. (2003). However, rather than using a parametric method, they estimated probability distributions for ITD, ILD and joint ITD-ILD features from the training data. The system exhibited an enhanced ability, compared with an unprocessed mixture, to recognise a target speech signal in the presence of reverberation.

Much research in binaural CASA has been devoted to combining spatial cues with other cues. In a study of Denbigh and Zhao (1992), harmonicity was used to separate the voiced speech of two talkers. In addition, the use of directional cues in the binaural mixture helped them to determine the allocation of pitch periods to different talkers in each time frame. Similarly, the system proposed by Woods et al. (1996) uses both pitch and localisation cues, this time based on a confidence score derived from the consistency of the cues used. Kollmeier and Koch (1994) also integrated low-frequency envelope modulations and binaural cues. Segregation systems based on harmonic tracking alone perform poorly when the fundamental frequencies of sound sources are very close. Using a residue-driven CASA architecture, Nakatani et al. (1996) and Okuno et al. (1999) used localisation cues to improve both the pitch estimation and the assignment of a pitch to one of two talkers . In more recent research, Jiang et al. (2014) developed a CASA binary classification system for reverberant speech segregation using integrated gammatone frequency cepstral coefficients (GFCCs) and binaural cues.

4.5.2 Binaural CASA structure

In this section, we introduce some of the fundamental ideas associated with most binaural CASA systems. They typically consist of the four stages shown in figure 4.3.

The first stage of a CASA system (see figure 4.3) typically comprises an auditory filterbank, often composed of gammatone filters (Patterson et al., 1987) to perform a time-frequency analysis of the incoming binaural mixture.

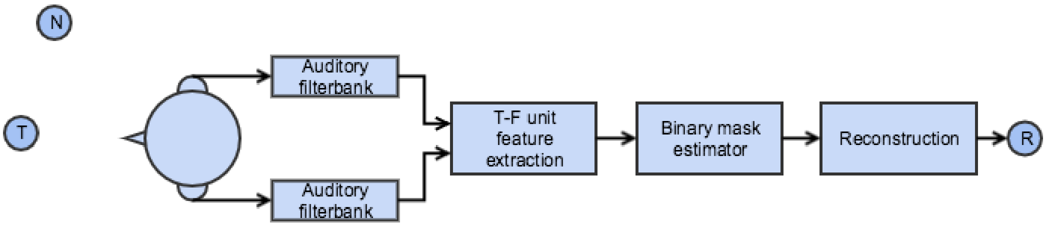


Figure 4.3: Schematic diagram of a typical binaural CASA system. T, N and R denote the target source, noise source (interferer) and the reconstructed target source, respectively. Summarised from Roman et al. (2003) and Jiang et al. (2014).

It is inspired by the frequency selectivity processing which takes place in the human auditory system. The order, bandwidth and frequency spacing of the filters are usually set to match measured psychophysical data (Brown and Wang, 2005). The response of each filter is half-wave rectified and compressed to simulate the neuromechanical transduction in the cochlea (Roman et al., 2003). Not all CASA systems employ auditory periphery processing, however, as the time-frequency analyser can be replaced by the short-time Fourier transform or discrete wavelet transform, for example (Brown and Wang, 2005).

In the second stage, binaural cues are extracted at the level of each time-frequency unit. There are two primary localisation cues: ITD at lower frequencies and ILD at higher frequencies. ITD is typically extracted using cross-correlation. Such a scheme was proposed by Jeffress (1948) and is expressed in equation 4.9. ILD is calculated using equation 4.10.

$$C(t,f,\tau) = \sum_{n=0}^{N-1} h_L(t-n,f)h_R(t-n-\tau,f)w(n) \tag{4.9}$$

$$ILD(t,f) = 10\log 10\frac{\sum_{n=0}^{N-1} h_R(t+n,f)^2}{\sum_{n=0}^{N-1} h_L(t+n,f)^2} \tag{4.10}$$

where $h_L(t,f)$ and $h_R(t,f)$ are the signal outputs from the auditory periphery for left and right channels, respectively, at time t and frequency bin f, and

$w(n)$ is a window function of length N . The ITD can be derived from the time lag corresponding to the position of the peak in $C(t, f, \tau)$ for each time-frequency unit.

Apart from these two binaural cues, many other monaural and binaural cues can be extracted in the second stage. Denbigh and Zhao (1992) proposed a related speech separation system which incorporated information on the fundamental frequency of a speech source. In addition, as described in section 3.1.2, interaural envelope difference (IED) is able to provide direction information at higher frequencies.

The computational goal of CASA is to estimate an ideal time-frequency mask for segregating the target signal in a mixture (Wang and Brown, 2006) and this leads to the third stage of the typical CASA structure. The binary mask estimator usually relies on supervised learning to combine the different features extracted from the input mixture. Binary mask estimation can be seen as a classification problem. In previous research, Roman et al. (2003) applied kernel density estimation, and Woodruff and Wang (2013) employed a set of multilayer perceptrons, whilst Jiang et al. (2014) used deep neural networks to estimate the mask. Once the binary mask has been obtained, the time domain signal for the target source can be reconstructed by inverting the auditory periphery processing.

CASA systems have been reported to perform well in speech segregation tasks, improving both speech intelligibility and quality (Jiang et al., 2014). The implementation of a very recent DNN-based binaural CASA system will be described in detail and compared with our new optimal cue mapping approach in Chapter 8.

4.6 Discussion

The human auditory system is extremely good at segregating and localising sounds in a binaural mixture and this is partly due to the integrated use of both ears in binaural hearing. Hence, an understanding of human listening performance in these conditions is essential in the development of a binaural hearing aid algorithm and has been a focus of this chapter.

Compared with monaural hearing, binaural hearing makes speech more intelligible in both anechoic and reverberant conditions involving multiple sound sources, provided that the target speech is spatially separated from the other sources. Segregation and localisation performance decrease as the number of competing sources increases and the presence of reverberation generally makes the task of source segregation still more difficult. When the number of interference sources becomes sufficiently high, the sources become known as speech babble, which changes the masking effect they create from being predominantly informational into purely energetic masking.

The robustness of speech recognition in background noise is commonly attributed to the glimpsing process in which human listeners detect and group the spectro-temporal regions where the target speech dominates. Low- and mid-frequencies provide more speech intelligibility information than high frequencies and so efforts to perform spatial segregation should be concentrated on the lower regions of the audio spectrum.

Insights gained in this chapter provide the motivation for the two-microphone approach introduced in the next chapter. In Chapter 6 we begin to evaluate the algorithm using the simplest underdetermined configuration; a three-source setup in anechoic conditions. The constraints are gradually released in Chapter 7, where reverberation is also considered in the evaluation. Increasingly realistic conditions are examined in Chapter 8, where the algorithms performance is evaluated using larger numbers of competing speech sources.

A variety of methods for evaluating speech intelligibility and speech quality have been reviewed. They include time-consuming subjective measurements on human listeners and much more rapidly determined objective measurements based on perceptual models. Experiments show that the STOI and PESQ metrics described in sections 4.2 correlate well with their correspoding subjective measurements. In the following chapters, in order to avoid the time and effort involved in evaluating the proposed algorithm subjectively, we adopt these objective evaluation methods for estimating speech intelligibility and quality.

Many machine-based techniques for segregating multiple sources have been briefly reviewed in this chapter. These were most broadly classified as blind source separation techniques, model-based separation systems and computational auditory scene analysis. In terms of function, they are all potentially useful for this research, as a perfectly segregated speech source would maximise speech intelligibility, the prime goal of our work. The model-based speech segregation systems reviewed here are generally too complex to implement, however, or require too much computation and electrical power to be supportable on a miniature, wearable DSP device. Some algorithms possess the fundamental difficulty that they cannot be run in realtime, for example because the segregation solution is arrived at iteratively. To satisfy the requirements of this research, a relatively simple, low complexity, real-time algorithm ultimately capable of operating on a hearing aid device is needed. The binaural CASA approach has been given special attention in this chapter, because it is most closely related to our optimal cue mapping (OCM) technique, which is described in the next chapter.

Chapter 5

Binaural Processing of Multiple Sound Sources: Optimal Cue Mapping

Due to the effects of masking, speech intelligibility can be significantly decreased in the presence of concurrent interfering sounds reaching a listener's ears. These acoustic interference sources make speech perception more challenging for people who suffer from a hearing deficit (Shinn-Cunningham and Best, 2008).

A popular method for segregating a wanted (target) speech source from a mixture of the target and one or more interfering sources (interferers) is to Fourier transform the mixture into the frequency domain to create a vector containing the mixture's spectrum as a set of magnitude and phase values, one pair for each frequency point. The fraction of the energy at each frequency point due to the target alone is estimated and placed in a second vector, known as the mask. Pairwise multiplication of these two vectors, that is, of the mixture spectrum and the mask, produces an estimate for the spectrum of the target. The success of the method depends on the accuracy with which the mask can be estimated. In ideal conditions, by which is meant that the mask can be estimated perfectly, the mask method of source segregation is capable of providing large improvements in intelligibility. A great advantage of the approach is that these improvements are retained even when only the magnitude values of the mask are estimated and then combined with the phase values of the original mixture spectrum.

Methods that estimate the magnitude multiplicative mask in the time and frequency domains show great promise in terms of speech intelligibility improvements. In recent speech perception research, Roman et al. (2003), Li and Loizou (2008) and Jiang et al. (2014) proposed target speech segregation models based on an ideal binary mask (IBM), which is a two-dimensional matrix of binary values with the same form as a spectrogram. When the estimated target speech is dominant in a particular time-frequency (T-F) unit, that mask element is assigned the value unity and zero otherwise. Research has shown that source segregation using a binary mask produces remarkable improvements in terms of speech intelligibility in noisy conditions, for both normal-hearing and hearing-impaired listeners (Li and Loizou, 2008; Wang et al., 2009; Kim et al., 2009). Moreover, listeners can also benefit from binary mask source segregation in reverberant environments (Roman and Woodruff, 2011; Jiang et al., 2014).

As the signal-to-noise ratio (SNR) of the target speech signal reduces, however, the IBM becomes increasingly sparse, causing the intelligibility of the target to deteriorate, albeit gracefully. In low SNR circumstances, it has been shown in automatic speech recognition tasks that the ideal ratio mask (IRM) performs better (Harding et al., 2006; Srinivasan et al., 2006; Narayanan and Wang, 2014). In a ratio mask the mask values may take any value between 0 and 1 and can be thought of as indicating the probability of target dominance for each T-F unit in the mask. In part, because errors in mask estimation are likely to have less of a perceptual impact, an IRM is able to perform better in terms of objective intelligibility than IBM (Narayanan and Wang, 2013). Also, as the SNR of the target deteriorates, ratio mask values diminish, but the mask does not become sparse in the sense that a binary mask does. The intelligibility advantage at low target SNRs of using a ratio mask rather than a binary mask is demonstrated graphically in figure 5.2 in section 5.1.2.

In acoustically challenging environments, listening to speech with both ears improves its intelligibility. Inspired by the advantage of binaural hearing, Roman et al. (2003) use the interaural time difference (ITD) and interaural intensity difference (IID) spatial cues to estimate a binary mask. Moreover, these cues have long been recognised as being important in localisation (Strutt, 1907; Rayleigh, 1907). However, acoustic signals are rich in other cues originating from the sources themselves and from the effects of the acoustic environment. With the assistance of these additional cues, the performance of source segregation can potentially be improved. In considering this possibility, however, it is desirable to be able to determine the usefulness of each cue in different contexts (e.g. anechoic or reverberation conditions), so that weak cues are not integrated, which could waste limited computational resources.

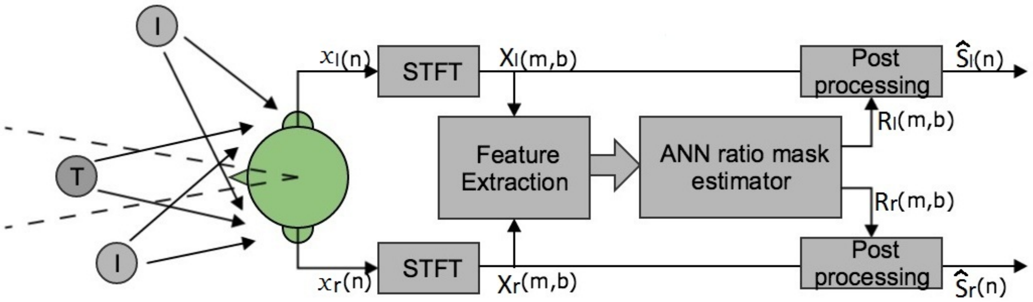


Figure 5.1: Block diagram of the proposed binaural signal processing method. $x_l(n)$ and $x_r(n)$ are the signals recorded at both ears. After applying the STFT, a pair of complex-valued spectra, $X_l(m, b)$ and $X_r(m, b)$, are obtained. Acoustic features are extracted for each T-F unit and form inputs to the ANN, which is trained to predict the energy ratio $R_l(m, b)$ and $R_r(m, b)$ of target speech in the left and right channels, respectively. The estimated target speech signals $\hat{s}_l(n)$ and $\hat{s}_r(n)$ are reconstructed via post processing for presentation to the listener.

In this chapter, the optimal cue mapping (OCM) model is proposed. It is a binaural signal processing strategy based on time-frequency mask estimation, as shown in the figure 5.1. The left and right channels, $x_l(n)$ and $x_r(n)$, respectively, of a binaural mixture signal are received at the two ears and transformed into the time-frequency domain using the short-time Fourier transform. This results in a pair of complex-valued spectra, $X_l(m, b)$ and $X_r(m, b)$, where m and b denote the time frame and frequency band, respectively. Information embedded in these signals is extracted and then data-driven machine learning techniques are used to estimate the ideal ratio mask. We use a set of simple two-layer, feed-forward artificial neural networks (ANNs) to estimate the energy ratio of target speech to target speech plus interference in the binaural mixture. The segregated target signals $\hat{s}_l(n)$ and $\hat{s}_r(n)$ are finally reconstructed by the post processing and presented binaurally to the listener.

By interrogating this machine learning model, the cues extracted from the binaural mixture can be ranked according to their estimated importance. The importance ranking is a measure of the contribution made by one input or by a group of inputs to the estimation of a mask. Depending on factors such as the restricted computational power and limited memory size of the hearing aid, or on the properties of the acoustic environment (e.g. interaural coherence in section 5.2 or other cues that have not addressed in this thesis), the least important input(s) to the ANN can be pruned out. In so doing, the trained system is able to segregate the target signal out of the binaural mixture in an efficient way and provide the best possible intelligibility and quality of a segregated target speech signal in the prevailing conditions.

5.1 Definitions

To ensure rigour and facilitate repeatability, in this section we define a number of processes which will be used when describing optimal cue mapping and when evaluating our results.

5.1.1 Signal-to-noise ratio definition

In our application, signal-to-noise ratio (SNR) is a measure of the total signal energy due to the target speech source compared with the total energy due to any interfering sound sources. The ratio is often expressed in decibels:

$$SNR = 10 \cdot \log 10 \frac{\sum_{n=1}^{M} s^2[n]}{\sum_{n=1}^{M} y^2[n]} \tag{5.1}$$

where M is the length of the signal, s and y are the target speech signal and interfering noise signal, respectively. In order to distinguish it from the refined definition described below, we use the term global SNR to describe equation 5.1.

To measure the SNR of a speech signal using the global SNR is problematic due to the highly non-stationary nature of speech. For instance, consider an utterance and white noise both having a certain energy E over a duration of 1 second. According to equation 5.1, the global SNR is 0 dB. If, on the other hand, the same utterance with the same overall energy, now includes a period of silence, such as a pause in the middle, then its duration increases. The energy E in this signal is spread over a longer period of time and so, compared with a continuous noise signal, will now appear to have a smaller global SNR. The example here illustrates the fact that the actual level of noise presented to achieve a certain global SNR strongly depends on the proportion of silence in the speech samples. For this reason, we apply an energy-based voice activity detector (VAD) in the SNR measurement to remove the silent or quiet portions of the speech before calculating the SNR. We use the term local SNR to describe this calculation, shown in equation 5.2:

$$local\ SNR = 10 \cdot \log 10 \left(\frac{\sum_{p=1}^{P} \left(\sum_{n=1}^{M} s_p^2[n] \cdot VAD_p \right)}{\sum_{p=1}^{P} \left(\sum_{n=1}^{M} y_p^2[n] \cdot VAD_p \right)} \right) \tag{5.2}$$

where P is the total number of frames and VAD_p is zero when the energy in the p-th frame of s_p is 40 dB less than the global maximum or the p-th frame of y_p has zero energy. For all other cases, VAD_p is set to one. In detail, the signal is broken into a contiguous series of fixed-length, overlapping frames. Following the practice adopted by May et al. (2015), and as indicated above, a frame of signal is excluded from the SNR calculation if the energy level of the target source is 40 dB less than the global maximum of the mixture. In this way, the SNR is calculated only in periods of the mixture signal where the target source is active. This leads to a consistent way of defining the SNR, which is independent of the extent of silences or low energy portions in the target speech.

5.1.2 Time-frequency mask

The conventional approach to ratio mask estimation is to transform the temporal frames of the left and right channel waveforms into a domain where the signals are represented more sparsely, which makes the target and interferers easier to separate. The aim is to separate the target talker signal from the interferer signals using spectro-temporal signal processing. In previous research a variety of transforms have been explored (e.g., Hyvärinen and Oja (2000); Akansu and Haddad (2001)). They tend to offer different trade-offs and to provide optimal performance in different circumstances. Their performance is often evaluated using the short-time Fourier transform (STFT) as a baseline. We anticipate that the results of our research will be broadly applicable and largely independent of the transform method and we therefore adopt the STFT to perform signal decomposition.

In reality, the target speech and interference will always arrive at the listener as a binaural mixture. The computational goal of the left and right channel T-F masks is to estimate the target signal-to-interferer energy ratio in each T-F unit. In later chapters of this thesis, we adopt a machine learning approach to performing mask estimation. Training these algorithms requires generation of the IRM (ideal ratio mask), for which perfect knowledge of the target and the interference signals is necessary. Throughout this thesis we are simulating the acoustic environments we investigate. This means that we have individual access to the target and interference signals before they are each binaurally spatialised and mixed. Thus, the IRM can be calculated. By applying the STFT to the target speech signal $t[n]$ and interference $i[n]$, two complex valued spectra $T(m, b)$ and $I(m, b)$ are obtained, where the integers n, m and b denote the time index, the time frame index and frequency bin, respectively. The ideal energy ratio mask R for the target is defined by accessing the target and interference signal before mixing according to:

$$R(m, b) = \frac{\mid T(m, b) \mid^2}{\mid T(m, b) \mid^2 + \mid I(m, b) \mid^2} \tag{5.3}$$

Although we have seen that the ideal ratio mask delivers a superior performance compared with a binary mask in automatic speech recognition (Harding et al., 2006; Srinivasan et al., 2006; Narayanan and Wang, 2014), we will also have need of the equivalent binary mask B as a baseline for comparison of speech intelligibility performance. This can be derived very simply from the ratio mask:

$$B(m, b) = \begin{cases} 1 & R(m, b) > \delta \\ 0 & otherwise \end{cases} \tag{5.4}$$

where δ is an adjustable threshold. δ is set to 0.5 for the results reported in this thesis. At this value, when the energy of the target signal exceeds the energy of the interference then the appropriate T-F unit in the binary mask is set to one, otherwise it is set to zero.

5.1.3 Intelligibility

The fundamental aim of this research is to increase the intelligibility of target speech which is being partially obscured by interfering sound sources and reverberation. The ultimate way to assess the effectiveness of an algorithm

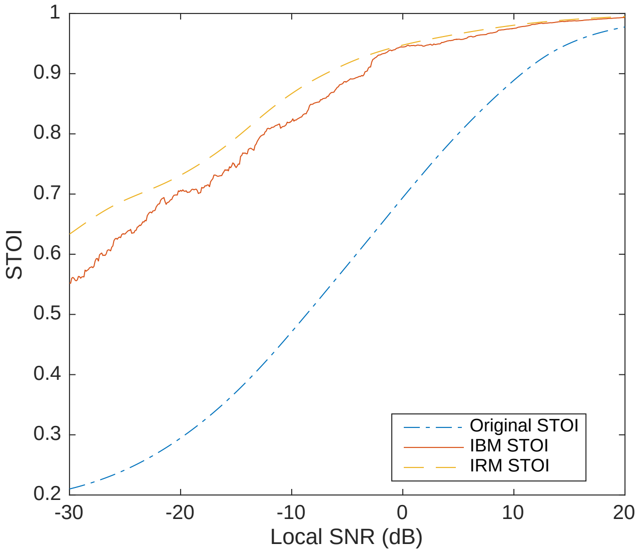


Figure 5.2: STOI performance of the original binaural mixture (left channel) and of the target signal reconstructed using the IBM and IRM, as a function of various local SNRs. Target speech and interference are located at 0° and $-30°$, respectively.

in this respect is to run a listening test, but these are very time-consuming and human-resource intensive. Hence, a number of automated methods have been devised for estimating the intelligibility of speech under various adverse conditions. In our work, we estimate the intelligibility of a variety of unprocessed and processed signals using the short-time objective intelligibility (STOI) metric (Taal et al., 2011), which correlates well with the intelligibility of noisy and T-F weighted noisy speech as measured experimentally using human participants.

Figure 5.2 shows the STOI score for an original binaural mixture signal (left channel) and the STOI scores after applying ideal ratio and binary masks to this. The graphs are plotted as a function of local SNR for a two-source configuration (target speech and one interferer) in anechoic conditions. (A more detailed description of the configuration will be provided in the next section.) We can observe that as the local SNR increases, the original mixture of target speech and interference becomes more intelligible since fewer and fewer T-F units are dominated by the interference signal. After application of the ideal ratio and binary masks, the STOI score becomes much better than the original mixture. The improvement is largest for very low local SNR values. Furthermore, in this ideal case, the IRM performs very similarly to the IBM when the local SNR is above -3 dB. However, the performance gap between IRM and IBM widens at lower local SNR values. This behaviour reflects the fact that the IRM performs better than the IBM, particularly for low SNR conditions where the binary mask becomes sparse. It is also noteworthy that the binary mask STOI score is not as smooth as the other two curves. This stems from the fact that the T-F units are turned on or off using a hard decision process which creates erratic signal reconstruction using the binary mask. For these reasons, our work primarily focuses on estimation of ratio masks rather than binary masks.

5.2 Cue harvesting

In the human auditory system, it is well known that binaural cues play a key role in determining the direction of arrival of a sound source. Differences between the signals at the two ears facilitate the separation of target speech and unwanted speech or noise based on their spatial locations. Principally, these cues are the time difference of arrival of the sound at each ear and the sound pressure level difference between the two ears. Exploiting these cues appropriately leads to improved intelligibility. Therefore, there are clear benefits in applying binaural signal processing methods based on these cues in hearing aids. The OCM (optimal cue mapping) algorithm is designed to improve speech intelligibility by processing the signals arriving at both ears jointly using a variety of cues. These cues can originate in the speech source (e.g. pitch and sibilance), the acoustic environment (e.g. the number of sources and the amount of reverberation) or in the listener (e.g. via their HRTFs).

The main cues used in the OCM algorithm and their statistical features are described in the following sections. They are predominantly binaural and an obvious choice since we are using two microphones in a binaural arrangement.

5.2.1 Statistics of IPD and ILD

We begin by considering the binaural cues used by Roman et al. (2003), namely interaural time difference (ITD) and interaural level difference (ILD). In fact, we use interaural phase difference (IPD), which is closely related to ITD at frequencies below approximately 1 kHz. Above this frequency, IPD values begin to wrap, first at highly lateral azimuths and then at gradually smaller deviations from the median plane as the frequency increases. Roman et al. (2003) derived ITD using cross-correlation, which we consider further in section 5.2.3.

Segregating two spatially separated sound sources in the horizontal plane using a binaural arrangement of two microphones is usually analytically tractable (i.e. a fully determined problem) and almost perfect separation under anechoic conditions is often possible. Nevertheless analysing the statistical characteristics of IPD and ITD when two sound sources located in different directions are simultaneously active provides valuable insights which are of value in more complex acoustic situations.

We begin by generating a binaural mixture of two talkers speaking simultaneously. The subsequent analysis will be based on this example. The source material for both talkers is selected from the TIMIT database (John et al., 1993). The HRTFs used to spatialise the speech are taken from set HA02 in the SYMARE database (Jin et al., 2014). The local SNR is set to 0 dB by scaling both the target and noise signal before spatialisation. The target source is from a male talker located at 0° azimuth, saying, "Don't ask me to carry an oily rag like that" *. The interfering speech is "Doctors

* File name is 'train/dr2/mdss0/sa2.wav'in the TIMIT database

prescribe drugs too freely" *, from a female talker who is located at −30° azimuth. Both sources are in the horizontal plane.

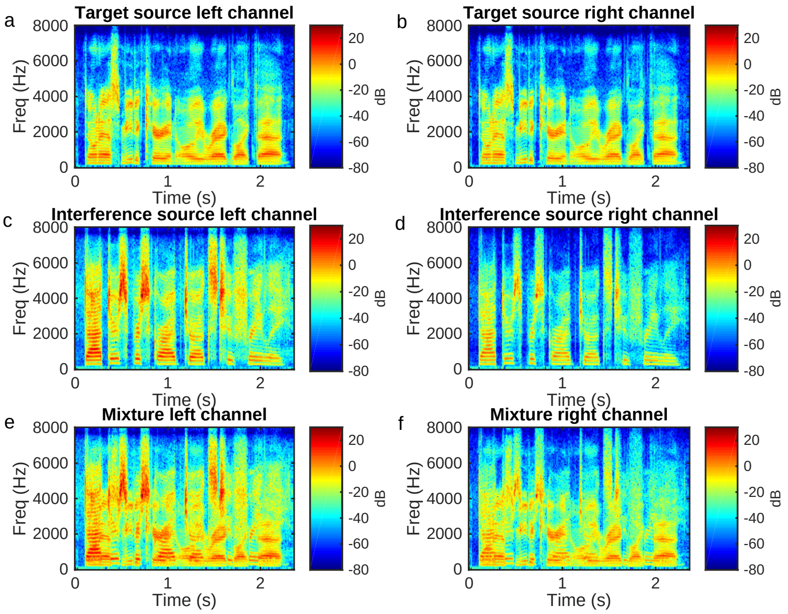


Figure 5.3: Spectrograms of the left and right channels for each source and the mixture in a two-source anechoic configuration. The target source is located at 0° azimuth, and the interference is placed at −30° azimuth.

Figure 5.3 shows the spectrograms of the left and right channels of each individual source and of the mixture. Figures 5.3 (a) and (b) appear the same, which is to be expected, since the level difference between both ears for the target speech is small due to its location directly in front of the listener. However, it is possible to tell from subplots (c) and (d) that the left channel of the interfering speech has more energy than the right which is because the interferer source is closest to the listeners left ear. After mixing the two sources binaurally (subplots (e) and (f)), it is still noticeable that the two channels have different magnitude spectra, especially near the beginning of the mixture sound and in some parts of the spectral region above 4 kHz. Comparison of the spectrograms reveals that there are T-F

[*] File name is 'train/dr1/fkfb0/sx78.wav' in the TIMIT database

units in the mixture where the interference is more energetic than the target and therefore there is the potential for masking to occur. In about the first 0.4 s, it is hard to discern the harmonics of the target speech, whereas they become easily distinguishable later, especially after 2 s.

In order to study IPD and ILD in more detail, figure 5.4 shows the phase difference between the left and right channels in the left-hand column (subplots (a), (c) and (e)) and the level difference subplots ((b), (d) and (f) are in the right-hand column. IPD and ILD are extracted according to:

$$IPD(m,b) = \phi X_l(m,b) - \phi X_r(m,b) \tag{5.5}$$

$$ILD(m,b) = 20 \cdot \log 10 \frac{\mid X_l(m,b) \mid}{\mid X_r(m,b) \mid} \tag{5.6}$$

where ϕ is the unwrapped phase of the spectra X_l and X_r which are the complex-valued spectra for the left and right channels, respectively.

The IPD (subplot(a)) and ILD (subplot(b)) of the target source are close to zero, since the target is positioned symmetrically between the two ears at 0° azimuth. Perhaps due to a small offset in the direction of the head during the measurement of the HRTFs, or because of a slight asymmetry in the shape of the listeners head, there is a constant small delay between the signals arriving at the ears from the target. This manifests itself in subplot (a) as a phase change which increases linearly with increasing frequency. The IPD of the interfering source at $-30°$ azimuth, on the other hand, changes much more rapidly with frequency and cyclically completes a 2π shift approximately every 3.3 kHz. This frequency corresponds to a period of 300 μs, which relates this particular IPD response to an equivalent near-constant ITD for the interferer of this value. The ILD at low frequencies is small due to diffraction of the interfering sound around the head. At higher frequencies, above approximately 4 kHz, head shadow effects become stronger and subplot (d) exhibits greater energy differences. For every frequency band, because the directions of the two sources are stable, the IPD and ILD values are constant as a function of time. In the mixture IPD plot (figure 5.4 (d)),

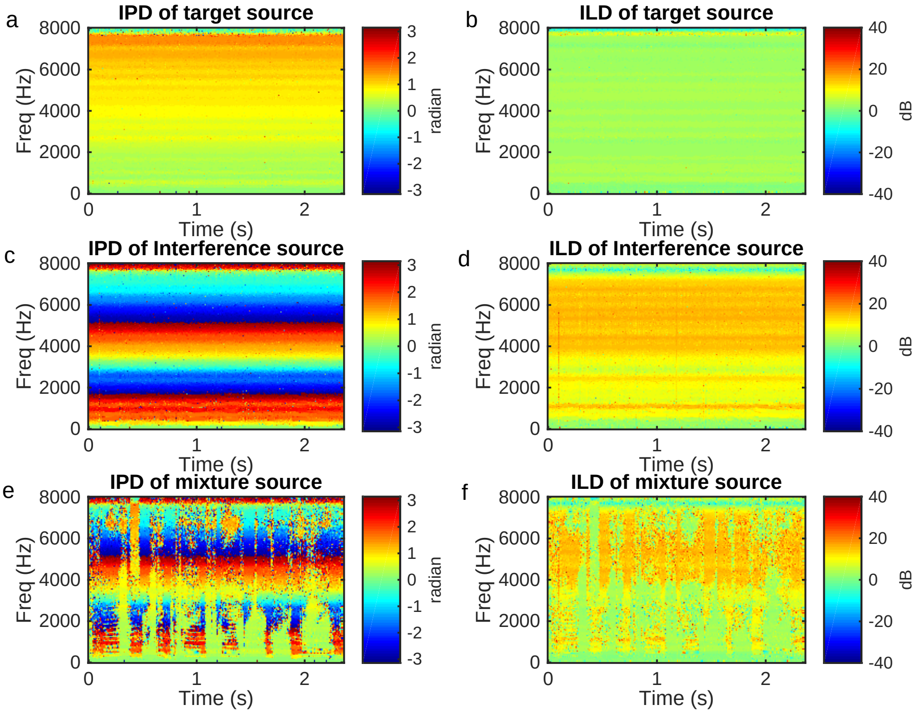


Figure 5.4: Spectrograms of IPD and ILD for each source and for the mixture using a two source anechoic configuration. The target source is located at 0° azimuth, and the interference is located at $-30°$ azimuth.

the regions where the IPD is close to zero belong to the target and are easily distinguishable from the the non-zero regions caused by the interference. Similarly, the ILDs of both sources are well reflected in the ILD plot for the mixture signal. The regions with approximately zero ILD values are associated with the target and the remaining regions derive predominantly from the interference source. Generally speaking, the regions with IPD and ILD values close to zero can be confidently marked as target source active and dominant.

Based on the above observations, we see that IPD and ILD are strong indicators of whether or not a T-F unit is dominated by the target. Roman et al. (2003) demonstrated that as the energy ratio R between the target and interfering source changes there is a corresponding systematic shift in the possible values for ITD and ILD. Specifically, reducing values for the ITD and ILD indicate an increasing probability that the target source is active and dominant. Conversely, a decrease in the value of ratio R results in the ITD and ILD spreading away from target-only values before converging on the values corresponding to the interferer alone. In a particular frequency bin, the ITD and ILD display a certain location-dependent statistical distribution corresponding the location of each source.

In order to study the relationship between IPD, ILD and energy ratio R, we extract IPDs and ILDs from 5,000 frames of binaural mixtures for the same two-source configuration. We also calculate the ideal energy ratio R between the target and interferer signals, since each source is individually accessible before mixing. Figure 5.5 and 5.6 demonstrate the relationships between the values of IPD, ILD and R. We obtain similar results to those described by Roman et al. (2003), who use an auditory filterbank for signal decomposition rather than the STFT.

Figure 5.5 (a) shows the scatter plot for the distribution of IPD with respect to R, as well as the mean and standard deviation of R. Subplot (c) is the histogram of the IPD values at the frequency bin corresponding to 1 kHz, created from the 5,000 frames of binaural mixture. Similar plots for ILD values at a frequency of 3 kHz are shown in figures 5.5 (b) and (d). It is worth noting that when IPD/ILD values cluster close to zero, R is close to unity. Similarly, when R decreases towards zero, IPD/ILD values shift towards their interference-only values. Therefore, in this two-source configuration, there are two peaks in the histograms and these correspond to the IPD/ILD target-only values and interferer-only values for this particular frequency bin.

Furthermore, to reveal the IPD and ILD statistical properties in a joint space, we extract both from the 5,000 frames of binaural mixture at 2 kHz to demonstrate them. There are 20 bins for IPD from $-\pi$ to π, and 20 bins for ILD from -20 dB to 20 dB, resulting in a 20 by 20 grid. Figure 5.6 shows the histograms of IPD and ILD samples in the joint space at 2 kHz. Each peak in

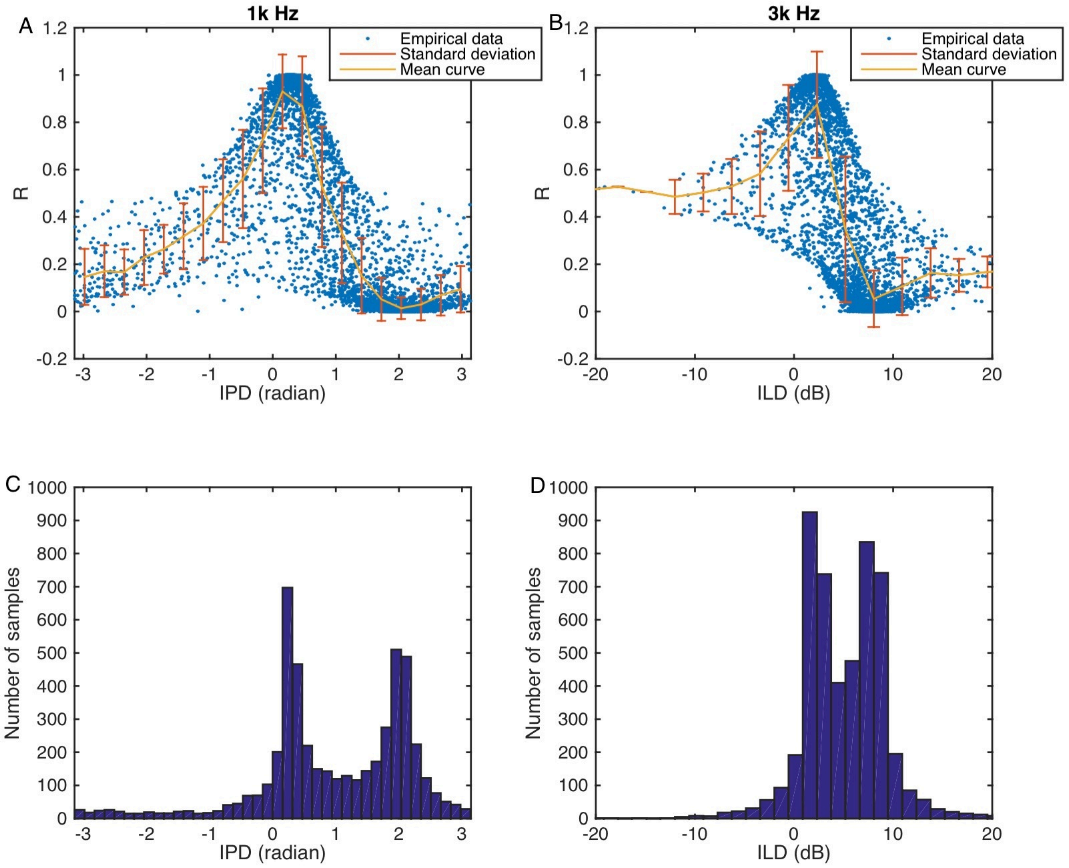


Figure 5.5: Statistical properties of 5,000 empirical data values for IPD and ILD as a function of energy ratio R in anechoic conditions. The target and interference are located at 0° and $-30°$, respectively. (a) The scatter plot of IPD against R showing the mean and standard deviation of R at 1 kHz. (b) The scatter plot of ILD against R with the mean and standard deviation of R at 3 kHz. (c) Histogram showing frequency of occurrence of IPD values at 1 kHz. (D) Histogram showing frequency of occurrence of ILD at 3 kHz.

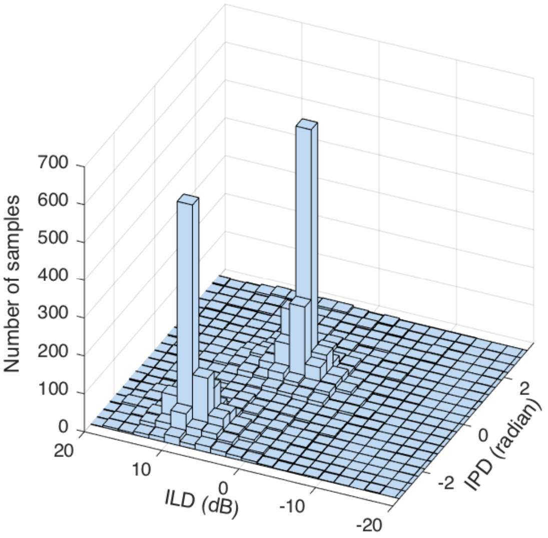


Figure 5.6: Histogram of IPD and ILD samples in the 20 by 20 joint grid space, generated from 5,000 frames of binaural mixture at 2 kHz in anechoic conditions. The empirical data are obtained with target and interferer located at 0° and $-30°$ azimuth, respectively.

the histogram represents a distinct active source coming from one direction. The target source generates a peak in the combined frequency of occurrence for IPD and ILD close to zero in both dimensions. The other peak occurs where IPD and ILD values are close to the interference-only value.

Based on the observations shown in figure 5.5, the mapping relationships between IPD/ILD and R have been revealed. In addition, IPD and ILD have characteristic distributions in which the peaks indicate the number and location of active sources, particularly apparent in figure 5.6. This analysis demonstrates that IPD and ILD are rich sources of information for estimating R in anechoic environments. Therefore, they are strong candidates for forming inputs to an algorithm for estimating the ratio mask to segregate a target speech source from one or more interfering sources in a binaural mixture. In later chapters these binaural cues do indeed form two of the inputs to our proposed optimal cue mapping algorithm.

5.2.2 Statistics of ΔIPD and ΔILD

In the previous section, we considered a simultaneously active target and interferer. Individually, these have stable IPDs (figure 5.4 (a) and (c)) and stable ILDs (figure 5.4 (b) and (d)). When these sources are combined, however, the resultant IPD and ILD are unstable (figure 5.4 (e) and (f)). In this section we look at the properties of this unstable behaviour and at the cues it carries to assist in estimating the ratio mask, R.

In an early study (Gao and Tew, 2015), delta features (denoted by Δ) of IPD and ILD are found to be useful in estimating the mask. A delta feature measures the difference between the current value of a property and the value in the previous frame, also known as the rate of change. In this section, we investigate why delta IPD and ILD can help in the estimation of the energy ratio R and explore their relationship to it.

The definition of ΔIPD and ΔILD are given in equations 5.7 and 5.8.

$$\Delta IPD(m,b) = IPD(m,b) - IPD(m-1,b) \tag{5.7}$$

$$\Delta ILD(m,b) = ILD(m,b) - ILD(m-1,b) \tag{5.8}$$

where integers m and b denote the time frame index and frequency bin, respectively. As discussed in section 5.2.1 and shown in figure 5.4, the IPD and ILD for each individual source varies relatively smoothly across frequency. Once in the mixture, the IPD and ILD tend to be either close to the target-only value or close to the interferer-only value.

Consider a three-source configuration, for example, where the target is located at 0° azimuth and the two interfering sources are placed symmetrically on both sides at $-30°$ and $30°$ azimuth, respectively. In this configuration, IPD and ILD values will be close to zero when the target source only is active (or dominant), and either positive or negative when either the left or the right interferers is active, respectively. Hence, even when only the two

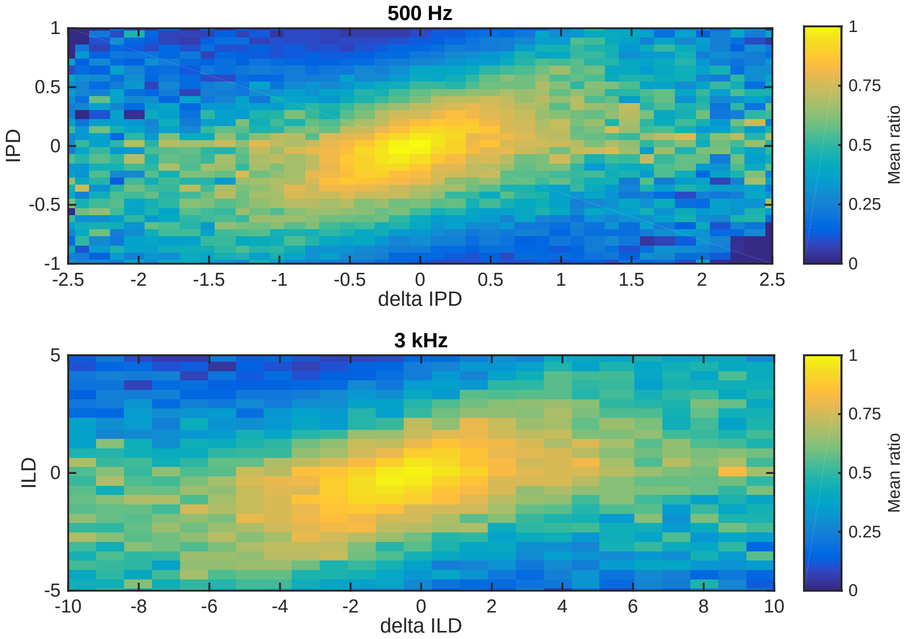


Figure 5.7: The relationship between ΔIPD and ΔILD and ratio mask value (indicated by the colour) particularly for small values of IPD and ILD at 500 Hz and 3 kHz, respectively.

interferers are active and the target is inactive, the IPD and ILD values of the binaural mixture will shift between positive and negative values and so inevitably pass through zero as they tend to shift towards the value of the currently dominant interferer.

Pursuing this further, when both IPD and ΔIPD are stable and approximately zero, it can be seen that the ratio R is close to unity, signifying the case that the target speech alone is active and is located at 0 azimuth. On the other hand, if instead two interferers are active, and the target is inactive, the IPD will vary erratically between positive and negative values, and will occasionally be close to zero. However, ΔIPD will generally be non-zero in this situation and the ratio R will be zero. In this scenario therefore, when IPD is close to zero, the value of ΔIPD is a strong indicator of the mask value and can be considered important. A similar argument applies to ΔILD.

Further analysis of figure 5.7 reveals that a small ΔIPD/ΔILD value particularly influences the mask value when the associated IPD/ILD value is small. The mask value is close to unity when both IPD and ILD and both of their deltas are zero, indicating that only the target is active; for other values of delta feature, the ratio mask value shifts towards to zero, indicating that multiple sources are active or, in the extreme condition, that the target source is actually inactive. In this way, not only are the importance of ΔIPD and ΔILD revealed, but also their role in enhancing the estimate of the mask.

5.2.3 Cross-correlation

This research aims to segregate a target sound source from spatially distinct interfering sources. Being able to determine the directions of the direct sounds arriving at the listener is potentially extremely valuable for estimating the proportion of energy in each T-F unit due to the desired target speech. ITD has been shown to be one of the most important cues for localising a source (Blauert, 1997). For this reason, many researchers have integrated ITD into their algorithms to perform localisation (Roman et al., 2003; Harding et al., 2006; Woodruff, 2012; Ma et al., 2015b).

Cross-correlation measures the similarity between two signals as a function of the lag τ of one relative to the other. For signals with a sufficiently narrow cross-correlation function it can be used to estimate the time difference of arrival of a sound received by two spatially separated microphones, as in the case of a binaural pair of hearing aids. In this section we consider basic approaches for estimating time difference of arrival and build up to the system which we go on later to apply in our optimal cue mapping method.

In section 3.1.1 it was shown that ITD can be approximately calculated by the Woodworth formula. Figure 5.8 shows ITD plotted as a function of azimuth for HRTF set HA02 in the SYMARE database (Jin et al., 2014). Points on the curve are calculated by cross-correlating the HRIR left/right pairs at each measured azimuth direction in the database. The time lag

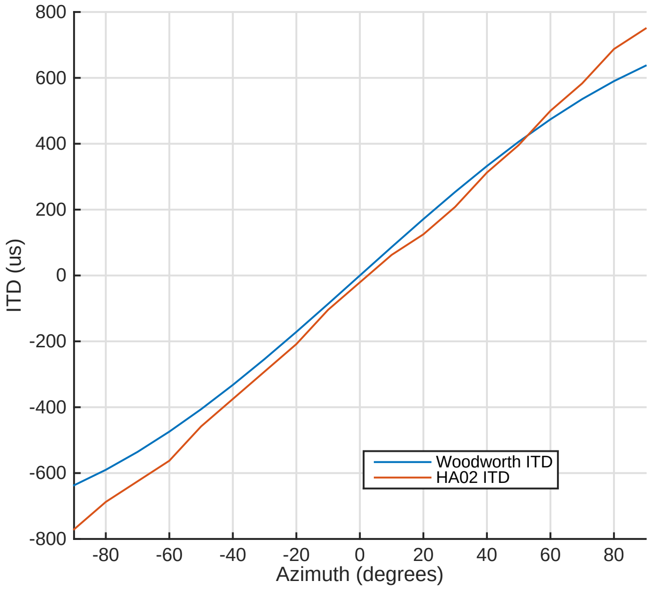


Figure 5.8: Comparison of ITD for θ between $-90°$ and $90°$, $\phi = 0°$, from set HA02 in the SYMARE database, with the Woodworth approximation for $r = 0.085\,\mathrm{m}$ and $c = 343\,\mathrm{m/s}$.

corresponding to the maximum value in the cross-correlation function is used to determine the ITD between two signals. For comparison, the Woodworth approximation is also plotted in figure 5.8. The head radius parameter in the model has been adjusted to create a good fit with the measured ITD. The graph demonstrates the near-linear relationship between ITD and the azimuth angle of a sound source and confirms the usefulness of this parameter in the localisation process. The figure also shows the excellent fit between the simplistic Woodworth formula and an ITD obtained through measurement of a real HRIR.

One weakness of estimating ITD by cross-correlating HRIR pairs is that it generates a frequency-independent result, when it has been shown that in reality ITD varies as a function of frequency (Kuhn, 1977). We accommodate this frequency dependency by calculating the normalised cross-correlation function within each frequency bin b of a gammatone filterbank (Roman et al., 2003). Although all the features described in the previous sections can be extracted from binaural mixtures just based on gammatone filterbanks, at this stage we aim to keep the spectral bandwidth as narrow as possible at higher frequencies, which can be provided by the STFT. Hence, we combine the two signal decomposition techniques by allocating the appropriate STFT frequency points to each gammatone filter band. The bandwidths of the gammatone filterbank adhere to the equivalent rectangular bandwidth (ERB) concept (see section 2.2) promulgated by Glasberg and Moore (1990). The ERB bandwidth calculation is repeated here. For each frequency point b in the STFT:

$$ERB(b) = 24.7(4.37fc(b) + 1) \tag{5.9}$$

where centre frequency $fc(b) = fs \cdot b/N$ (kHz), fs is the sampling frequency and N is the number of points in each STFT frame. Using this information the impulse response of the associated 4^{th} order gammatone filter is calculated for every frequency point according to:

$$g(n,b) = \begin{cases} n^3 e^{-2\pi ERB(b)n} cos(2\pi f_c n) & if n \geq 0 \\ 0 & otherwise \end{cases} \tag{5.10}$$

The gammatone-filtered ear signals are denoted as $l(m,b)$ and $r(m,b)$ for left and right ear, respectively. Then the normalised cross-correlation function (CCF) between the two channels can be described as:

$$CCF(m,b,\tau) = \frac{\sum\limits_n (l(m,b)[n] - \overline{l(m,b)})(r(m,b)[n-\tau] - \overline{r(m,b)})}{\sqrt{\sum\limits_n (l(m,b)[n] - \overline{l(m,b)})^2} \sqrt{\sum\limits_n (r(m,b)[n-\tau] - \overline{r(m,b)})^2}} \tag{5.11}$$

where τ is limited in range to encompass the natural range of ITD values, approximately -800 μs to 800 μs, and n is the sample index in the time-domain filtered signal. The bars denote the mean value. A sampling frequency of 16 kHz results in a cross-correlation output vector of length 27 for each T-F unit. The 3D structure of cross-correlation output values for all frequencies and all time frames is called the cross-correlogram (Roman et al., 2003). Then the overall time delay in the mth frame $SCCF(m,\tau)$ can be estimated by summing $CCF(m,b,\tau)$ across all frequency points, as shown in equation 5.12

$$SCCF(m,\tau) = \sum_{b} CCF(m,b,\tau) \tag{5.12}$$

which is the form used in (Roman et al., 2003). We include the expression for the mean time delay here for completeness and it is not actually used in our OCM algorithm. As explained above, time difference of arrival varies with frequency and so we compute it on a per frequency point basis according to equation 5.11.

To visualise the time difference of arrival information in the cross-correlation function between the left and right input channels, the signal being analysed here is the example mixture used in section 5.2.1. The mixture contains two talkers, the target and an interferer located at 0° and −30° azimuth, respectively, in anechoic conditions. Figure 5.9 (a) and (b) show the cross correlation vector values for each source for all frequency bands over a duration of 1.4 s. The cross-correlograms summed across all frequencies within the same time frame are presented in subplots (c) and (d) and the cross-correlograms for each signal are shown in subplots (e) and (f).

The time differences for the target signal in all frequency bands are close to zero, as indicated by the time lag for which the peaks in each frequency band are aligned in figure 5.9 (a). This alignment creates the peak at 0 s observed in the summation shown in figure 5.9 (c). It is clear that the horizontal line running along the time dimension in figure 5.9 (e) indicates that the time difference between the two channels is, as expected, about 0 s throughout the whole duration of the target signal. A similar pattern applies for the plot of the interference signal in the right column of figure 5.9. The time delay between the two channels is now about -300 μs, which is similar to the time difference for the azimuth angle −30° in figure 5.8. Since the interference is on the left side of the listener, we use a negative time delay to signify that the sound reaches the left ear sooner than the

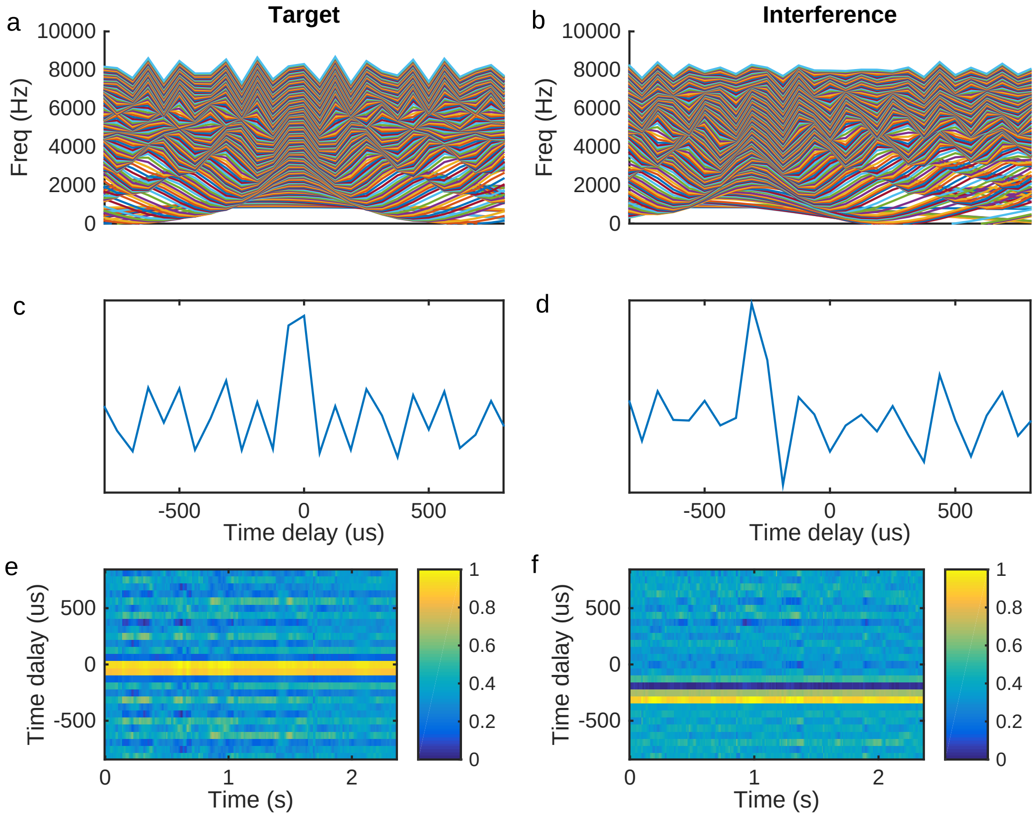


Figure 5.9: Example of source localisation methods for target and interferer sources, respectively. (a) Cross-correlation coefficients for the target speech within each frequency band at the time of 1.4 s. (b) Cross-correlation coefficients for the interferer speech within each frequency band at the time of 1.4 s. (c) Summed cross-correlation coefficients across all frequency points for the target. (d) Summed cross-correlation coefficients across all frequency points for the interferer. (e) Cross-correlogram for the whole time duration of the target. (f) Cross-correlogram for the whole time duration of the interferer.

right ear. Multiple peaks can be seen in the high frequency channels for both target and interferer. This ambiguity is due to the shorter acoustic wavelength of signal components at higher frequencies creating periodicities. Psychophysical models generally apply envelopes to the responses in this frequency range to overcome the ambiguity (Blauert, 1997). In this study, we currently do not employ cross-correlation envelopes. Note that the time resolution of the computed time differences is limited by the sampling interval. For finer temporal resolution, the time differences could be interpolated.

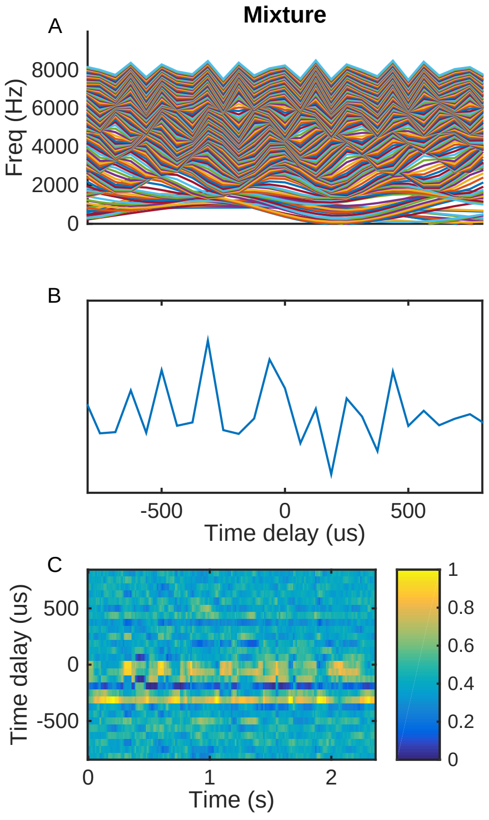


Figure 5.10: Example of source localisation for the mixture. (a) Cross-correlation coefficients within each frequency band at the time of 1.4 s. (b) Summed cross-correlation coefficients across all frequency points. (c) Cross-correlogram for the whole time duration.

A similar set of plots, for the binaural mixture of the target and interferer combined, is shown in figure 5.10. Two broken horizontal lines are now apparent in the cross-correlogram in subplot (c), representing the time difference of arrivals for the two sources in the figure. As expected, the line relating to the target is located at time differences close to 0 s and the line for the interferer lies at approximately -300 μs. For frequency channels where either the target or the interferer is much stronger than the other source, the observed peak lies closer to the source's true time difference of arrival. When both sources have similar energy levels, two principal maxima are observed indicating that multiple sources are active simultaneously, although other peaks also occur, creating ambiguity.

Because the cross-correlation function has the potential to estimate source direction, we include it as a further input to our optimal cue mapping method. Since our method estimates ratio mask values for every frequency point, the cross-correlation coefficients for each frequency point are evaluated, instead of using their summation across all frequencies.

5.2.4 Interaural coherence

So far we have only considered cue extraction in anechoic conditions. In many realistic situations, direct sounds are contaminated by room reflections. A room impulse response (RIR) evolves from the deterministic direct sound and early reflections through to the stochastic reverberant tail. The varying statistical properties of the reverberant sound require different treatments to reduce their effect on intelligibility.

In the preceding sections, we have considered a series of cues that create potentially useful dissimilarities between a signal arriving at the left and right channels of a binaural hearing aid. We now look at a process which approaches the problem from another perspective. Many researchers have demonstrated the usefulness of a metric known as interaural coherence (IC) for improving estimates of ITD and ILD in reverberant conditions (e.g., Faller and Merimaa (2004a); Jeub et al. (2010a); Alinaghi (2013)).

Interaural coherence measures the similarity between two signals. It can be described as:

$$IC(m,b) = \frac{E_{l,r}(m,b)}{\sqrt{E_{l,l}(m,b)E_{r,r}(m,b)}} \tag{5.13}$$

where $E_{l,l}(m)$ and $E_{r,r}(m)$ denote the recursive auto-power spectral densities (APSD) for each T-F unit in the left and right channels, respectively, and $E_{l,r}(m)$ is the recursive cross-power spectral density (CPSD) of two time-aligned channels. They are calculated by means of the recursive relation:

$$E_{l,l}(m,b) = \alpha E_{l,l}(m-1,b) + (1-\alpha) \mid X_l(m,b) \mid^2 \tag{5.14}$$

$$E_{l,r}(m,b) = \alpha E_{l,r}(m-1,b) + (1-\alpha) X_l(m,b) X_r^*(m,b) \tag{5.15}$$

where α is the smoothing factor with values in the interval [0,1]. $X_l(m,b)$ and $X_r(m,b)$ are the Fourier-transformed signal in the left and right channels of the mth time frame and bth frequency bin.

Equation 5.13 generates IC values between 0 and 1. Perfectly coherent left and right channels yield the IC value 1. The value of the IC drops as the behaviour of a T-F unit from one frame to the next becomes increasingly dominated by reverberation and other (generally unwanted) chaotic influences. This is illustrated in the figure 5.11. Subplots (a) and (c) show spectrograms of a speech signal in anechoic and reverberant conditions, respectively. Here, the binaural room impulse response, obtained from the AIR database (Jeub et al., 2009), is measured in a stairway hall and the speech source is placed at 0° azmuth. The IC is almost 1 in the anechoic condition (figure 5.11 (b)), since the left and right channels are similar with approximately zero time difference in arrival time. Figure 5.11 (c) shows the same speech source spatialised within the reverberant environment. Temporal features have become smeared by the approximately exponential decay of the room impulse response. The original patterns of energy variation in the spectrogram of the anechoic target can be recognised in the plot of IC in reverberation in figure 5.11 (d). This is because the regions of late reverberation produce IC values that shift away from 1 due to the fact that the late reflections come from multiple directions with different times of arrival in each ear. This incoherence occurs in both channels, and the red circle shows a clear example of how high values of IC tend to pick out the T-F units which contain relatively high energy direct sound whereas reverberant tails tend to produce low IC values.

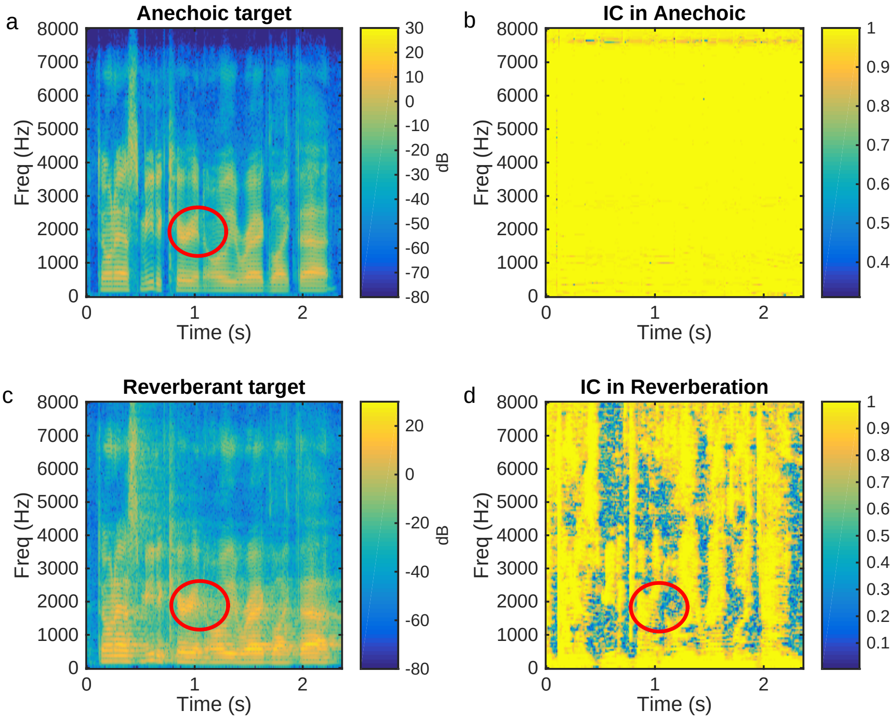


Figure 5.11: Comparison between the spectrograms of some anechoic and reverberant speech (left channel shown only) and their corresponding IC patterns. (a) Spectrogram of anechoic speech. (b) Interaural coherence of the anechoic speech. (c) Spectrogram of the same speech in reverberation. (d) Interaural coherence of the reverberant speech.

The reliability of estimated IPDs and ILDs in reverberation is controlled by setting a threshold for the IC value. IPD and ILD values for T-F units with an IC above a suitable threshold indicate a reliable estimate (Faller and Merimaa, 2004b). It is necessary to have a different threshold for each frequency band. This is due to the high variation in the characteristics of reverberation with frequency, most notably the reverberation time (Jeub et al., 2009). Finding a suitable set of thresholds is further complicated by the large differences in the reverberant properties of different rooms.

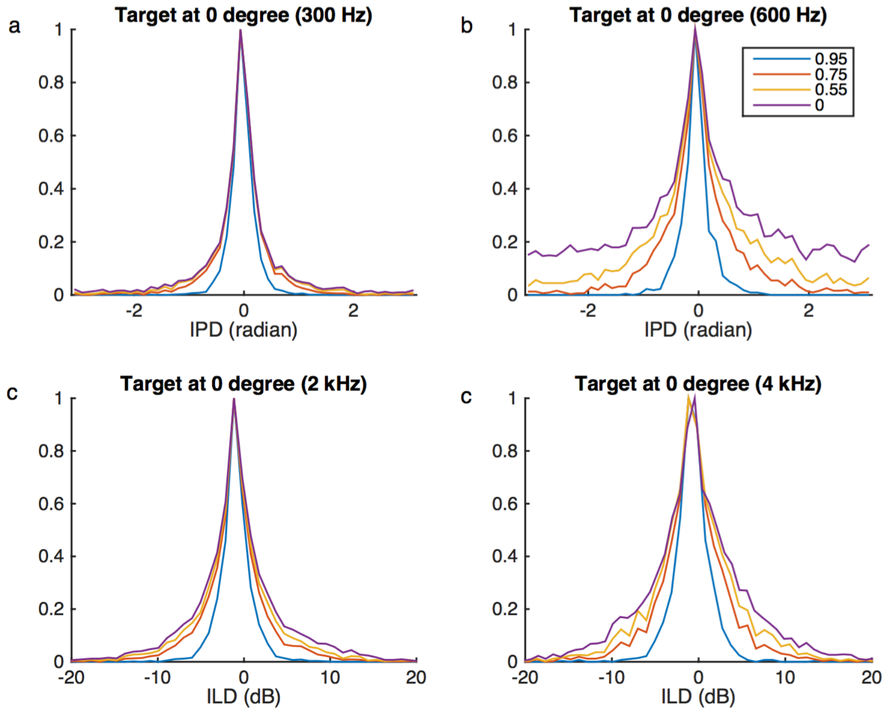


Figure 5.12: The envelopes of normalised histograms showing the distributions of selected IPDs and ILDs in reverberation for four IC threshold values. The speech source is at 0° azimuth. (a) and (b) IPD distributions at 300 Hz and 600 Hz, respectively, (c) and (d) ILD distributions at 2 kHz and 4 kHz, respectively.

To reveal the relationship between binaural cue estimation accuracy in reverberation and IC threshold, we generate 5,000 observations for target speech at 0°. Again, the binaural room impulse response measured in a stairway hall has been used from the AIR database (Jeub et al., 2009). Figure 5.12 (a) plots the envelopes of normalised histograms showing the relative frequency of occurrence of IPD values for four different IC threshold values at 300 Hz. Subplot (b) shows the result of the same analysis at 600 Hz. ILDs at 2 kHz and 4 kHz are analysed in a similar way in figures 5.12 (c) and (d). Setting the IC threshold to zero means that all the IPDs or ILDs are selected and contribute to the histograms. The majority of IPDs or ILDs

	IC threshold			
	0.95	0.75	0.55	0
IPD at 300 Hz	0.0711	0.2562	0.4203	0.6132
IPD at 600 Hz	0.0784	0.5841	1.1848	2.0512
ILD at 2 kHz	4.1286	15.9367	22.9968	32.3888
ILD at 4 kHz	4.7197	17.8085	25.7629	39.6461

Table 5.1: Variance of selected cues for different IC threshold values and different frequencies for speech in reverberation at 0° azimuth.

cluster around zero since the speech source lies in front of the listener. As the threshold is raised the distribution tends to compress and become more sharply peaked. Raising the threshold also has the effect of selecting fewer cues, though this is obscured by the normalisation applied in these plots. The variances of the selected cues corresponding to figure 5.12 are shown in table 5.1. Higher IC threshold values create smaller cue variances. For the same threshold value, the variance of the selected cues increases with increasing frequency. Therefore, in order to achieve the same variance across frequency, a frequency-dependent IC threshold is needed.

	IC threshold			
	0.95	0.75	0.55	0
IPD at 300 Hz	0.1042	0.3873	0.7007	1.1464
IPD at 600 Hz	0.118	0.9796	0.7007	2.5664
ILD at 2 kHz	3.9837	20.6285	32.4524	47.6133
ILD at 4 kHz	5.5505	17.3251	27.1951	40.6805

Table 5.2: Variance of selected cues for different IC threshold values and different frequencies for speech in reverberation at $-30°$ azimuth.

Altering the azimuth angle of the source also affects the shape of the distribution created using a particular threshold. The details are presented in figure 5.13 and table 5.2 for a sound source located at $-30°$ azimuth. In the figure, it can be seen that the expected shift in the peaks of the distributions has occurred, corresponding to the change in source direction. However, the distributions are not as symmetrical as when the source is located at 0° azimuth. Compared to the variances of the selected cues in table 5.1, when

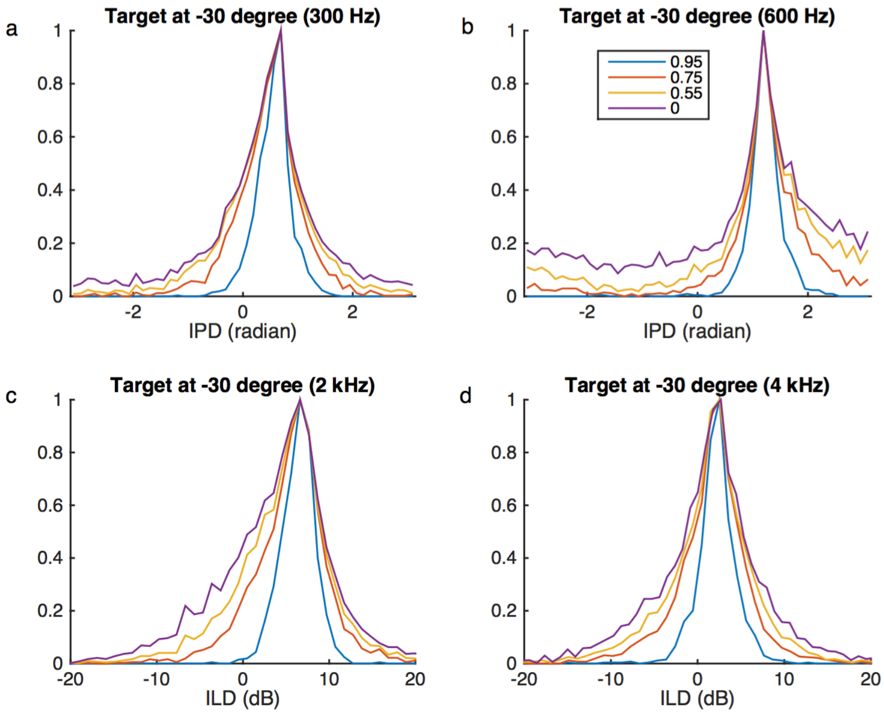


Figure 5.13: The envelopes of normalised histograms showing the distributions of selected IPDs and ILDs in reverberation for four IC threshold values. The speech source is at −30° azimuth. (a) and (b) IPD distributions at 300 Hz and 600 Hz, respectively, (c) and (d) ILD distributions at 2 kHz and 4 kHz, respectively.

the source lies at 0°, the variances of the selected cues in table 5.2 tend to be larger for a given IC threshold value. The variation of the direct-to-reverberant energy ratio (DRR) at each ear for different azimuth angles leads to this direction-dependent behaviour in the variances of IPD and ILD.

It can be concluded from this analysis that the estimation of IPD and ILD cues in reverberation can be further improved with the assistance of interaural coherence when an appropriate threshold is applied. Since IPD and ILD are powerful cues for ratio mask estimation, the inclusion of IC into our optimal cue mapping algorithm can be expected to improve its ability to segregate target speech in reverberant conditions.

5.3 Neural network-based mask estimator

The main aim of this research is to improve estimation of the binaural ratio mask for segregating a target sound source in the presence of a spatially separate interfering mixture of sources and reverberation. In addition, the main novel contribution of this work is the systematic integration of many cues for improving the estimate in an efficient way. One approach for doing this is to apply machine learning techniques. In this section we demonstrate the method using a conventional and relatively simple type of artificial neural network (ANN). A simple ANN has been used so that attention can be focused on the selection and integration of the input cues.

5.3.1 Artificial neural network topology

Every artificial neural network (ANN) requires input data, either in the form of training data or of test data. Training data includes input (or observation) data and also the corresponding expected output (or label) data. The training data provides the ANN with examples which are used to adapt the weights of its internal connections so that its output gradually produces better and better estimates of the correct output. Training includes testing how well the ANN is learning the relationship between input and output. This usually involves presenting the ANN with previously unseen data and comparing its output with correct results. A trained network is able to fulfil certain tasks, such as mapping or classification.

Prior to training an ANN, the input and expected output data need to be transformed so that they possess, as closely as possible, a normal distribution (Akansu and Haddad, 2001). This results in changes in input values causing a similarly sized change at the output. To standardise the measurement scales of the inputs, they are transformed to possess zero mean and exhibit a spread in values of one standard deviation (standard score or z-score) using the equation:

$$input_n = \frac{x_n - \hat{x_n}}{\sigma_n} \tag{5.16}$$

where $input_n$ is the standardised nth input, x_n is the original observed feature input data, $\hat{x_n}$ and σ_n denote the mean and standard deviation of the input x_n, respectively.

The nonlinear transfer functions used to build the ANN are sigmoids. In order to satisfy the demands of the transfer function, the expected output has to be constrained to the range [0 1]. Due to the nature of the ratio mask in our application, its values are already located in this range, hence no transformation is needed for the expected output in this study.

For the training stage, the learning method adopted here utilises the back-propagation algorithm (Fausett, 1994). The label data is the desired output from neural network and is presented together with the corresponding observation data. The learning algorithm continually updates the weights and biases in the ANN after evaluating a new observation and computing the difference between each actual output and the label.

Figure 5.14 shows the topology of a simple ANN. The number of neurons (or nodes) in the input layer of the ANN is determined once the shape of the training data is known. Generally, the number of neurons is equal to the dimension (or length) of the input features. Taking the features IPD and ILD as an example, the number of neurons required is therefore two, as shown in figure 5.14. Further design details are given in the experimental work described in Chapter 6 and 7.

In a similar way to the input layer, the number of neurons in the output layer equals the number of outputs. The ratio mask R is the computational goal in this experiment, so there is a single node in the output layer for each ANN and each one calculates the ratio for one frequency point. The ANN in figure 5.14 is an example of a simple network architecture with 3 layers. The inputs to the ANN are normalised IPD and ILD values for the frequency corresponding to the network. The hidden neurons are labelled A, B and C and there are only three in order to display the architecture clearly. In a practical network the number of hidden neurons is generally much larger than three. The input-hidden-output neuron connection weights are denoted by W. Subscripts are used to indicate the origin and destination of each connection (e.g. W2B signifies that the connection comes from input 2, the ILD input, and goes to hidden layer node B)

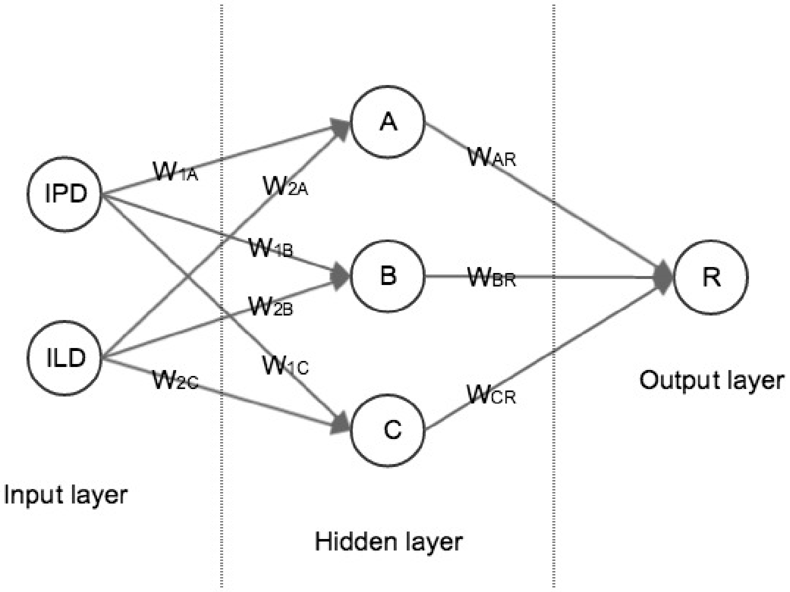


Figure 5.14: Example of a simple three-layer ANN architecture. It has two input neurons, three hidden neurons and one output neuron.

5.3.2 Network optimisation

In any application of neural networks for the mapping or classification of data, the number of hidden layers and the number of hidden neurons needs to be decided. There exists no simple method for determining this. Early research by Irie and Miyake (1988) and Hornik et al. (1989) has shown that a three-layer (one hidden layer) feed-forward network with an arbitrarily large number of nodes is a universal function approximator. Recently, Gao and Tew (2015) demonstrated for our application that an ANN with one hidden layer has the ability to learn the mapping between various acoustic cue inputs and appropriate ratio mask output values. Therefore, we continue to use one hidden layer in this research. With regard to defining the number of neurons in the hidden layer, the strategy we adopt is an exhaustive search. Although this approach is computationally highly expensive it allows the optimal network topology to be determined with a high degree of confidence.

To estimate neural network generalisation error and determine the optimum number of neurons in the hidden layers for the ANN at each frequency point, we employ the cross-validation method developed by Weiss and Kulikowski (1991) and Plutowski et al. (1994). In N-fold cross-validation, the training data is split into N subsets of approximately equal size. The neural network is trained using $N-1$ subsets of the data and is subsequently evaluated using the remaining one subset. This procedure is repeated N times, on each occasion using a different subset for the evaluation. Therefore, each evaluation uses a unique subset of training data and it explains why the method is also known as leave-one-out cross-validation (Kohavi et al., 1995). The average performance of the N neural networks is an indication of their ability to generalise (that is, to provide reliable estimates for previously unseen input data). We use 10-fold cross-validation in this study and the performance of each ANN is measured by averaging the mean square errors (MSEs) between the ideal ratio mask values and the estimated ratios.

5.4 Cue importance ranking

Identifying cues which have the potential to improve estimates of the ratio mask is the first step in creating an effective solution to the speech segregation problem. However, efficiency is also crucial in the envisaged application of this research in hearing aids, which have very limited computational resources. For this reason it is helpful to rank the cues in importance, so that unnecessary computation is avoided by excluding inputs which do little or nothing to improve the estimation of the binaural ratio mask.

Up to this point, we have identified a variety of features (or acoustic cues) in the binaural mixture which are potentially useful for estimating the ratio mask. They include IPD, ILD and their deltas, cross-correlation, and interaural coherence. In this section, these inputs are presented as inputs to ANNs and the networks are trained to estimate ratio mask values. The trained ANNs are then analysed to identify the degree to which these features contribute to the mask estimation by the network. The goal is to reveal the relative importance of each input feature. Here we employ the connection weights method (Olden et al., 2004) and Garson's method (Garson, 1991) to analyse each of the neural network inputs. With reference to the demonstrator ANN in figure 5.14 the analysis proceeds as follows:

- 1: For each input neuron, measure the contribution of input neuron to output neuron via each hidden neuron by calculating the product of the weights along every connection path.

  $$c(i,j) = w(i,j)w(j,R) \tag{5.17}$$

  where $i \in \{1,2\}$ denotes either the IPD input neuron ($i = 1$) or ILD input neuron ($i = 2$), and $j \in \{A,B,C\}$ denotes the index of a hidden layer neuron. R denotes the output neuron.

- 2: Sum the products $c(i,j)$ across all the hidden neurons for each input type to create the overall connection weights $P(i)$:

  $$P(i) = \sum_j c(i,j) \tag{5.18}$$

- 3: Measure the relative contribution $r(i,j)$ of each input feature to the output layer via each hidden neuron:

  $$r(i,j) = \frac{|c(i,j)|}{\sum\limits_i |c(i,j)|} \tag{5.19}$$

and find the sum $S(i)$ of contribution of each input neuron:

$$S(i) = \sum_{j} r(i,j) \tag{5.20}$$

- 4: Garson's relative importance metric RI for input feature i is then calculated by:

$$RI(i) = \frac{S(i)}{\sum\limits_{i} S(i)} \tag{5.21}$$

The connection weights method is composed of steps 1 and 2. The overall connection weight indicates the relative importance (Olden et al., 2004). Steps 1 to 4 constitute Garson's algorithm. It is noteworthy that Garson's algorithm does not retain the sense of the relative contribution of each feature since the absolute value is used in step 3. The relative importance estimate resulting from both methods will be discussed in the next chapter.

5.5 Post-processing

A trained network yields ratio mask estimates for a single frequency in either the left or the right channel. Hence, the estimated spectrum of the target signal, $\hat{T}_{l|r}(m,b)$, is obtained by multiplying the coefficients of mixture $M_{l|r}(m,b)$ by the estimated ratio mask $E\hat{R}M_{l|r}(m,b)$:

$$|\hat{T}_l(m,b)| = |M_l(m,b)| E\hat{R}M_l(m,b) \tag{5.22}$$

$$|\hat{T}_r(m,b)| = |M_r(m,b)| E\hat{R}M_r(m,b) \tag{5.23}$$

For each time frame, the complex spectrum of target signal $T_{l|r}(m,b)$ can be calculated by combining the phase of the original mixture with the modified magnitude spectrum:

$$\hat{T}_{l|r}(m,b) = |\hat{T}_{l|r}(m,b)|e^{j\omega\angle M_{l|r}(m,b)} \tag{5.24}$$

This complex spectrum can be converted back to the time domain using an N-point inverse discrete Fourier transform (IDFT). Finally, the segregated target signal is synthesised using the overlap-and-add method with a Hann window.

An alternative way to synthesise the segregated target signal is to calculate the N-point IDFT of the ratio mask, $erm_{l|r}(m,b)$, and convolve the result with the N time-domain mixture samples in the frame, $m_{l|r}(m,b)$.

$$\hat{t}_l(m,b) = m_l(m,b) \bigotimes erm_l(m,b) \tag{5.25}$$

$$\hat{t}_r(m,b) = m_r(m,b) \bigotimes erm_r(m,b) \tag{5.26}$$

where $\bigotimes$ denotes convolution operation. This produces $2N-1$ output samples and preserves the convolution tail correctly. The overlap-and-add process can then be applied without further windowing.

5.6 Summary

This chapter outlines optimal cue mapping (OCM), which is a novel signal processing algorithm intended to improve the intelligibility of a target speech source in the presence of multiple interfering sounds in anechoic and reverberant conditions. This is to be achieved by exploiting binaural cues to segregate the wanted speech from the unwanted interference. The principle novelty of the proposed algorithm has two aspects. Firstly, it is a relatively simple algorithm based on a conventional ANN architecture, in contrast to more modern machine learning algorithms, such as the deep learning neural network (see Chapter 8). Only information in the previous and current time frames are required. This means that the algorithm can potentially be run in real-time, which is a necessity in a hearing aid. Secondly, a way to understand the contribution of diverse cues for estimating the mask is provided. In this way, cues can be integrated dynamically, depending on the acoustic conditions. This facilitates the removal of cues which do not contribute to the mask estimation process, potentially increasing algorithm efficiency still more.

It has been demonstrated that spectro-temporal binary masks can achieve substantial intelligibility improvements for both normal-hearing and hearing-impaired listeners in adverse conditions (Li and Loizou, 2008; Wang et al., 2009; Kim et al., 2009; Roman and Woodruff, 2011). However, the ideal binary mask with a fixed threshold becomes increasingly sparse in low SNR situations, resulting in the deterioration of intelligibility. Recent research into the use of a ratio mask approach has been demonstrated to improve performance in automatic speech recognition tasks (Harding et al., 2006; Srinivasan et al., 2006; Narayanan and Wang, 2014). Furthermore, the ratio mask has been shown to yield greater intelligibility improvements than a binary mask in similar conditions (Gao and Tew, 2015).

In this initial work, the cues are acoustic in nature and predominantly binaural. Spatial cues are extracted from a multi-source binaural mixture. We demonstrate how IPD and ILD are two powerful cues for indicating the direction of a dominant sound source in the mixture, and we show that there is a strong relationship between these cues and the proportion of target energy in each T-F unit, the precursor to estimating a ratio mask. Extra cues, ΔIPD and ΔILD, are also extracted and we explain how these assist the mask estimation process. The cross-correlation function yields a series of coefficients which we demonstrate are rich in information about the direction of multiple sources in a binaural mixture. Hence, we also plan to integrate these into the ratio mask estimation algorithm. In reverberant conditions, this chapter suggests how interaural coherence can help identify T-F units which contain reliable IPD and ILD values. In addition, by analysing the importance of each type of cue in the estimation, we gain deeper insight into how these cues should be integrated into the system. The importance analysis also provides the opportunity for further optimisation in terms of satisfying the computational limitations of different applications with the minimum loss of source segregation performance.

In optimal cue mapping, ANNs are trained to estimate the spectral energy fraction of a wanted speech source at each frequency point in the input mixture. Once trained, the ANN outputs form a spectral ratio mask which is applied frame-by-frame to the mixture to approximate the magnitude spectrum of the wanted speech. Due to careful cue selection, OCM is a potentially real-time binaural method, since it only uses information from the current frame of mixture samples, such as IPD and ILD cues, and information from the preceding frame, e.g. to calculate ΔIPD and ΔILD values. Therefore, it is potentially feasible to implement OCM in time critical applications, such as in hearing aids.

The extent to which OCM can provide binaural unmasking and intelligibility improvements for a target speech source in a mixture needs to be investigated. Experiments to discover its effectiveness form the subject of the next chapter.

Chapter 6

Pilot Study

The main research aim in this project is to improve the speech intelligibility and quality of speech in complex binaural sound mixtures. As mentioned in previous chapters, the ideal binary mask (IBM) has been shown to be capable of producing intelligible speech in adverse noisy conditions (Li and Loizou, 2008; Wang et al., 2009; Kim et al., 2009; Roman and Woodruff, 2011; Jiang et al., 2014). In an IBM, the time-frequency (T-F) units are classified as either being totally associated with the target signal or with the interferer signal by assigning each unit with the value 1 or 0, respectively. Hence this formulation can be thought of as a classification problem. A more complicated form of T-F mask is the ratio mask which was proposed by Barker et al. (2000). In the T-F unit of an ideal ratio mask (IRM), it is the proportion of energy attributable to the target in the mixture which is stored (see section 5.1.2). In this case, the value of the T-F unit lies on a continuum between 0 and 1. Evaluating a ratio mask turns the binary mask classification problem into a probability estimation exercise. In addition, it has been shown that the IRM performs particularly well in automatic speech recognition tasks (Harding et al., 2006; Srinivasan et al., 2006).

In Chapter 5, the optimal cue mapping (OCM) algorithm, based on machine learning, was introduced. The computational goal of this algorithm is to estimate the ideal ratio mask that makes best use of the available cues whatever the acoustic environment. In this chapter, a pilot study is described based on the OCM algorithm. A viable source segregation system is systematically built up by utilising a list of acoustic cues (see section 5.2) extracted from the binaural input mixture.

The OCM algorithm has the potential to assist in establishing the relative importance of acoustic spatial cues, of properties of the acoustic space and of sound source features in the binaural mixture. This process was described in section 5.4. Hence, in this chapter we also analyse the importance of each cue that we consider for inclusion in the mask estimation process and present the results.

Some of the work presented in this chapter has previously been published by Gao and Tew (2015).

6.1 Experimental setup

In this development of the OCM algorithm we assume that the target speech source lies within a few degrees of 0° azimuth, i.e. in front of the listener. This is based on the anecdotal observation that a listener tends to look at the talker with whom they are conversing as a way of reducing listening effort through reading the talkers lips and body language. It also serves as a way of showing respect and of indicating attentiveness. We consider this assumption to be a reasonable starting point, although we recognise that there are many occasions in real life when it is invalid. With appropriate training, the machine learning systems at the heart of the OCM method can be adapted to work with the target sound source in other directions. The scope of this research does not, however, extend to exploring methods for identifying and tracking the target and so we restrict ourselves here to proving the principle of optimal cue mapping for a single target source direction.

With a two-microphone array in most anechoic conditions, it is theoretically possible to separate two point sources of sound which lie in different directions. This is an example of an overdetermined system (i.e. the number of microphones is equal to or exceeds the number of sound sources). It may arise, for example, out of doors when listening to a target talker in the presence of one interfering conversation. The situation changes, however, when there are two interfering conversations, which typically will introduce two interfering talkers speaking at the same time as the target talker. This resulting binaural mixture is now underdetermined. In this case it is no longer possible to separate the target talker analytically using linear filtering of two microphone signals because there is no longer enough information in the two microphone signals to be able to reverse the mixing process.

With appropriate training of the ANNs used in OCM, there is no hard constraint on the number of interferers and their directions. Variable-direction interferer setups in both anechoic and reverberant conditions are described in Chapter 7 and yet more configurations are used in Chapter 8. In Chapter 8, underdetermined configurations with up to five sources (including the target speech) are evaluated and compared to one state-of-the-art signal processing approach using deep learning neural networks.

In this pilot study, we first train and test an OCM system for a simple underdetermined case in anechoic conditions. Thus, the training and test mixtures that are evaluated in this chapter are a three-source configuration which contains two interfering talkers and one target talker. Specifically, the two interferers are positioned asymmetrically either side of the listener to simulate a plausible natural situation. However, as stated in our early work, reported in (Gao and Tew, 2015), one of our initial aims was to compare the performance of the OCM method with work reported by Roman et al. (2003). Hence, in this section we inherit their setup, which employs a symmetrical arrangement of two interfering sound sources located at azimuth angles of $-30°$ and $30°$, respectively. As previously discussed, the target is placed at $0°$ azimuth.

Signal analysis is based throughout on the short-time Fourier transform (STFT). The STFT frame duration is set to 20 ms, which corresponds to 320 samples using a sampling frequency of 16 kHz. This sampling frequency was determined by the sampling frequency of the speech corpus employed (John et al., 1993) for creating the binaural training and test mixtures. Each frame is Hann windowed using 50% overlap with its neighbours. Each STFT frame of the binaural input mixture therefore consists of 161 frequencies, from 0 Hz up to and including the Nyquist frequency at 8 kHz.

In the OCM algorithm, to exploit the cues listed in section 5.2, input data for each ANN is generated from two T-F units of the same frequency, one drawn from the current frame and the other from the preceding frame. For training purposes, as well as generating the input features for each ANN, the corresponding ideal energy ratio value is calculated for segregating the target speech at the ANNs frequency. The method for computing the ideal ratio mask is described in section 5.1.2. There is a left channel and right channel pair of ANNs for each frequency and hence there are 322 ANNs in total.

All the ANNs used in the various experimental configurations are trained on the York Advanced Research Computing Cluster (YARCC) (Smith, 2015). The full cluster consists of 28 nodes, 58 processors and 528 cores. Depending on the level of competing demand for resources, we can employ up to half of the cluster at any given time. It takes under 24 hours to train any of the sets of ANNs described in this chapter on a single core or cluster. The training time is dependent on the computing speed of the different cores.

6.1.1 Training and testing setup

Speech for training is randomly selected from the training set in the TIMIT database (John et al., 1993). Speech for testing purposes is randomly selected from the database's test set. Therefore, there is no overlap between the two sources of material and so testing of the ANNs is conducted using previously unseen speech clips. For the anechoic simulated conditions, we use the HRIRs of SYMARE subject HA02 (Guillon and Zolfaghari, 2012) to spatialise the sound sources in the mixture. Note, there is no restriction on the choice of HRIRs. For example, we have also used the HRIRs for subject HA01 in the SYMARE database and obtained similar results. Varying the elevation of the sound sources is not currently considered in this research. Hence all the sources are placed in the horizontal plane at an elevation angle of zero degrees. The HRIRs used in this chapter were originally sampled at 48 kHz. Hence, we down-sample the HRIRs to 16 kHz to match the sampling rate of the speech signals.

To simulate the condition that people are talking with approximately the same sound intensity, the training mixtures are adjusted to have 0 dB local SNR (section 5.1.1) before they are spatialised. This is considered to be a worst case scenario, since in reality, individual interfering talkers will generally be at a somewhat lower level than the target. Convolving the sound sources by their respective pairs of HRTFs will affect their original local SNRs and create an imbalance of levels between the left and right channels for lateral sources. Rather than train the ANNs with a fixed local SNR in the left or the right channel, no further attempt is made to control the local SNRs in the training and test mixtures, on the basis that these SNRs vary in a similar way under natural acoustic conditions. Indeed, a property which we explore later is the ability of the mask estimator to segregate the target speech successfully when the local SNR is deliberately altered.

Irrespective of the experimental configuration, we generate 5,000 training data items for each distinct training case (examples of different training cases include varying the number of active sources and/or varying the direction of the active sources). It is important to make sure the training data is sufficient to ensure that the ANN weights have adapted. Therefore, we also train the ANNs using double-sized training data. If this produces similar results in the subsequent testing phase (within the limits of the experimental noise floor) then this confirms that 5,000 items is sufficient for training. If further adaptation is observed using the larger training set, then the checking process is repeated with a larger number of items in both sets. Thus, 5,000 training data items is chosen as a starting point and the number is increased if necessary.

It is also important to ensure that sufficient test data is employed. To demonstrate that, we compare the results of using differently sized test mixtures. We find that increasing the size of the test mixture from 50 test items up to 100 items reduces the mean-square output error from the ANNs by less than 2%, which we consider to be a sufficiently small improvement to allow the use of 50 items in each test case. Further support for this decision comes from Jiang et al. (2014), who also set the size of their TIMIT test mixtures to 50.

6.1.2 Evaluation metrics

The OCM approach aims to improve the intelligibility of speech in adverse multi-talker conditions. Initially we model an anechoic environment and subsequently extend the problem to include room reverberation. We evaluate the OCM algorithm in respect of two properties of speech: speech intelligibility and speech quality and discuss each in turn in this section. The detailed evaluation methods for speech enhancement systems have been discussed in section 4.2.

Experimentally, perhaps the simplest way to measure speech intelligibility is as a rate of success for correctly identifying spoken words read from a list (e.g. Kalikow et al. (1977)). All perceptual testing involving human participants is, however, potentially time consuming. This has motivated the development of automated methods for estimating speech intelligibility. These can be a valuable and efficient indicator of performance before time and effort is committed to confirming the results by experiment.

Taal et al. (2011) propose the short-time objective intelligibility (STOI) metric. This has been shown to correlate well with the intelligibility of speech in noise and T-F weighted noisy speech. The STOI model yields scores between 0 and 1, where a higher score indicates higher intelligibility. Therefore, we adopt the STOI metric to evaluate the performance of the OCM method in terms of intelligibility.

The second property we consider is speech quality, which reflects the realism and naturalness of the speech. Increasing the quality of the speech does not necessarily lead to a rise in its intelligibility (Gold et al., 2011). Generally, highly intelligible speech exhibits a subjectively good speech quality and vice versa. However, highly intelligible speech also can be of low quality and *vice versa*. Ramírez and Górriz (2011) report that an improvement in speech quality can reduce listener fatigue and, even for this reason alone, speech quality is an important consideration in speech processing. Again, to avoid the time and effort involved in measuring the speech quality of the speech produced by the algorithms investigated in this research, we turn to an automated method for estimating this attribute. The perceptual evaluation of speech quality (PESQ) metric described by Rix et al. (2001) has been shown to correspond well with subjective speech quality scores for speech separation and enhancement systems (Yi and Loizou, 2008). We therefore adopt this metric in our evaluations. The PESQ metric returns scores which range between -0.5 and 4.5, where higher scores suggest better perceptual speech quality.

6.1.3 Definitions

Before starting to describe the pilot study, the short labels for ANN topologies and mask estimators that will be used in this chapter are summarised in table 6.1. All the systems are trained at an SNR of 0 dB.

Table 6.1: System definitions and training details for each mask set.

Interferer configuration	HRIR	System	Mask set
Two interferers in fixed	SYMARE	Ideal	IBM IRM
directions −30° and 30°	HA02 Anechoic	F2 2-inputs	EBMF2 ERMF2
Two interferers in fixed directions −30° and 30°	SYMARE HA02 Anechoic	F6 6-inputs	EBMF6 ERMF6
Two interferers in fixed directions −30° and 30°	SYMARE HA02 Anechoic	F7 7-inputs	EBMF7 ERMF7

6.1.4 Feature selection for system F2

As described in section 3.1, IPD and ILD are two primary binaural cues for sound localisation in the horizontal plane. Hence we first train OCM system F2 using as inputs IPD and ILD features only. These features are similar to the ITD and IID features employed by Roman et al. (2003). The reason for the use of slightly altered cues is discussed in section 5.2.1. Here, we use EBMF2 and ERMF2 to denote the estimated binary and ratio mask sets created by OCM system F2 for the two inputs, IPD and ILD, only. As described in section 5.1.2, the binary mask is derived by quantising the ratio mask to two levels, 0 and 1, with the threshold set to 0.5. Thus, the same ANN is used in both cases and only its output is modified, depending on whether a ratio mask or a binary mask is desired. The performance of the binary mask using the inputs IPD and ILD will form the baseline with which more sophisticated systems will be compared.

6.1.5 Additional feature selection for systems F6 and F7

In section 5.2, several other spatial cues and binaural features are introduced. These extra cues are potential candidates for improving the accuracy of mask estimation in various conditions. Based on system F2 with two primary localisation cues, we add four further inputs in a new system, F6. These cues are ΔIPD, ΔILD, magnitude and interaural coherence (IC). Here, EBMF6 and ERMF6 denote the binary and ratio masks, respectively, which are produced using system F6. Furthermore, system F7 has one more extra input which is a set of cross-correlation coefficients. EBMF7 and ERMF7 denote the corresponding binary and ratio masks.

We assess the relative importance of each input in three systems, F2, F6 and F7. Of particular interest is how well the extra inputs have been integrated into the estimation process. The role of all these cues are described below.

ΔIPD and ΔILD

The delta features of IPD and ILD, ΔIPD and ΔILD, measure the difference between the current value of that feature and its value in the previous time frame. In section 5.2.2 we investigated the impact of these delta features on the ratio mask and concluded that they will be useful in mask estimation. We therefore incorporate inputs ΔIPD and ΔILD into the enhanced system, F6.

Magnitude

The magnitude input is simply the absolute value of the current T-F unit at the frequency associated with the ANN. It therefore indicates the level of the mixture signal at this frequency. Its primary effect on mask estimation is anticipated to occur when both the target and the interferer signal levels at this frequency are low. In this situation the ratio is subject to increased error due to the influence of the noise floor and is therefore set to zero, since there is no target signal to be segregated.

Interaural coherence

As a test of the OCM method, we include IC as the sixth and final input to system F6. IC has been demonstrated to be an important cue in many source separation algorithms involving dereverbveration (Westermann et al., 2013; Alinaghi, 2013). Therefore, we anticipate that IC will contribute little to the estimation of the mask values in anechoic conditions and that we will observe a rise in its importance in the presence of reverberation.

Cross-correlation

In section 5.2.3 cross-correlation coefficients were identified as a potentially useful cue for segregating sources based on their direction. This feature is likely to be most useful for sources with variable direction, whereas the directions of the interferers are known and fixed at $-30°$ and $30°$ in this experiment. For this reason the coefficients have not been included in the current system. In addition, cross-correlation coefficients are formed into a feature vector containing 27 coefficients. This sets it apart from the other input features which only consist of 1 element. Cross-correlation coefficients will be used, however, in a later simulation (see section 6.3), where the directions of the interferers are allowed to vary and it will become apparent that the multi-element nature of the feature requires special treatment during analysis.

6.2 Comparison of OCM performance for F2 and F6

In this chapter, as a starting point in the development of ecologically valid solutions for the speech segregation problem, we concentrate on entirely anechoic simulations. Furthermore, as stated in section 6.1, we begin in this section by considering a three-source configuration, which is one of the simplest underdetermined cases which can be addressed. The target is placed directly in front of the virtual listener at 0° azimuth and the two interference sources are located at $-30°$ and $30°$ azimuth, respectively. The main aim of this section is to train and compare the performance of two OCM systems.

6.2.1 ANN topologies for F2 and F6

In order to compare the performance of the masks generated by systems F2 and F6, it is important to ensure that the underlying ANNs reach a near-minimum level of mean-square error (MSE) during the training phase. In so doing, they will produce close to the best possible mask estimation results. We follow the procedure previously described in section 5.3.2 to determine the number of neurons in the hidden layer of each ANN. The goal is to achieve a sufficiently low MSE using the fewest possible neurons.

In defining the ANN topologies for F2 and F6, 16 frequency points are selected out of the 161 linearly-spaced frequencies available between 0 Hz and the Nyquist frequency (see section 6.1). The 16 frequency points are chosen to be approximately equally spaced on the equivalent rectangular bandwidth frequency scale to create an analysis of ANN performance across the audio bandwidth. Under this configuration of three sound sources with fixed-directions, the number of neurons is defined using the 10-fold cross-validation method described in section 5.3.2. The resulting MSEs for F2 and F6 are plotted in figure 6.1 as a function of the number of hidden units in the

ANNs. The 2D graphs in subplots (c) and (d) include a dashed line showing the size of the hidden layer which has been chosen.

It can be seen from figures 6.1 (a) and (c) that the minimum MSEs for F2 vary across frequency, although they all follow a similar trajectory. For the sake of clarity, the 2D projections of subplots (a) and (b) are shown in figures 6.1 (c) and (d), respectively. For all 16 frequency points, substantial reductions in MSE are observable up to 10 neurons. The performance is fairly stable between 10 and 20 neurons, beyond which there is no further improvement. Indeed, the MSEs tend to increase after 30 neurons, which is caused by overfitting. In the 10-fold cross-validation method, ANNs are trained using 9 subsets of training data and tested using only one. As the number of neurons in the ANN increases, the ANNs become better at learning the training data and weaker in their ability to generalise. Hence, in the test phase, larger ANNs yield greater MSEs when presented with the previously unseen validation data. ANNs with 15 neurons exhibit a MSE which is within 0.1% of the asymptotic value and hence this network topology is used in system F2 with inputs IPD and ILD only. In figures 6.1 (b) and (d) a similar pattern is observed for system F6 and therefore ANNs with 15 neurons are used in the six-input case also.

6.2.2 Relative importance

Relative importance (RI) measures the contribution of each type of cue to the mask estimation process (see section 5.4 for information on how RI can be calculated). A knowledge of the RI of each input provides insights into the acoustic conditions in which they are of greatest value and provides an indication of how well each has been integrated into the estimation process.

The level of contribution for each type of input is examined here for both F2 and F6. For the sake of simplicity, rather than measuring RI across the entire frequency band, we measure RI for a subset of 64 frequency points only. The selected frequencies are equally spaced on the equivalent rectan-

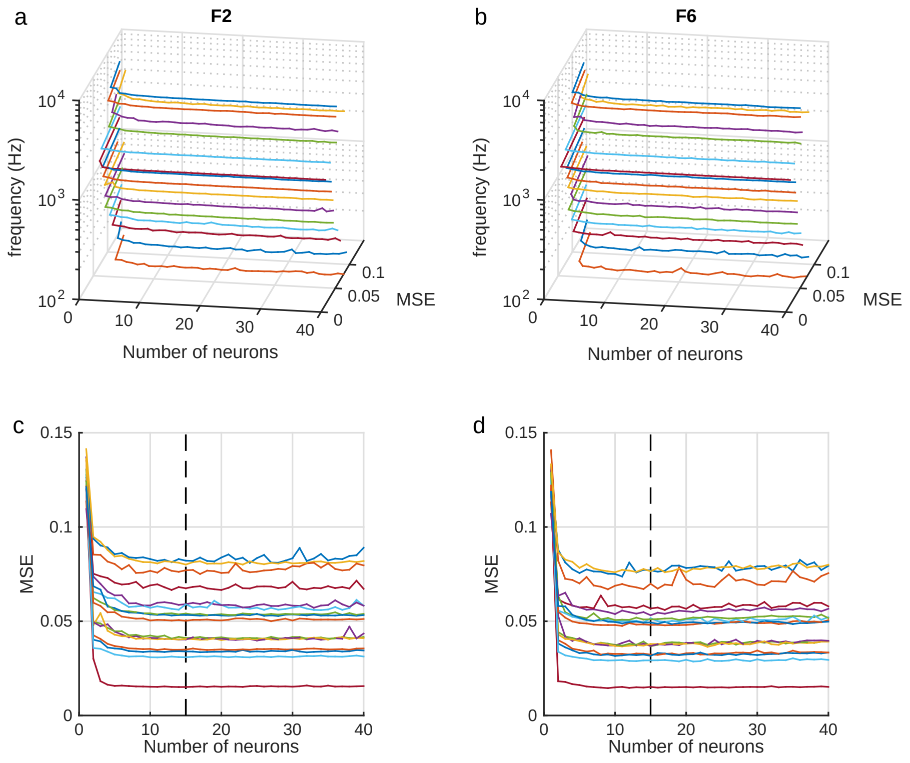


Figure 6.1: MSE performances of ANNs across frequency as a function of the number of neurons in the fixed three-source anechoic configuration with target at 0° azimuth and two interferers at −30° and 30° azimuth, respectively. (a) MSE performance across frequencies as a function of the number of neurons for system F2. (b) MSE performance across frequencies as a function of the number of neurons for system F6. (c) The 2D projection of (a). (d) The 2D projection of (b).

gular bandwidth scale, ranging from 100 Hz up to 8 kHz. These frequencies are mapped onto the linear 320-point STFT frequency scale comprising 161 points from 0 Hz up to 8 kHz, described in section 6.1. Due to the linear frequency spacing of the FFT, at higher frequencies multiple STFT frequency points fall within each ERB. After removing the duplicated frequency points 55 frequency points remain out of the original 64. In order to obtain a sufficiently statistically reliable measurement, 100 ANNs are trained for each frequency point, resulting in 5,500 ANNs in total for F2 and F6. We compare Garson's method and the connection weights approach (see section 5.4) to analyse the relative importance of each type of input based on all 5,500 ANNs. The importance of each of the inputs is ranked and the plausibility of the results from the two methods is scrutinised to determine whether one method is more appropriate than the other in this application.

6.2.2.1 Garson's method

Using Garson's method (see section 5.4), the relative importance of each input is calculated for both systems over all 55 frequency points. Figure 6.2 (a) plots the envelopes of two histograms. These show the number of ANNs in each of 100 bands of relative importance values for the inputs IPD and ILD. The results in figure 6.2 (a) indicate that ILD is more important than IPD since the peak of its distribution occurs at an RI value 10% greater than the peak for the IPD distribution. The plots have even symmetry because the two inputs share a total possible contribution of 100%. Thus, a rise in the RI of IPD causes a commensurate drop in the RI of ILD and vice versa. For system F6 with six inputs, figure 6.2 (b) shows that IPD contributes most strongly and ILD lies in second place, followed by ΔIPD and ΔILD. As anticipated, interaural coherence is determined to be least important in this anechoic situation, with magnitude doing little better, probably because of the infrequency with which both sources are simultaneously at very low levels.

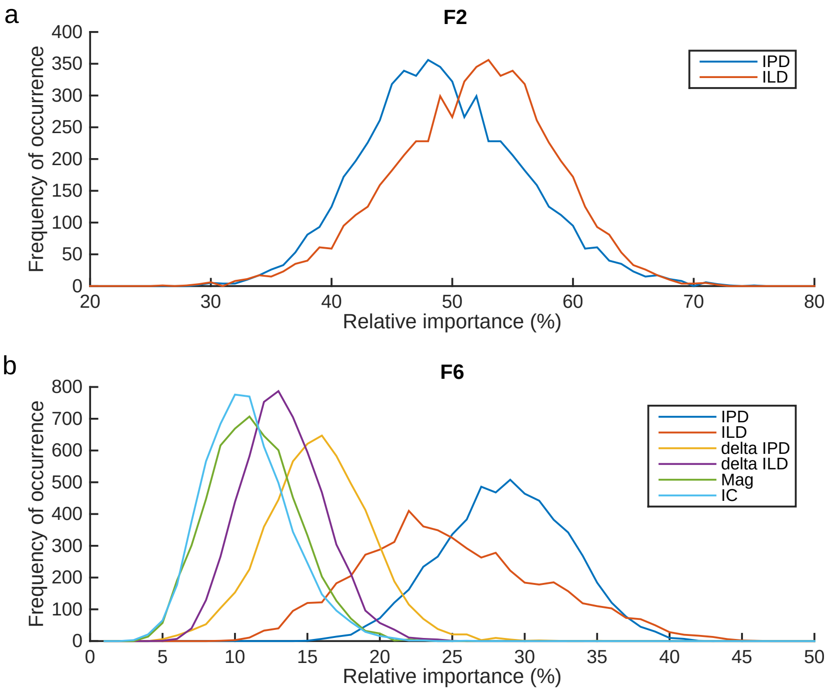


Figure 6.2: Number of ANNs within each relative importance bin for each input computed over all frequencies using Garson's method. The setup consists of three sources: target at 0° azimuth and two interferers at $-30°$ and 30° azimuth, respectively. (a) The results over all frequencies for inputs IPD and ILD in the two-input system F2. (b) The results over all frequencies for inputs IPD, ΔIPD, ILD, ΔILD, magnitude (Mag) and interaural coherence (IC) in the six-input system.

Figure 6.3 breaks Garson's analysis down further and shows the relative importance of IPD and ILD against frequency. Each frequency point is the mean of the relative importance for all 100 ANNs which were trained per frequency point. Figure 6.3 (a) shows that for system F2 IPD is more important than ILD below approximately 650 Hz and more important above about 2.8 kHz. The transition in importance from IPD to ILD is gradual and the results encouragingly reflect the well known fact that IPD (and ITD) are the dominant localisation cues at low frequencies and ILD is the dominant cue at higher frequencies.

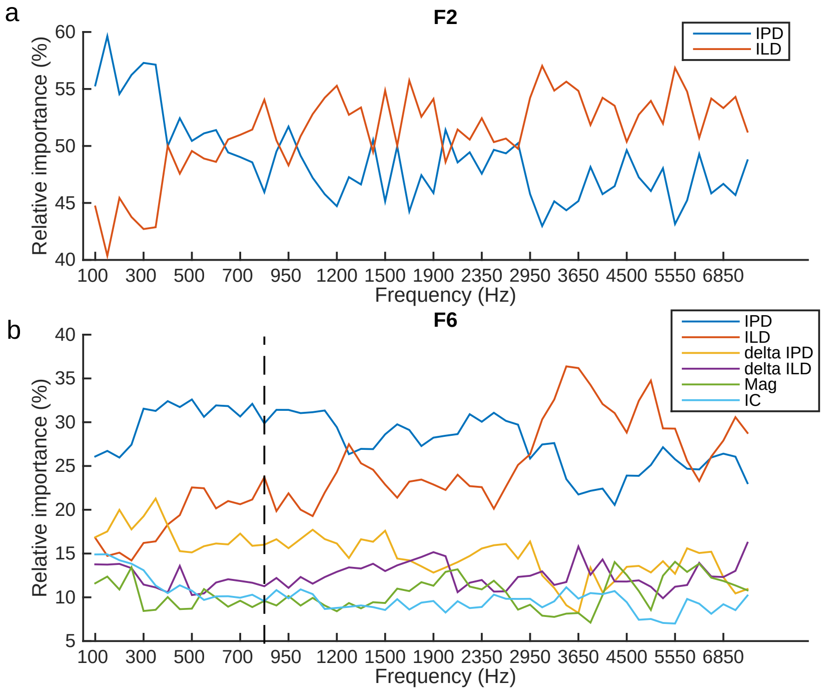


Figure 6.3: Garson's relative importance metric for each type of input for all ANNs from 100 Hz to 8 kHz on an ERB scale in the fixed three-source anechoic setup with target at 0° azimuth and two interferers at −30° and 30° azimuth, respectively. (a) The results for inputs IPD and ILD in the two-input system. (b) The results for inputs IPD, ΔIPD, ILD, ΔILD, magnitude (Mag) and interaural coherence (IC) in the six-input system.

The corresponding results for system F6 are shown in figure 6.3 (b). The relative importance of IPD and ILD cross over between approximately 1.2 kHz and 3 kHz. IPD is more important at frequencies below 1.2 kHz while ILD becomes dominant when the frequency is greater than 3 kHz. Although the two curves touch at approximately 6 kHz, this does not affect the general tendency for ILD to be of greater importance than IPD at higher frequencies. The figure confirms that the extra cues ΔIPD and ΔILD do play a role in mask estimation, even though they contribute less to the estimation overall. In similar fashion to IPD, ΔIPD rises in importance for frequencies below about 1.8 kHz although the opposite trend is not seen in ΔILD. The first crossover between ΔIPD and ΔILD occurs at approximately 1.8 kHz and above this frequency they tend to exhibit similar importance. As expected, the importance of the magnitude and interaural coherence (IC) inputs are the least significant. Above 3.8 kHz, however, the importance of the magnitude input rises to the level where it is approximately equal to that of ΔIPD and ΔILD.

6.2.2.2 Connection weights approach

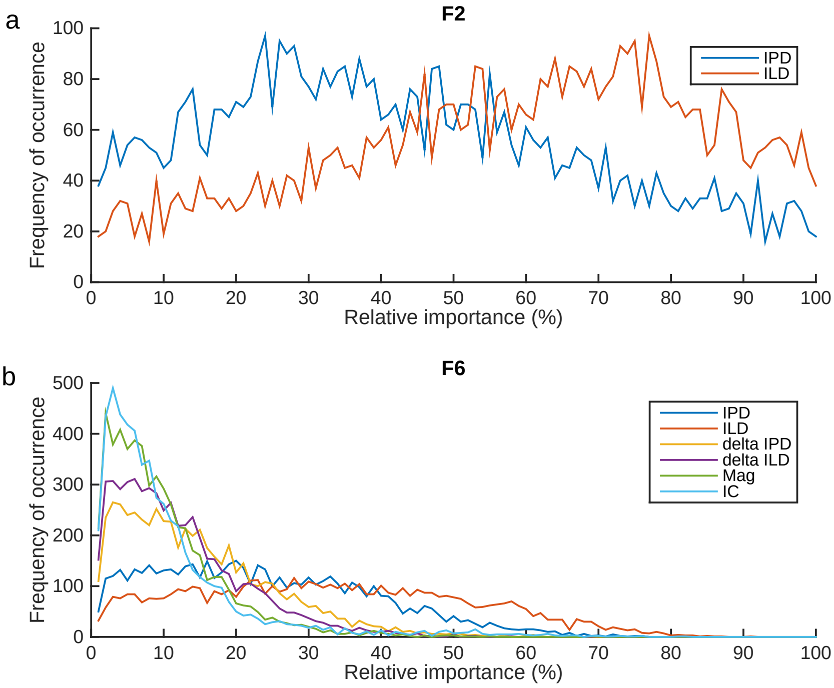


Figure 6.4: Number of ANNs within each relative importance bin for each input computed over all frequencies using the connection weights approach. The setup consists of three sources: target at 0° azimuth and two interferers at $-30°$ and 30° azimuth, respectively. (a) The results over all frequencies for inputs IPD and ILD in the two-input system F2. (b) The results over all frequencies for inputs IPD, ΔIPD, ILD, ΔILD, magnitude (Mag) and interaural coherence (IC) in the six-input system.

We also measure the contribution of each input using a different method, the connection weights approach (see section 5.4). The envelopes of his-

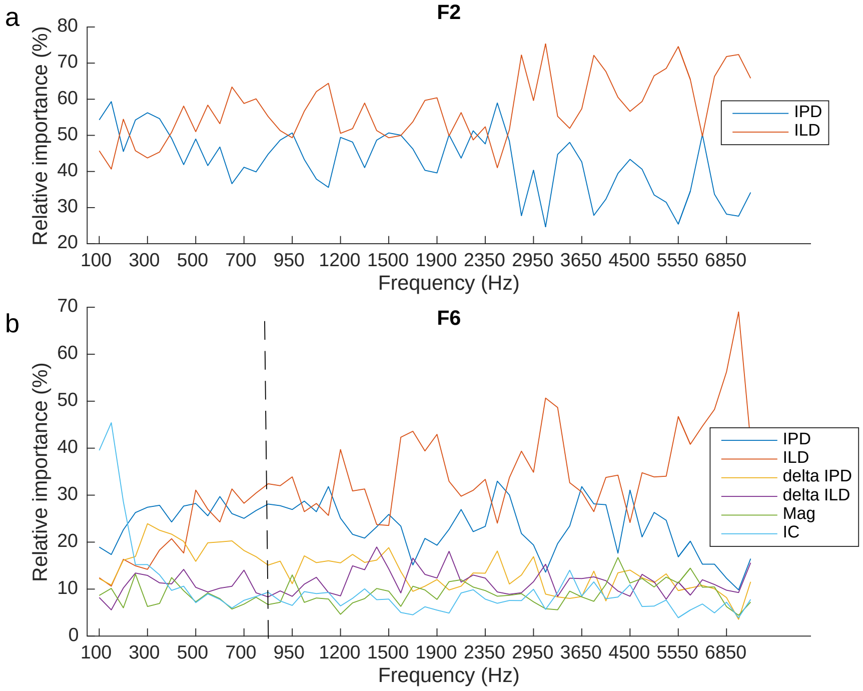


Figure 6.5: Relative importance of inputs for all ANNs from 100 Hz to 8 kHz on an ERB scale using the connection weights approach in the fixed, three-source anechoic setup. The target is at 0° azimuth and the two interferers are at $-30°$ and 30°, azimuth, respectively. (a) The results for inputs IPD and ILD in two-input system. (b) The results for inputs IPD, ΔIPD, ILD, ΔILD, magnitude (Mag) and interaural coherence (IC) in six-input system.

tograms, generated in a similar way to those in figure 6.2, are shown for both systems in figure 6.4. As when using Garson's method, similar results are obtained for system F2 to those in figure 6.4 (a) in that the relative importance values for the IPD inputs are grouped towards the lower end of the RI scale, which indicates that ILD cues contribute less to mask estimation than IPD cues. Figure 6.4 (b) shows the contribution of each type of cue to mask estimation in system F6. Both ILD and IPD lead in importance, but their positions are reversed, with ILD showing as the most dominant cue, followed by IPD. Both have broader distributions than in figure 6.2 (b). As with Garson's method, the places of third and fourth in importance are taken by ΔILD and ΔIPD, respectively. Also, magnitude and interaural coherence are again the least important cues.

Figures 6.5 (a) and (b) compare the relative importance, calculated using the connection weights approach, between each input as a function of frequency for systems F2 and F6, respectively. For F2, the relative importance trend of IPD suggests a slight dominance over ILD only below a few hundred hertz. The ILD cue becomes clearly dominant above 2.8 kHz. There is a wide transition region over the frequency range between 400 Hz and 2.8 kHz. These results are broadly similar to the results obtained using Garson's method. For system F6, the first difference compared with that method is that the relative importance of IPD and ILD crosses at a lower frequency of around 500 Hz. Secondly, interaural coherence is the most dominant cue at very low frequencies below 200 Hz.

6.2.2.3 Discussion

It can be seen that there are both similarities and differences between the results obtained using the two relative importance metrics. In this section, we investigate and discuss these differences.

In figure 6.2, the relative importance for systems F2 and F6 calculated by Garson's method indicates that the most important cue for F2 is ILD, whereas the most important cue for F6 is IPD. This can be accounted for by considering the relative importance of each cue as a function of frequency, shown in figure 6.3. In figure 6.3 (a), the bandwidth dominated by IPD is approximately 650 Hz (11 frequency points). The bandwidth from 2.8 kHz to 8 kHz (44 frequency points) is dominated by ILD. Therefore, many more frequency points in the 55 that were analysed are dominated by ILD and this causes ILD to appear more important than IPD in the RI histogram (figure 6.2 (a)). However, the number of analysed frequency points which are dominated by IPD (36) is more than the number dominated by ILD (19), as shown in figure 6.3 (b). Therefore, in figure 6.2 (b) IPD is ranked as the most important cue. This explanation does not provide a reason for the differences observed between the two systems in the bandwidths dominated by each cue, which remains an open question.

We next address the different results obtained by the two methods. As we described in section 5.4, Garson's method employs the absolute values of connection weights to quantify the importance of each ANN input. Therefore, it ignores the opposing influences of any connections with negative connection weights (Olden et al., 2004). In order to understand the disadvantage of Garson's method that leads to mis-ranking between IPD and ILD at some frequency points, we demonstrate the phenomenon at 800 Hz, where the importance of ILD is 10% smaller than IPD in figure 6.3 (b) (using Garson's method), and 4% larger than IPD in figure 6.5 (b) (using the connection weights method). A dashed vertical line has been drawn at 800 Hz in both of these figures.

To demonstrate the origin of the different results obtained by the two methods, ANN connection weight products for IPD and ILD, averaged over all 100 ANNs at 800 Hz, are listed in table 6.2. There are 15 neurons in the system, as shown in figure 6.1. In table 6.2, considering the IPD input column, it is apparent that eight hidden neurons positively influence mask estimation and the sum of their weights is 1.93. The remaining seven hidden neurons are all negative and have an opposing influence on mask estimation. They sum to -2.24. In a similar way, the ILD connection weights contribute a total positive influence of 1.05 to the estimation, and a total negative influence of 3. Both the negative and positive influences are considered by the connection weights method. This is represented in table 6.2 by the rows labelled 'sum of positive weights' and 'sum of negative weights' and their combined influence in the row 'sum of all weights'.

Garson's method, however, does not account for the effects of inhibition. The fact that some hidden neurons have negative weights is lost by using absolute values in the calculation. This leads to different estimations of ANN input importance by the two methods (Olden et al., 2004). The effect

Table 6.2: Averaged connection weight products for inputs IPD and ILD for the 100 F6 ANNs trained at 800 Hz. The sum of all weights corresponds to the connection weights result, and the bottom row corresponds to the Garson result.

Neuron ID	Input	
	IPD	ILD
1	0.24	-0.57
2	0.72	0.77
3	0.44	0.11
4	0.14	-0.22
5	-0.30	-0.38
6	0.18	0.08
7	0.05	-0.34
8	-0.24	-0.09
9	-0.69	-0.60
10	-0.09	-0.24
11	0.01	0.02
12	-0.27	-0.29
13	-0.39	0.06
14	-0.28	-0.22
15	0.14	-0.03
Sum of positive weights	1.93	1.05
Sum of negative weights	-2.24	-3.00
Sum of all weights	-0.31	-1.95
Sum of absolute weights	4.17	4.04

of Garson's method is simply illustrated by the bottom row 'sum of absolute weights' in table 6.2.

Scrutiny of the values in the bottom row of table 6.2 reveals that the IPD absolute sum (4.17) is larger than the absolute sum for the ILD weights (4.04). Therefore, Garson's method ranks IPD as more important than ILD. On the other hand, in the 'sum of all weights' row, positive and negative connection weights tend to cancel. This results in a sum of -0.31 for IPD, which is considerably less negative than the sum of -1.95 for ILD. Therefore, the connection weights approach uses raw connection weights which preserves the reinforcement/inhibition information and is why the method is preferred by Olden et al. (2004). However, the approach tends to lose information about the strength of the mapping between an input and the output of the ANN, since weights with high positive values and others with high negative values tend to cancel and this appears to be a distinct advantage of Garson's method.

In figure 6.5 (b) it can be observed that, according to the connection weights approach, interaural coherence (IC) becomes the most important cue below about 200 Hz and peaks at 150 Hz. To investigate this more closely we plot in figure 6.6 the relationship between IC and the ratio mask value R at 150 Hz.

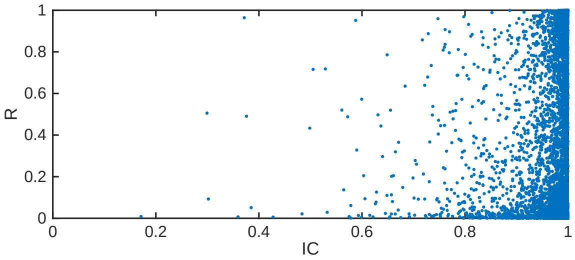


Figure 6.6: Relationship between interaural coherence (IC) and ratio (R) at 150 Hz in ANN training data.

IC measures the similarity between the signals reaching the two ears. At low frequencies, there is very little acoustic difference between the left and right channels. Therefore, in figure 6.6, most values of IC are close to 1, independent of the value of the ratio, R. From this we conclude that IC must be unimportant at this frequency, because it is impossible to predict the value of R from IC. From this point of view, Garson's method produces a more believable result, because figure 6.3 (b) shows IC to be the least important input at low frequencies.

For both Garson's method and the connection weights method, we also notice that some of the importance curves are smooth and some appear to be noisy. An important point to discuss is where the noise is coming from. One possible reason is that it has something to do with variations in the properties of the training and test data. Another reason may be that some ANNs complete the training phase closer to the MSE global minimum than others. It is likely, for example, that it is the noise in the RI metric which causes the two importance curves, IPD and ILD, to touch at approximately 6 kHz in figure 6.3 (b) (Garsons method), and is why the RI scores for IPD and ILD cross over at 200 Hz as shown in figure 6.5 (a) (connection weights method).

It is clear from this discussion that both RI metrics have their own strengths and weakness and so in subsequent RI analyses we continue to show the results for both approaches. Furthermore, to reduce the impact of noise, we apply smoothing to the RI curves presented in later sections of the thesis to reveal the underlying trends in RI more clearly.

6.2.3 Target speech segregation STOI performance

In previous sections a suitable ANN topography has been established for the purpose of estimating the ratio mask in our optimal cue mapping approach to target speech segregation. We have also presented a method for ranking potentially useful cues in order of importance as a function of frequency.

In this section we begin to evaluate the perceptual performance of optimal cue mapping. We do so by applying standard models for estimating the intelligibility gains delivered by OCM.

Target speech that the ANNs have not been exposed to during the training phase is evaluated in terms its STOI scores. The test data employed here consists of 50 test mixtures, described in section 6.1, with local SNRs of -5 dB, 0 dB and 5 dB. By comparing the segregated target speech with the original clean target speech before it was mixed with the interferers, a STOI score is obtained. The STOI results for four masks types are presented in table 6.3: EBMF2, ERMF2, EBMF6 and ERMF6 (see section 6.1.3). The baseline results using the ideal binary mask (IBM) and ideal ratio mask (IRM), the definitions for which can be found in section 5.1.2, are presented to indicate the upper bound for possible intelligibility improvements.

Before discussing the STOI scores in detail, we use this opportunity to compare two different post-processing methods. Synthesis of the segregated target speech is implemented using both of the methods described in section 5.5 in order to examine how much each method influences intelligibility performance. Both are based on the short-time Fourier transform (STFT). In the convolution (CONV) method, however, the convolution tail created in each frame is preserved and added to the following frame, whereas in the IFFT method the tail is discarded and each frame is Hann windowed to avoid possible discontinuities at frame boundaries.

As shown in table 6.3, the convolution method yields better STOI scores than the IFFT method in all scenarios. Furthermore, in terms of the STOI score improvements, the binary masks benefit more than the ratio masks from using the convolution method. This result is likely due to the fact that the binary mask forces hard decisions at each T-F boundary which inherently causes the resynthesised signal to be less smooth, and hence less intelligible, than when the ratio mask is used. Keeping the convolution tail removes small but audible amplitude modulations of the audio output between neighbouring time frames. There is more to gain by using the convolution method in both of these situations. Since the convolution method is shown to be superior, it is used in the post processing for all subsequent evaluations.

Table 6.3: STOI scores using four different masks with two different post-processing methods at three different SNRs. The configuration is three-source anechoic, with the target at 0° azimuth and two interferers at $-30°$ and 30° azimuth, respectively. Numbers in bold are the best segregation results for each test condition and numbers in italics are the best ideal results.

Systems	Methods	STOI					
		-5 dB		0 dB		5 dB	
UPM		0.5448		0.6819		0.8266	
		Binary mask results					
EBMF2	IFFT	0.7367	19.19%	0.8279	14.60%	0.9132	8.66%
	CONV	0.7444	19.96%	0.8313	14.94%	0.9150	8.84%
EBMF6	IFFT	0.7482	20.34%	0.8384	15.65%	0.9187	9.21%
	CONV	**0.7552**	**21.04%**	**0.8421**	**16.02%**	**0.9202**	**9.36%**
IBM	IFFT	0.8284	28.36%	0.8955	21.36%	0.9469	12.03%
	CONV	*0.8377*	*29.29%*	*0.9015*	*21.96%*	*0.9500*	*12.34%*
		Ratio mask results					
ERMF2	IFFT	0.7812	23.64%	0.8572	17.53%	0.9308	10.42%
	CONV	0.7817	23.69%	0.8579	17.60%	0.9312	10.46%
ERMF6	IFFT	0.7869	24.21%	0.8622	18.03%	0.9338	10.72%
	CONV	**0.7880**	**24.32%**	**0.8634**	**18.15%**	**0.9344**	**10.78%**
IRM	IFFT	0.8843	33.95%	0.9249	24.30%	0.9604	13.38%
	CONV	*0.8844*	*33.96%*	*0.9258*	*24.39%*	*0.9619*	*13.53%*

From a practical point of view, if an ANN can generalise from a restricted set of training data then this will reduce the training data size and training time required. For this reason it is important to know how a system trained in one signal-to-interferer SNR condition performs in different SNR conditions. To more easily assimilate the data in table 6.3, from the point of view of the generalisation ability of each ANN, it is presented graphically in figure 6.7. The figure shows the STOI scores for each mask at three different SNRs compared with the scores for the unprocessed mixture (UPM).

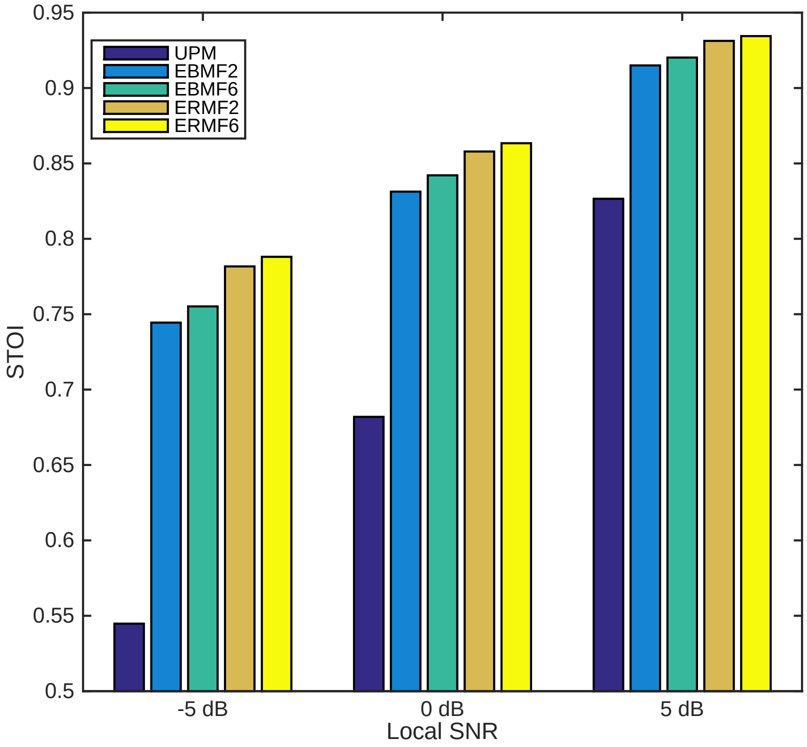


Figure 6.7: STOI scores for the masks EBMF2, EBMF6, ERMF2 and ERMF6 compared with the unprocessed mixture UPM and the corresponding ideal masks, IBM and IRM. The comparison is performed for local SNRs of -5 dB, 0 dB and 5 dB in anechoic conditions with target speech at 0° azimuth and two interferers at $-30°$ and 30° azimuth, respectively.

Figure 6.7 shows that, for binary mask, EBMF2, compared with the unprocessed mask, UPM, the intelligibility improves by approximately 20%, 14.9% and 8.8% at local SNRs of -5 dB, 0 dB and 5 dB, respectively. In addition, there are 23.7%, 17.6% 10.5% improvements using the ratio mask ERMF2 for the three SNR test conditions.

By incorporating more input features in system F6, masks EBMF6 and ERMF6 perform better than their two-input counterparts using F2. For the binary mask based on six inputs, there is approximately a further 1.1%, 1.1% and 0.5% STOI score improvement over the equivalent two-input system for the -5 dB, 0 dB and 5 dB SNR conditions, respectively. The improvement for the corresponding ratio masks is smaller; approximately 0.6%, 0.6% and 0.3%, respectively.

As anticipated, the overall ratio mask performance is superior to that of the binary mask, particularly at low SNRs. Ratio mask ERBF2 achieves an improvement in STOI score of almost 3.7% at -5 dB SNR and 1.6% at 5 dB over the binary mask EBMF2. For the six-input system, compared to EBMF6, ERMF6 gains a further 3.3% improvement at -5 dB and 1.4% at 5 dB.

Overall, the mask which performs best in all test conditions is ERMF6. It yields an 18.2% STOI score improvement compared to the STOI score for the unprocessed mask UPM at 0 dB SNR. For the lower SNR of -5 dB, the mask ERMF6 creates an even greater improvement of 24.3% in the STOI score. At the highest SNR (5 dB) the improvement drops to 10.8%. In general, the STOI score is most improved for lower SNRs, and asymptotically approaches the score for the ideal upper bound for higher SNRs. The greater improvement at poor SNRs, particularly for the ratio masks, is especially advantageous in a hearing aid application.

In general, the STOI score is most improved at the lower SNR, and asymptotically approaches the score for the ideal upper bound at higher SNR conditions. The greater improvement at poor SNRs, particularly for the ratio mask estimators, is especially advantageous in a hearing aid application.

Table 6.4 presents an alternative measure for the intelligibility performance improvements produced by the masks. It compares the actual improvement in STOI score compared with the unprocessed mixture with the ideal maximum STOI score improvement. The performance gain describes the position of real results in the improvement space using the difference between the STOI score of the mixture and the ideal mask as a reference. We define the STOI performance gain using the following formula:

$$gain_{STOI} = \frac{STOI_{real} - STOI_{UPM}}{STOI_{ideal} - STOI_{UPM}} \tag{6.1}$$

Hence, a performance gain close to 1 indicates that the STOI score using the mask is close to the ideal upper bound. A gain of zero means that there is no improvement.

Table 6.4: The STOI performance gain of the masks for a three-source anechoic configuration with the target at 0° azimuth and two interferers at $-30°$ and $30°$ azimuth, respectively. Numbers in bold are the best results for the binary and ratio masks in each test condition.

Systems	Performance gain		
	-5 dB	0 dB	5 dB
EBMF2	0.68	0.68	0.72
EBMF6	**0.72**	**0.73**	**0.76**
ERMF2	0.70	0.72	0.77
ERMF6	**0.72**	**0.74**	**0.80**

Table 6.4 shows that, for all SNR conditions, the six-input masks, EBMF6 and ERMF6, perform better than their two-input counterparts, EBMF2 and EBMF6. In all cases, although the STOI improvement percentage increases as the SNR decreases, the performance gain in the improvement space reduces. This is reasonable; it is clearly more difficult to segregate the target signal in low SNR conditions and the performance gain reflects this fact. The improvement space for lower SNR conditions is wider than for high SNR conditions, hence a smaller performance gain at low SNRs may nevertheless result in a relatively large percentage STOI improvement.

6.2.4 Target speech segregation PESQ performance

As described at the beginning of this chapter, there is no simple connection between speech intelligibility and its quality. However, processed speech of good quality can bring hearing aid users a higher level of listening comfort. Therefore, it is worth examining the speech quality yielded by the masks. The PESQ score for each mask is shown in table 6.5. Similar trends to those seen in the STOI results are apparent. Speech quality improves more with the inclusion of a richer set of input features and with the use of a ratio mask as opposed to a binary mask.

We define the PESQ performance gain using a formula similar to the one for STOI performance gain:

$$gain_{PESQ} = \frac{PESQ_{real} - PESQ_{UMP}}{PESQ_{ideal} - PESQ_{UPM}} \tag{6.2}$$

It is the proportion defined by the difference between the PESQ score for the processed signal and the PESQ score for the unprocessed mixture, compared with the maximum possible difference.

Table 6.6 shows that, for each SNR condition, EBMF6 and ERMF6 exhibit the highest PESQ performance gains out of the two binary masks and two ratio masks, respectively. Although mask EBMF6 has a higher PESQ performance gain than ERMF6, it yields a lower percentage improvement in Table 6.5 than ERMF6 does. It is noteworthy that the speech quality improves more for lower SNR conditions, but that the performance gain drops due to the segregation challenge becoming greater. Overall, comparison of the PESQ scores in table 6.5 between the binary mask estimator and the ratio mask estimator shows that the output quality of the ratio mask estimator is superior in matching conditions.

Table 6.5: PESQ scores using four different masks with two different post-processing methods at three different SNRs. The configuration is three-source anechoic, with the target at 0° azimuth and two interferers at $-30°$ and 30° azimuth, respectively. Numbers in bold are the best segregation results for each test condition and numbers in italics are the best ideal results.

Systems	PESQ					
	-5 dB		0 dB		5 dB	
UPM	1.4608		1.8324		2.1730	
	Binary mask results					
EBMF2	2.2746	16.28%	2.6014	15.38%	2.9655	15.85%
EBMF6	2.3052	16.89%	2.6700	16.75%	3.0115	16.77%
IBM	*2.7957*	*26.70%*	*3.0994*	*25.34%*	*3.3520*	*23.58%*
	Ratio mask results					
ERMF2	2.3692	18.17%	2.7044	17.44%	3.0876	18.29%
ERMF6	**2.4137**	**19.06%**	**2.7500**	**18.35%**	**3.1157**	**18.85%**
IRM	*3.1360*	*33.50%*	*3.3822*	*31.00%*	*3.6082*	*28.70%*

Table 6.6: The PESQ performance gain of the masks for a three-source anechoic configuration with the target at 0° azimuth and two interferers at $-30°$ and 30° azimuth, respectively. Numbers in bold are the best results for the binary and ratio masks in each test condition.

Systems	Performance gain		
	-5 dB	0 dB	5 dB
EBMF2	0.61	0.61	0.67
EBMF6	**0.63**	**0.66**	**0.71**
ERMF2	0.54	0.56	0.64
ERMF6	**0.57**	**0.59**	**0.66**

6.3 Incorporation of cross-correlation in different SNR conditions

The performance of the OCM algorithm in terms of target speech intelligibility and quality for three different SNRs has been evaluated in the preceding sections. However, we have so far restricted the number of interferers to two and they have fixed directions of ±30° azimuth. The purpose of this section is to investigate relaxing the constraints on the interferers, initially to allow their directions to vary. To do so, we integrate cross-correlation coefficients (see Section 5.2.3) into the inputs to the mask estimator as a means of providing the ANNs with more information about the direction (and ultimately the number) of sources in the binaural mixture.

A sound source from a particular azimuth direction has a characteristic ITD associated with it. This converts to a time lag when the left and right binaural channels are cross-correlated and this forms the basis for the cross-correlogram (section 5.2.3). Human ITDs range from approximately -680 μs to +680 μs (see section 3.1.1). Given the sampling frequency of our current model (16 kHz), this conveniently results in 27 discrete time lags (13 negative, zero and 13 positive) over the range -800 μs to +800 μs. As a starting point, therefore, these 27 cross-correlation coefficients will be applied as additional inputs to the ANNs in the mask estimators. Aims of this section include determining whether these cross-correlation coefficients will assist the ANN in estimating masks and whether 27 coefficients is a suitable number.

Based on system F6's superior intelligibility according to the STOI analysis in section 6.2.3 and higher quality using the PESQ analysis in section 6.2.4, we next evaluate whether cross-correlation coefficients enhance the ratio estimation of this system. The new system with seven types of inputs is denoted by F7. F7, like F2 and F6 (defined in section 6.1.3), has a binary mask variant, EBMF7, and a ratio mask variant, ERMF7. Hence, it results in a new system with 33 dimensions of input features.

The same training and testing strategy used in section 6.2 is applied here. A baseline analysis is performed first, where the fixed, two-interferer setup is applied to the new S7 ANN. Only if the results are as expected, i.e. they are similar to the results for system F6, is it safe to move on to more challenging setups.

6.3.1 Establishing the ANN topologies

To ensure that the new networks contain sufficient neurons to reach a near-minimum level of MSE, the same process (see section 6.2.1) is employed to define the 33-input ANN topology.

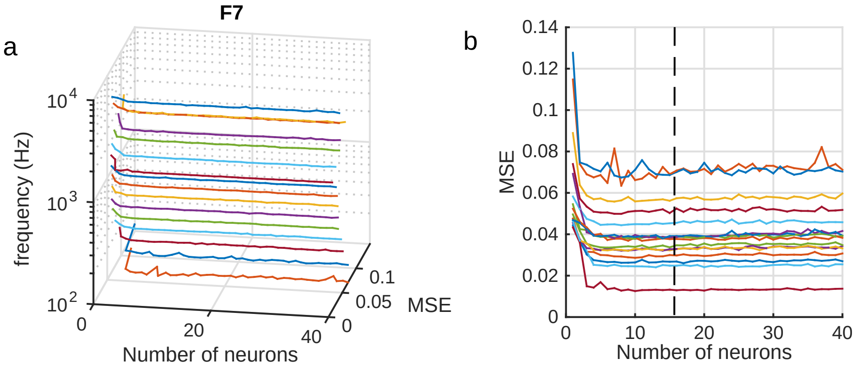


Figure 6.8: MSE performances of ANNs across frequency as a function of the number of neurons in the fixed three-source anechoic configuration with target at 0° azimuth and two interferers at $-30°$ and 30° azimuth, respectively. (a) MSE performance across frequencies as a function of the number of neurons for system F7. (b) The 2D projection of (a).

The size of the F7 ANNs is determined, as before, using 10-fold cross-validation. The mean of the MSEs for the 16 frequency bands analysed are plotted in figure 6.8. According to the performance of the ANNs using different numbers of neurons, we again select ANNs with 15 hidden neurons for this system. The internal network topology is unchanged, because the training data setup has not been altered from the fixed-source conditions used with systems F2 and F6 in section 6.2.

6.3.2 Relative importance

The purpose of introducing cross-correlation is to enable the ANN to estimate the ratio mask better for interferers of arbitrary direction. Therefore, evaluating its relative importance for the fixed 2-interferer setup is expected to cause the cross-correlation input to have a very low importance. We use this configuration as a baseline for assessing the relative importance of the cross-correlation input using increasingly complex source arrangements.

In previous relative importance measurements, described in section 6.2.2, all input cues, such as IPD and ILD, involved presenting a single input to the ANN. Although cross-correlation coefficients are considered to form one type of input, they involve multiple inputs which are fed in parallel into the ANN. This alters the way in which the relative importance of the cross-correlation coefficients are computed. In order to evaluate a vector input, the mean value of all the corresponding connection weight products is used. The relative importance measured by using Garson's method is shown in figure 6.9 and the corresponding results using the connection weights method is shown in figure 6.10.

The envelopes of seven histograms, one for each input type, are presented in figure 6.9 (a) and 6.10 (a), showing the frequency of occurrence of the RI scores for each input in each of 10 bins. Both sets of results indicate that IPD and ILD still contribute most strongly to mask estimation. ΔIPD, ΔILD, magnitude and interaural coherence appear to contribute to mask estimation as well, but to a much lesser extent. This also agrees with our earlier results in section 6.2.2. As expected, the cross-correlation coefficients cue is the weakest in this setup for F7. This is because the system has been trained and tested with fixed source directions. In these circumstances, very little extra information is provided by the cross-correlation coefficients, which have therefore not been exploited by the networks.

More detailed relative importance results, averaged across all 100 ANNs

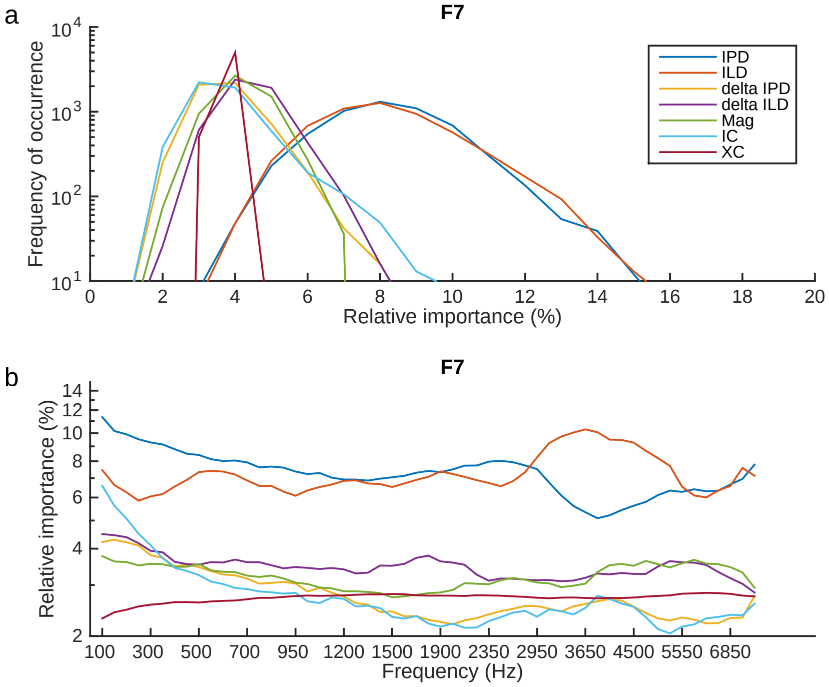


Figure 6.9: Relative importance calculated using Garson's method for the inputs IPD, ΔIPD, ILD, ΔILD, magnitude (Mag), interaural coherence (IC) and cross-correlation coefficients (XC) for all ANNs in the fixed, three-source anechoic setup for system F7. The speech target is at 0° azimuth and the two interferers are at −30° and 30° azimuth, respectively. (a) Relative importance histogram envelopes for system F7. (b) Smoothed relative importance of inputs for all ANNs from 100 Hz to 8 kHz on an ERB scale.

trained at each frequency, is shown in figures 6.9 (b) and 6.10 (b) for each of the RI metrics, respectively. As discussed in section 6.2.2.3, smoothing the results of relative importance across frequency reveals the underlying trends more clearly that we are trying to identify. Hence, a 3-point moving average filter is applied to smooth the curves. Note too that the vertical scales are now logarithmic to help distinguish the curves with lower RI scores.

Garson's method suggests that IPD is the most important cue below 1.2 kHz and ILD is mostly dominant above 3 kHz. The connection weights

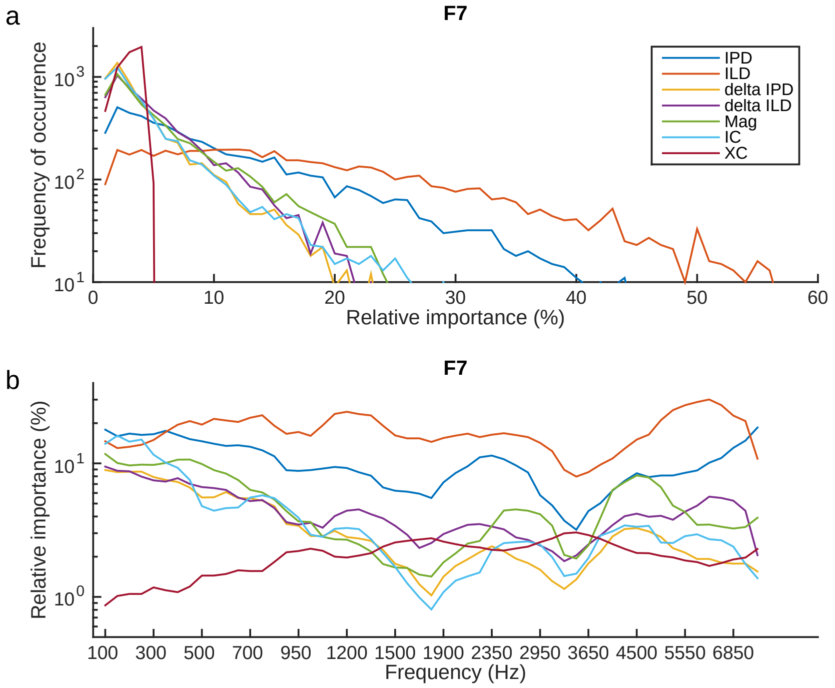


Figure 6.10: Relative importance calculated using the connection weights approach for the inputs IPD, ΔIPD, ILD, ΔILD, magnitude (Mag), interaural coherence (IC) and cross-correlation coefficients (XC) for all ANNs in the fixed, three-source anechoic setup for system F7. The speech target is at 0° azimuth and the two interferers are at −30° and 30° azimuth, respectively. (a) Relative importance histogram envelopes for system F7. (b) Smoothed relative importance of inputs for all ANNs from 100 Hz to 8 kHz on an ERB scale.

method, however, indicates that the RI scores for IPD and ILD cross over at the lower frequency of 400 Hz. Although, the transition frequency is different, it can still be seen that IPD is the dominant cue at low frequencies and ILD dominates at higher frequencies. In addition, the importance of ΔIPD decreases compared with F6 in figures 6.3 (b) and figure 6.5 (b). Both methods also suggest that the contribution of interaural coherence in the estimation is very low for frequencies above 1.2 kHz.

6.3.3 Target speech segregation STOI and PESQ performance

The STOI and PESQ measurement procedures employed for systems F2 and F6 in sections 6.2 are repeated for F7 to discover if the addition of cross-correlation coefficients improves the performance of the resulting binary and ratio masks. The F7 results, using the previously defined fixed-source, two-interferer setup are shown in table 6.7. For comparison purposes, the results of F6 are repeated alongside them. As shown in the table, ERMF7 yields slightly higher scores for both the STOI and the PESQ analyses for all SNR conditions. At an SNR of -5 dB there is another 1% improvement in the STOI rating compared to ERMF6. For an SNR of 5 dB the improvement falls to 0.3%. There is also a small improvement in PESQ scores of approximately 0.02 - 0.03. It is interesting to note that although the STOI score improves using the binary mask EBMF7, the PESQ score consistently decreases slightly. The reason for this is unclear.

Table 6.7: STOI and PESQ scores for systems F6 and F7 for different SNRs using a three-source anechoic configuration. The target is at 0° azimuth and the two interferers are at $-30°$ and 30° azimuth, respectively. Numbers in bold are the best results for each test condition.

Systems	-5 dB		0 dB		5 dB	
	STOI	PESQ	STOI	PESQ	STOI	PESQ
EBMF6	0.7552	2.3052	0.8421	2.6570	0.9202	3.0115
EBMF7	0.7637	2.2994	0.8467	2.6354	0.9230	2.9962
ERMF6	0.7880	2.4137	0.8634	2.7500	0.9344	3.1157
ERMF7	**0.7985**	**2.4486**	**0.8697**	**2.7720**	**0.9375**	**3.1362**

Although we have shown that the relative importance of the cross-correlation inputs is, as expected, small, it is not zero. This may account for the small but consistent further increase in performance of F7 over F6, even for this two-source, fixed direction setup.

6.4 Conclusion

In this chapter we considered the training of three families of ANNs, with topographies F2, F6 and F7. They were trained using different numbers of input features. The training data was a binaural mixture consisting of a mixture of target and interferer speech fragments. A left and a right channel ANN was assigned to each frequency point. Each ANN was trained to estimate the ideal ratio mask value for its associated frequency and training was terminated once the output MSE using unseen test data fell below a previously determined threshold.

The importance of every input for each system was ranked. Both similarities and differences are observable between the results obtained using the two relative importance metrics; Garson's method and the connection weights method. Garson's method does not account for the effects of inhibition of the hidden neurons. The metric uses the absolute values of connection weights. By contrast, the raw connection weights are used in the connection weights method, which therefore preserves the reinforcement/inhibition information. High positive and negative weights tend to cancel, however, and when this occurs, the strength information in the mapping is lost between the inputs and the output of the ANN. Due to the complementary strengths and weakness of the two methods, we use both approaches to demonstrate the importance of each cue.

Both of these relative importance metrics produce noisy results. This noise may stem from variations in the properties of the training/test data and the different local minima attained by each ANN at the completion of its training phase. Smoothing is applied to the relative importance results to reduce the impact of the noise.

Integrating extra input features always led to better performance in terms of speech intelligibility and generally improved the quality metric. The STOI and PESQ scores confirm the results of the importance analysis in section

6.2.2, which predicted that the additional inputs ΔIPD and ΔILD would provide a small, but clear, improvement in the estimation of the binary and ratio masks. The importance of each of the inputs for each system was ranked and the plausibility of the results for the six-input system was scrutinised. Based on the importance analysis for system F6 in section 6.2.2, IPD is the dominant cue at low-frequencies and ILD is the most important cue at higher frequencies. The interaural coherence and magnitude inputs appear to be redundant. Only the additional inputs ΔIPD and ΔILD are likely to provide significant improvements when integrated with IPD and ILD.

System F7 was introduced to release the constraints on the direction of interferers by integrating cross-correlation coefficients as additional inputs. As expected, testing F7 at fixed directions of interferers produced similar results to those for F6, though a small improvement in terms of both STOI and PESQ scores was nevertheless observed. This confirmed that there is no harm in integrating these coefficients and they are expected to become of more use in subsequent simulations when we introduce interferers which vary in direction in next chapter.

In addition, we have demonstrated that the systems trained for the 0 dB SNR condition also have the ability to generalise to SNR conditions 5 dB above and 5 dB below this level.

In this chapter, all the systems were evaluated using anechoic mixtures in which the interference directions were fixed. In the next chapter we release these constraints and the proposed algorithm is evaluated with variable interference directions in both anechoic and reverberation mixtures. A systematic evaluation of the ability of the ANNs to generalise when tested on binaural mixtures with a different SNR from the one they were trained on is also presented next.

Chapter 7

Detailed Evaluation of the OCM Approach

In the previous simulations, the systems F2, F6 and F7 were trained using a three-source configuration in anechoic conditions, with the target placed straight in front of the virtual listener (i.e. in the direction 0° azimuth) and two fixed interferers were placed one on either side of the listener at $-30°$ and $30°$ azimuth, respectively.

Now, with the aim of applying optimal cue mapping in increasingly realistic situations, we make the problem space more complicated. In this chapter the direction of the interferers is varied in 10 degree increments. We now allow the direction of one interferer to range between $-90°$ and $-10°$ azimuth and the other interferer to range between $10°$ and $90°$ azimuth. Thus, there will be always one interferer on both sides of the listener. With 10° steps, there are 91 possible combinations of direction. A vector of coordinates is used to represent the permitted interferer configurations. For example, $[-30, 30]$ indicates one interferer is at $-30°$ azimuth and the other one is at $30°$ azimuth. As in previous simulations, the target is always located at 0° azimuth and hence the position of target is omitted from the configuration vector. The configurations used in previous simulations form one of these 91 combinations so that performance comparisons can easily be made between them and the results presented in this chapter.

7.1 Variable direction interferers in anechoic environments

Since the problem space is now much bigger, the training strategy will be different. Most importantly, the training phase will be considerably longer and the ANNs will be considerably larger than for the previous systems for which the two interferers were fixed. In the next section we consider the impact of varying the training strategy both on the ANN topology required and on the performance as measured using STOI and PESQ.

7.1.1 Planning the training and testing strategies

We consider three training strategies, TS1, TS2 and TS3, of varying complexity. Each of these strategies is likely to require a different network topology for effective learning and we define these as systems V7-1, V7-2 and V7-3, respectively. All three systems produce ratio masks with the generic label ERMV7 and a binary mask counterpart EBMV7. These systems include the cross-correlation coefficients defined in section 5.2.3. In the current systems, however, the key point is that the seven inputs are used to estimate the ratio mask for segmenting a target speech source from two interfering speech sources whose positions are no longer fixed. For each direction combination, we generate a set of training data containing 5,000 examples. This number was previously established in section 6.2 to be sufficient for successful training.

Training strategy TSV7-1

The most obvious training strategy is to choose to train a system using training data which covers all possible interferer directions and combinations. This results in 91 sets of 5,000 training examples (455,000 examples in total).

Training strategy TSV7-2

To reduce the size of the training set, a second system is trained with the target at $0°$ azimuth, but in this case the two interferers are paired in symmetrical directions and stepped in $10°$ increments: $[-10, 10]$, $[-20, 20]$, $[-30, 30]$, $[-40, 40]$, $[-50, 50]$, $[-60, 60]$, $[-70, 70]$, $[-80, 80]$, $[-90, 90]$. This reduces the training data size to nine sets of 5,000 examples (45,000 examples in total).

Training strategy TSV7-3

In the final strategy, the system is trained as for TSV7-2, but using only four pairs of directions: $[-20, 20]$, $[-40, 40]$, $[-60, 60]$, $[-80, 80]$. Thus, there are 20,000 training examples in total for strategy TSV7-3.

Definitions

The three training strategies yield three distinct families of ANNs and their related masks. The labelling system, which will be used throughout this chapter is defined in table 7.1.

Once trained, all the masks, ERMV7-1, ERMV7-2 and ERMV7- 3, are tested for all 91 configurations. We categorise the test data into three types for the three systems. The first category is the matched case, which means that the system was trained using the same source direction configuration

Table 7.1: System definitions and training strategies for each mask set in anechoic conditions.

Training strategy	HRIR	System	Mask set
TSV7-1	SYMARE HA02 Anechoic	V7-1	EBMV7-1 ERMV7-1
TSV7-2	SYMARE HA02 Anechoic	V7-2	EBMV7-2 ERMV7-2
TSV7-3	SYMARE HA02 Anechoic	V7-3	EBMV7-3 ERMV7-3

being tested. Hence, when testing mask ERMV7-1, trained using strategy TSV7-1, all the test examples belong to the matched case. Configuration [−20, 20], for example, belongs in the matched case category in all three systems because this pair of directions appears in the training data for all of them. The second category of test data is the semi-matched case. Here, only one interferer direction out of the two directions in each pair has been used in training the ANN. For example, when testing system V7-2, the configuration [−10, 20] is a semi-matched type test data, since both −10° and 20° azimuth angle appear in the training data, but never as a pair (i.e. only the symmetrical pairs [−10, 10] and [−20, 20]) are present in the training data, but not [−10, 20] together. The last category of test data is the unmatched case. For example, [−10, 10] and [−10, 30] are both unmatched cases in the third test strategy, TSV7-3, because neither of these directions are presented during the training phase. This approach to testing not only assesses how well each system has learned each matched case, but also reveals how well it can estimate the ratio mask for their previously unseen semi-matched and unmatched input conditions.

7.1.2 Establishing the ANN topologies

The same procedure previously described in section 6.2.1 is applied here to define the appropriate number of neurons in the ANN topology for the three

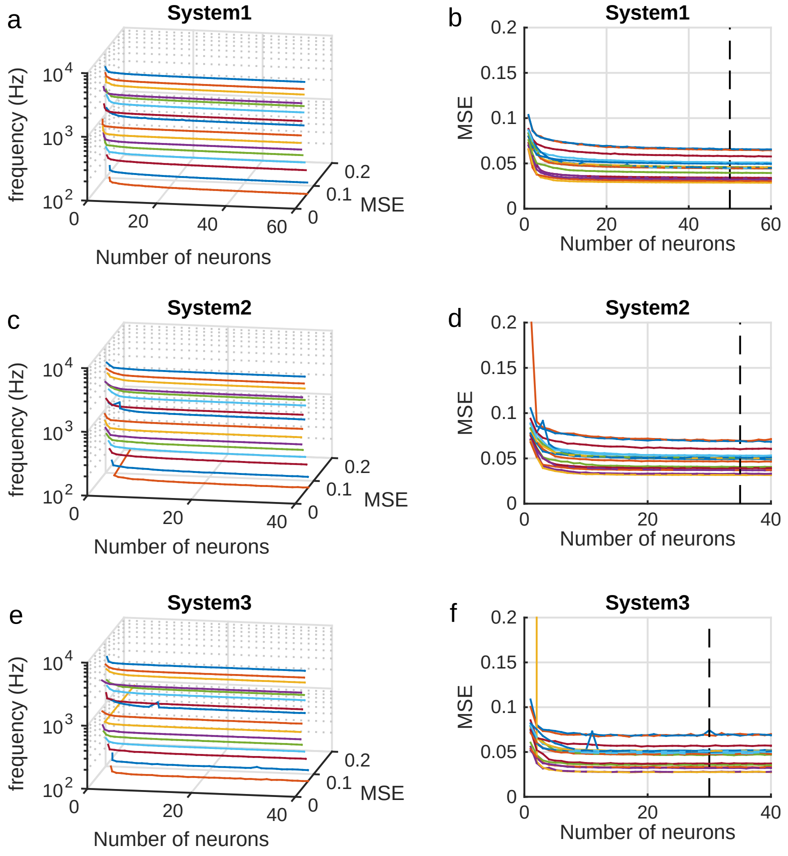


Figure 7.1: MSE performance of ANNs across frequency as a function of the number of neurons under three sources configuration for three systems in anechoic environment. (a) The system is trained at all the possible combinations of interferer's directions. (b) The projection of MSE performance of system 1 in 2D. (c) The system is only trained at all the symmetrical directions of interferers. (d) The projection of MSE performance of system 2 in 2D. (e) The system is trained partially at symmetrical directions of interferers. (f) The projection of MSE performance of system 3 in 2D.

systems. This is to ensure that all the networks contain sufficient neurons to reach a near-minimum level of mean-square error (MSE). Figure 7.1 shows the resulting MSEs from applying each training strategy to a multi-layer perceptron with one hidden layer for which the number of hidden units is systematically varied. Each graph includes a dashed line showing the size of the ANN chosen, based on 10-fold cross-validation (section 5.3.2). The size of the hidden layers for systems V7-1, V7-2 and V7-3 is 50, 35 and 30 neurons, respectively.

7.1.3 Relative importance

For the purpose of assessing the relative importance of each input, 100 copies of each of the ANN topologies in the three systems are trained using their respective training strategies. The local SNR of the training data is set to 0 dB. The training samples are drawn from the training speech corpus in TIMIT (John et al., 1993) and prepared using the HRIRs for subject HA02 from the SYMARE database (Guillon and Zolfaghari, 2012), as described in section 6.1.1.

To determine the relative importance of each input type, the ANNs are then analysed using both Garson's method and the connection weights method, described in section 5.4. The results using Garson's method are shown in figure 7.2 and those for the connection weights method are shown in figure 7.3. Figures 7.2 (a) and 7.3 (a) show the histogram envelopes, obtained using the same procedure as in section 6.2.2, for all seven types of inputs. Figures 7.2 (b) and 7.3 (b) show the relative importance of each type of input as a function of frequency on an ERB scale. It can be seen that ILD dominates IPD at mid-to-high frequencies, which is consistent with the comparable result for the seven-input topology using fixed interferers in figure 6.10 (section 6.3.2).

With variable direction inputs, however, we see that below 500 Hz interaural coherence has become a highly dominant cue in the connection weights

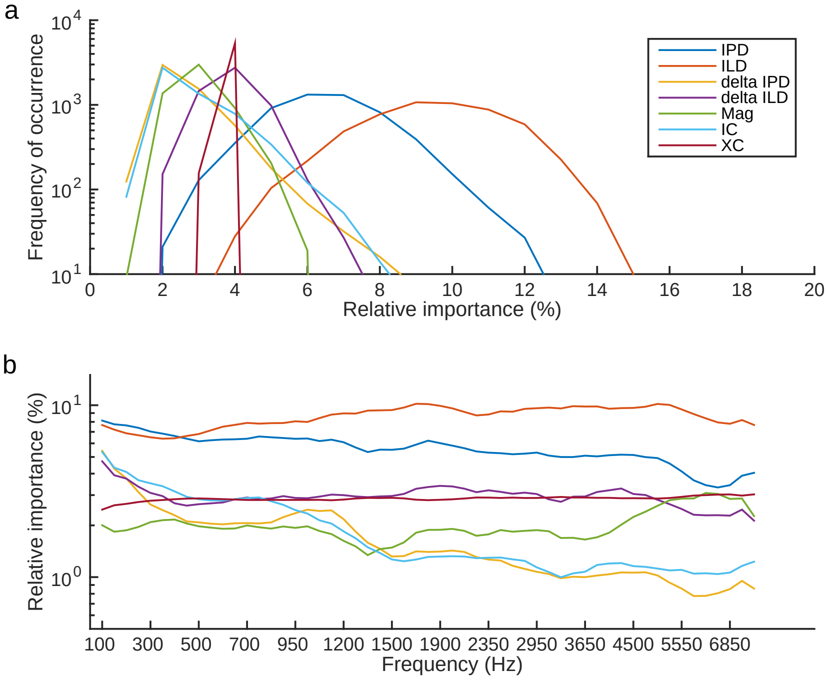


Figure 7.2: Relative importance of the inputs IPD, ΔIPD, ILD, ΔILD, magnitude (Mag), interaural coherence (IC) and cross-correlation coefficients (XC) using Garson's method in the variable three-source configuration in anechoic conditions. (a) Relative importance histogram envelopes for system V7-1. (b) Smoothed relative importance of each type of input from 100 Hz to 8 kHz on an ERB scale.

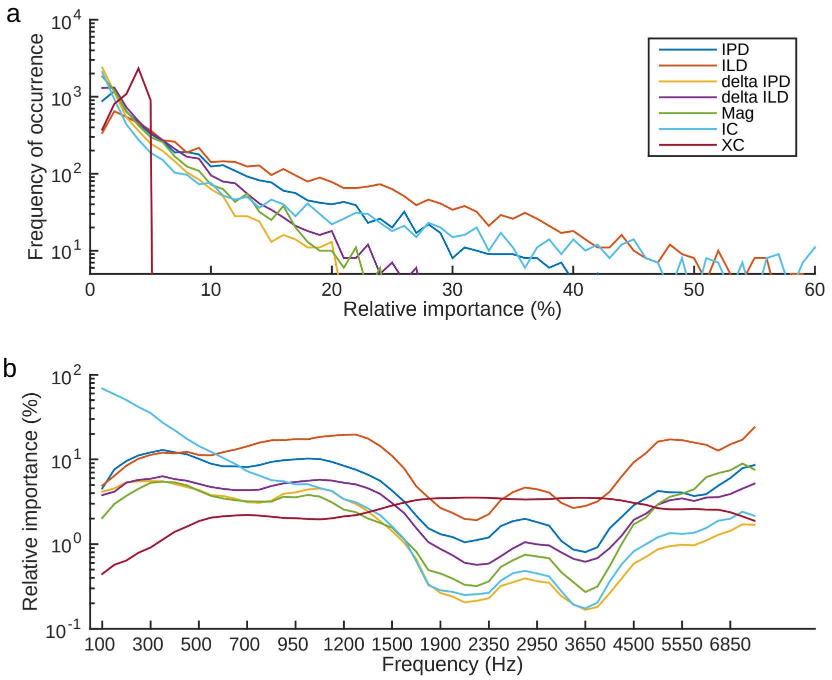


Figure 7.3: Relative importance of the inputs IPD, ΔIPD, ILD, ΔILD, magnitude (Mag), interaural coherence (IC) and cross-correlation coefficients (XC) using the connection weights method in the variable three-source configuration in anechoic conditions. (a) Relative importance histogram envelopes for system V7-1. (b) Smoothed relative importance of each type of input from 100 Hz to 8 kHz on an ERB scale.

method. Apart from this, both RI metrics indicate that interaural coherence contributes very little to mask estimation. The cross-correlation coefficients maintain their elevated importance compared with the fixed direction scenario throughout the upper range of the spectrum. This confirms that the 27 cross-correlation coefficients clearly contain strong directional cues at mid and high frequencies. Their relatively low importance at low frequencies is likely to be due to the poor time lag resolution of the cross-correlation function in this region of the spectrum.

7.1.4 Target speech segregation STOI performance

Once again, we apply the STOI performance measure to estimate the intelligibility improvement of the target speech afforded by the three variants of the V7 network topologies, V7-1, V7-2 and V7-3, using training strategies TSV7-1, TSV7-2 and TSV7-3, respectively. There are 91 different direction combinations when independent movement of the two interferers is allowed in 10° steps. For each of these 91 direction combinations, we generate a test case. Each case contains 50 mixtures and they all have a local SNR of 0 dB in the left channel before mixing.

As a preliminary check that the results of the three new systems appear reasonable, we first carry out a comparison of their STOI scores with the fixed direction system F7, originally shown in table 6.7. The results for F7 are repeated in table 7.2 and compared with the scores for the three new variable-direction mask estimators using only the fixed interferer directions. Mask estimator ERMV7-1 produces the highest STOI score of the three systems and this matches the STOI score for ERMF7 very closely. This confirms that system V7-1 has successfully learnt the fixed-direction pair, despite also being presented with 90 other direction pair combinations.

The STOI scores for the unprocessed mixtures (UPMs) for all direction combinations are shown in figure 7.4. The scores are measured in the left channel only, since the results for the right channel are similar. They lie in

Table 7.2: Comparison of STOI scores between the fixed direction system ratio mask ERMF7 and the three variable-direction masks ERMV7-1/2/3 for the same interferer directions.

Systems	STOI		
	-5 dB	0 dB	5 dB
ERMF7	0.7985	0.8697	0.9375
ERMV7-1	0.7868	0.8613	0.9293
ERMV7-2	0.7808	0.8576	0.9274
ERMV7-3	0.7559	0.8444	0.9223

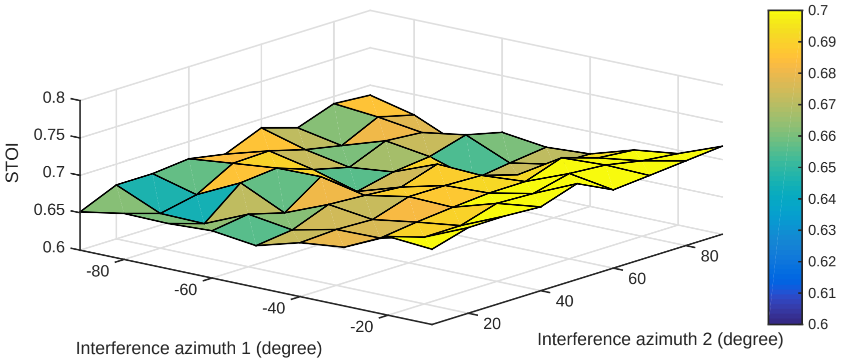


Figure 7.4: STOI scores for the unprocessed mixtures using all 91 test cases defined in TSV7-1 in the three-source anechoic configuration. The mean value is 0.6814 and standard deviation is 0.0187. The mixtures have 0 dB local SNR before spatialisation.

the range 0.65 to 0.7 with a mean value of 0.6814 and a standard deviation 0.0187. Visual inspection of the figure suggests that the variation in STOI scores displays a dependence on interferer position.

Mask ERMV7-2 compared with ERMV7-1 shows a small decline in STOI score of about 1%. In particular, the performance for the semi-matched cases is very close to the fully trained system V7-1. This is also confirmed in figure 7.5 which shows the similarity between the results of all the 91 test cases for both ERMV7-1 and ERMV7-2. For mask estimator ERMV7-3, however, it can be seen that relatively low STOI scores occur when either interferer 1 is at $-10°$ or when interferer 2 is at $10°$, which are boundary unmatched cases

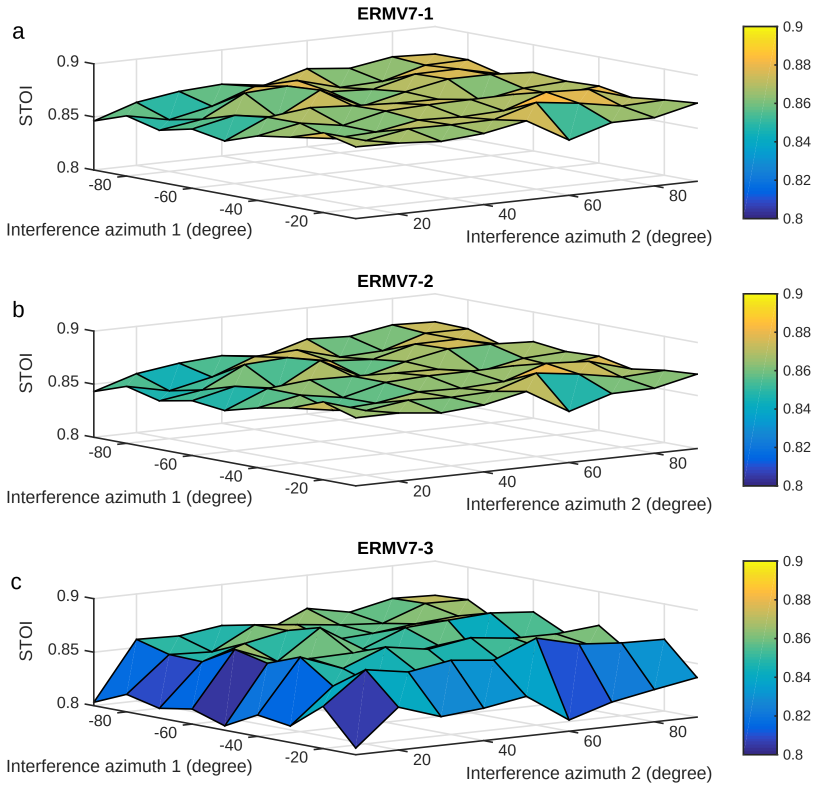


Figure 7.5: STOI scores for the three mask estimators ERMV7-1/2/3, using all 91 test cases defined in TSV7-1 in the three-source anechoic configuration. (a) Results for mask ERMV7-1 trained using strategy TSV7-1. The mean value is 0.8676 and standard deviation is 0.0077. (b) Results for mask ERMV7-2 trained using TSV7-2. The mean value is 0.8644 and standard deviation is 0.0079. (c) Results for mask ERMV7-3 trained using TSV7-3. The mean value is 0.8527 and standard deviation is 0.0189.

adjacent to the target. Furthermore, due to the relatively poor performance for the unmatched test cases in general, the STOI scores for ERMV7-3 exhibit a higher standard deviation of 0.0189.

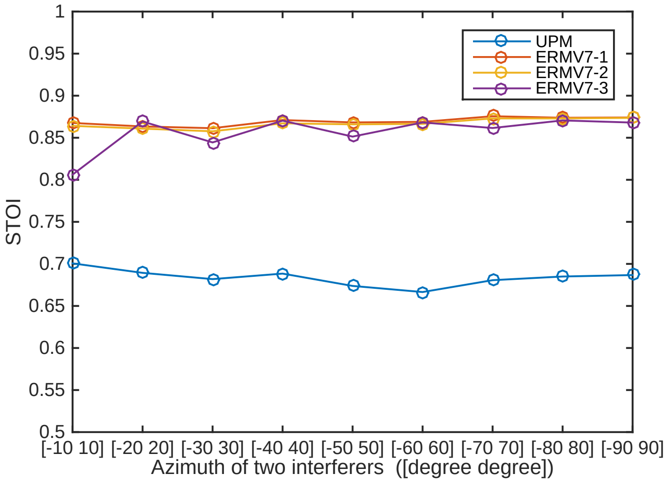


Figure 7.6: STOI scores for the original mixture with 0 dB local SNR before spatialisation and the mask estimators ERMV7-1/2/3 for the test cases on the symmetric diagonal in the three-source variable direction anechoic configuration.

To give a more detailed view of the results, figure 7.6 shows a 2D section of the STOI scores through the symmetric diagonal ([−10, 10] [−20, 20] [−30, 30] ... [−90 90]) of the 3D plots in figure 7.5. Both ERMV7-1 and ERMV7-2 have been trained at these nine direction combinations. Hence tests at these configurations represent matched cases. The results show they perform very similarly, with STOI scores which are close to 0.88 at all nine diagonal positions and the two lines in the figure almost entirely coincide. As expected, mask ERMV7-3 also yields similar STOI scores in the matched case directions. However, the results are slightly worse in the unmatched ones. When both interferers are close to the target signal, at [−10, 10], the STOI score falls sharply to about 0.8. This is not only because it is an unmatched case, but also because the system exploits several binaural features, which become very similar in value to the target speech being segregated. Therefore it becomes harder for the ANN in system V7-3 to distinguish the target from the interferers due to the small binaural cue differences. When the interferers have a greater angular separation from the target the differences in STOI score between matched cases using mask ERMV7-1 and the corresponding unmatched cases using ERMV7-3 decrease. This can be accounted for in terms of the greater differences in the binaural cues in these directions. The difference is smallest for the interferer direction pair [−90, 90], for which the interferers are angularly most separated from the target and from each other.

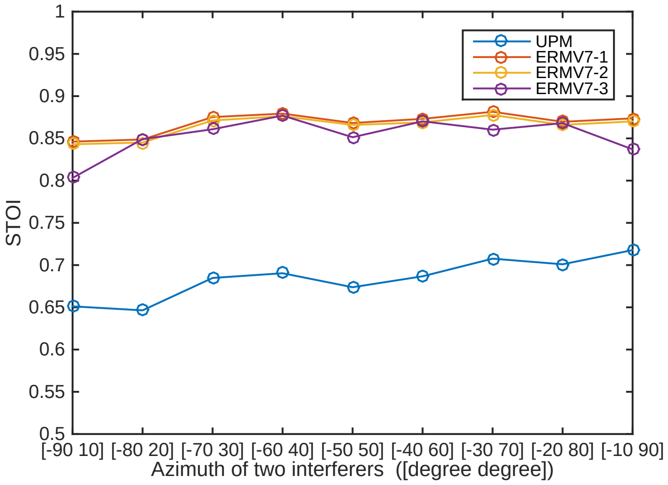
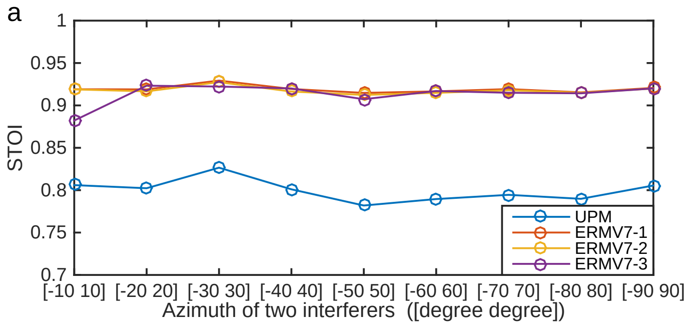


Figure 7.7: STOI scores for the original mixture with 0 dB local SNR before spatialisation and after using the masks ERMV7-1/2/3 for test directions on the antisymmetric diagonal in the three-source anechoic configuration.

The STOI scores for all three masks along the anti-symmetric diagonal ([−90, 10], [−80, 20] [−70, 30] ... [−10, 90]) are shown in figure 7.7. This plot contains examples of matched, semi-matched and unmatched test directions. These nine direction pairs are matched cases for ERMV7-1, and semi-matched cases for ERMV7-2. Again, the STOI scores from both of these masks are very similar and they coincide on the plot. For ERMV7-3, as expected, the STOI scores for the semi-matched directions follow the same pattern as ERMV7-2 and they are very close to the score obtained using mask ERMV7-1. The STOI scores at both corners (i.e. [−90, 10] and [−10, 90]) are lower than the others. Not only are these examples of unmatched directions, but also one interferer is close to the target so that it becomes more difficult for the ANN in system V7-3 to segregate the target. Overall, mask ERMV7-3 displays a symmetrical pattern of STOI scores on the surface plot in figure 7.5.



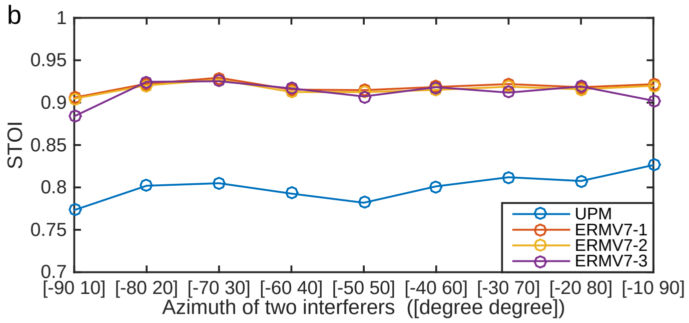


Figure 7.8: STOI scores for the original mixture with 5 dB local SNR before spatialisation and for the three masks ERMV7-1/2/3 for test direction pairs along the (a) symmetric and (b) anti-symmetric diagonals in the three-source anechoic configuration.

To examine the ability of each estimator to generalise the training data for input mixtures with different local SNRs, we further evaluate the performance of the three ratio masks at 5 dB and -5 dB. As usual, the local SNRs are measured before spatialising each sound source. Surface plots showing the STOI scores for all 91 test cases are available in appendix figure A.1 and A.2 for the 5 dB and -5 dB local SNRs, respectively.

2D cross-sections extracted from figure A.1 for the 5 dB SNR case are shown in figure 7.8 and for the -5 dB case in figure 7.9. Figure 7.8 (a) shows the STOI scores for the symmetric diagonal. ERMV7-1 and ERMV7-2 yield

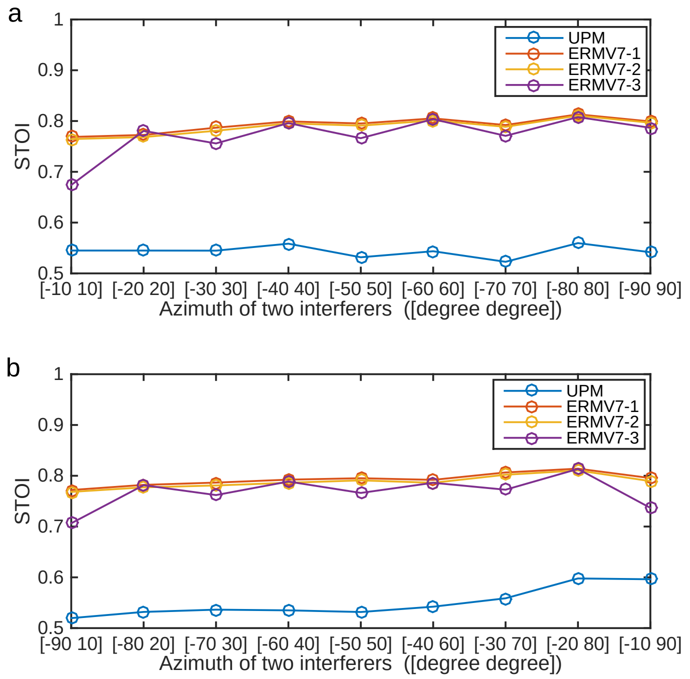


Figure 7.9: STOI scores for the original mixture with -5 dB local SNR before spatialisation and for the three masks ERMV7-1/2/3 for test direction pairs along the (a) symmetric and (b) anti-symmetric diagonals in the three-source anechoic configuration.

very similar results which consistently improve the scores by more than 10% from 0.8 to 0.9 for each pair of interferer directions. Mask ERMV7-3, however, improves by only approximately 8% for the test direction pair $[-10, 10]$, because it is an unmatched test case for this mask and both interferers are also very close to the target source. For the other test directions for ERMV7-3, the difference between systems is smaller. In figure 7.8 (a), all masks improve the STOI score by more than 10%, except ERMF7-3 for test cases $[-90, 10]$ and $[-10, 90]$. The same reason applies; the test cases for a system increase the difficulty of segregating the target source when they are unmatched and one interferer is very close to the target. For the -5 dB local SNR test condition, a similar pattern of results is obtained. Now, however, for both masks ERMV7-1 and ERMV7-2, the STOI score improvement is greater, at over 20% for all directions on both diagonals.

7.1.5 Target speech segregation PESQ performance

Figure 7.10 shows the PESQ score which is obtained for the original unprocessed mixture for every direction-pair combination. All the PESQ scores lie in the range 1.5 to 2, substantially below the ideal score of 4.5.

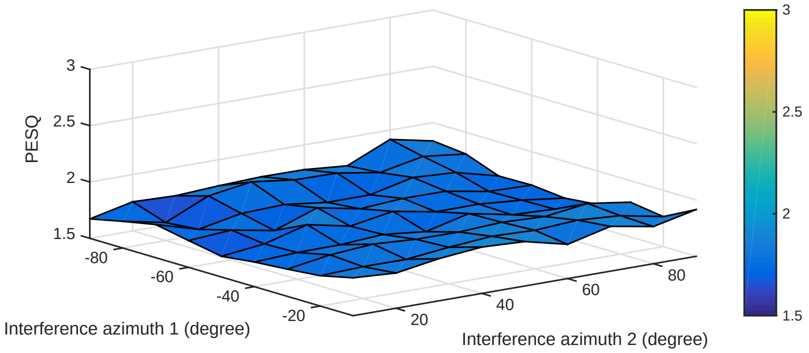


Figure 7.10: PESQ scores for the unprocessed mixtures using all 91 test cases defined in TSV7-1 in the three-source anechoic configuration. The mean value is 1.7852 and standard deviation is 0.0611. The mixtures have 0 dB local SNR before spatialisation.

As shown in figure 7.11, masks ERMV7-1 and ERMV7-2 both output similar speech quality. ERMV7-3 exhibits the poorest performance, which occurs at the two edges where one or both interferers lie close to the target speech. The PESQ score is improved to above 2.5 for all systems except for the unmatched cases in ERMV7-3.

Similarly, sections through the symmetric and antisymmetric diagonals from the surface plot in figure 7.11 are compared in more detail in figures 7.12 and 7.13. In figure 7.12, for both mask ERMV7-1 and ERMV7-2 the PESQ score is improved by 0.8 at $[-10, 10]$, and the improvement increases up to 1 when the interferers are far from the target at $[-90, 90]$. For all masks, the test results for matched direction pairs are very similar, which indicates

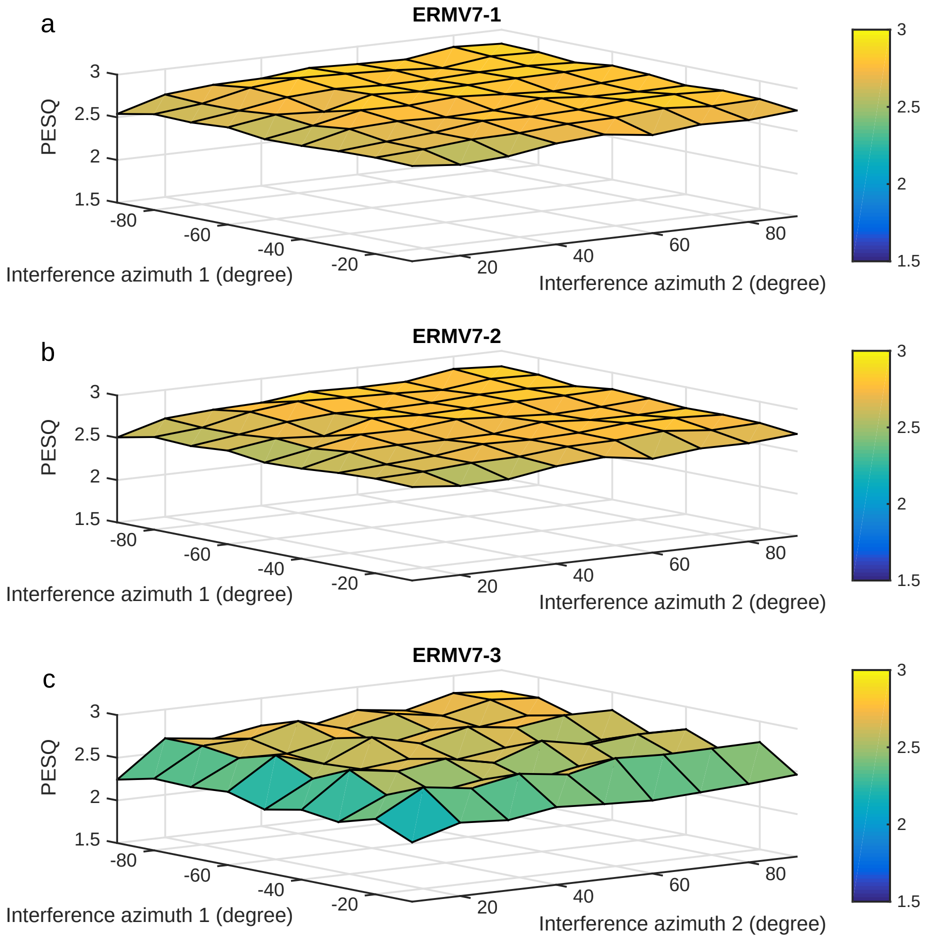


Figure 7.11: PESQ scores for the three mask estimators ERMV7-1/2/3, using all 91 test cases defined in TSV7-1 in the three-source anechoic configuration. (a) Results for mask ERMV7-1 trained using strategy TSV7-1. The mean value is 2.7480 and standard deviation is 0.0762. (b) Results for mask ERMV7-2 trained using TSV7-2. The mean value is 2.7162 and standard deviation is 0.0765. (c) Results for maskERMV7-3 trained using TSV7-3. The mean value is 2.5960 and standard deviation is 0.1584.

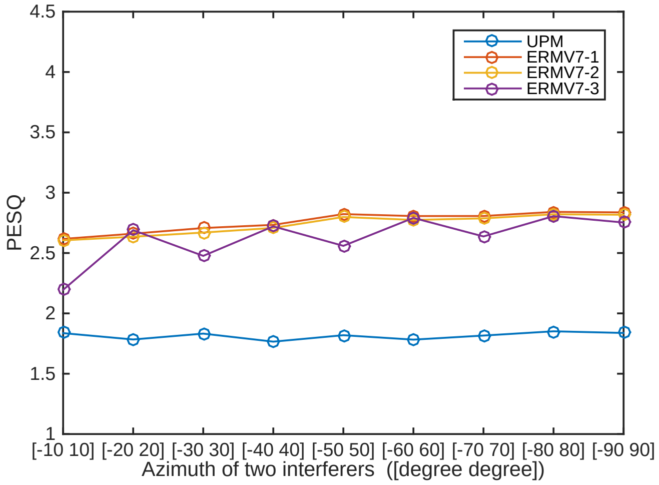


Figure 7.12: PESQ measurement of the unprocessed mixture with 0 dB local SNR before spatialisation and the three masks ERMV7-1/2/3 for test cases on the symmetric diagonal in the three-source anechoic configuration.

that all systems perform equally well for the conditions where learning is near-optimal. As expected, the performance of mask ERMV7-3 for the unmatched case ($[-10, 10]$) is significantly worse than the same direction pair for masks ERMV7-1 and ERMV7-2, for which this is a matched case. As noted before, the performance gap between matched and unmatched cases starts to decrease when the interferers are further away from the target since the binaural cues steadily become more distinct.

The results for the antisymmetric diagonal are shown in figure 7.13. It shows that all three masks exhibit similar PESQ scores for the matched cases that they have in common. The similarity in scores between ERMV7-1 and the corresponding semi-matched cases for ERMV7-2 demonstrate that the latter system has generalised well for this condition. There is a difference of approximately 0.5 between the PESQ scores for the matched case in ERMV7-1 and the unmatched case in ERMV7-3, reveals some short-comings in generalisation where neither interferer direction was used during training.

The PESQ performance of the masks for the 5 dB and -5 dB local SNR

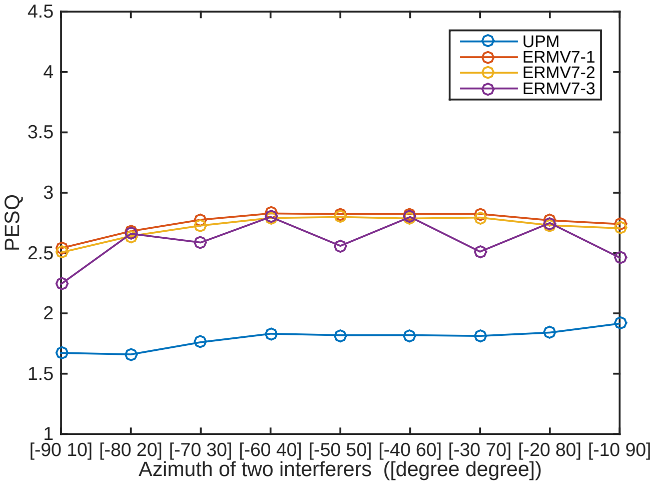


Figure 7.13: PESQ scores for the original mixture with 0 dB local SNR before spatialisation and the three estimators for test cases on the antisymmetric diagonal in the three-source anechoic configuration.

conditions are also evaluated. The STOI scores for all 91 test direction pairs are shown in figures A.3 and A.4 for 5 dB and -5 dB SNRs, respectively, in appendix A. The results along the symmetric and antisymmetric diagonals are extracted and are shown in figures 7.14 and 7.15. For the 5 dB SNR test condition, the PESQ score is improved by at least 0.8 (16%) for each test direction pair, from approximately 2 to 2.8. For the -5 dB SNR condition, the PESQ score is improved slightly less, but the improvement is at least 0.4 (8%) for each test direction pair, and lifts the score from approximately 1.5 up to at least 1.8 for unmatched cases, slightly over 2 for semi-matched cases and over 2.3 for matched cases. The results for all three estimators follow a similar pattern to the results for the 0 dB SNR test condition. ERMV7-1 and ERMV7-2 perform similarly, and the unmatched test cases for mask ERMV7-3 produce the lowest scores.

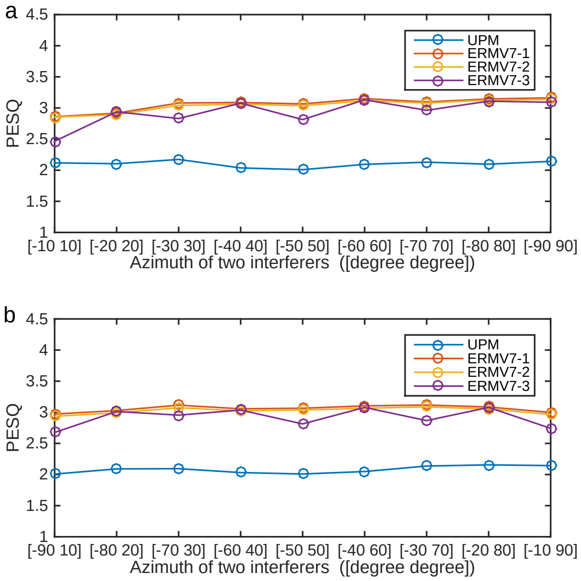


Figure 7.14: PESQ scores for the original mixture with 5 dB local SNR before spatialisation and using the three masks for test direction pairs on the two diagonals in the three-source anechoic configuration. (a) on the symmetric diagonal and (b) on the antisymmetric diagonal.

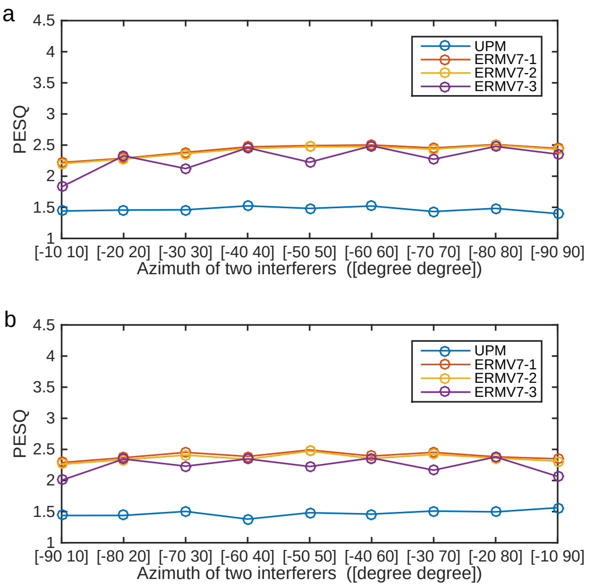


Figure 7.15: PESQ scores for the original mixture with -5 dB local SNR before spatialisation and using the three masks for test direction pairs on the two diagonals in the three-source anechoic configuration. (a) on the symmetric diagonal and (b) on the antisymmetric diagonal.

7.2 Variable direction interferers in reverberation environments

In previous sections, we have demonstrated that OCM improves speech intelligibility and quality using the STOI and PESQ metrics in simulated anechoic environments. However, most real environments involve substantial reverberation. Therefore, in this section we evaluate how OCM performs in reverberant conditions and how the contributions to mask estimation by the features in the binaural mixture are changed by the new acoustic conditions.

To create the reverberation test conditions, we use a binaural room impulse response (BRIR) from the AIR database (Jeub et al., 2009). It is measured in a stairway hall of width 5.2 m and length 7 m. The BRIR measurements were made using the HMSII.3 dummy head (Jeub et al., 2010b) oriented in 15° azimuthal steps in the range $-90°$ to $90°$ relative to the sound source at a distance of 1 m. Other binaural impulse responses in anechoic and reverberant conditions are used in the next chapter.

7.2.1 Training and testing strategies

The ANN topology V7-2 (section 7.1.2), previously applied in an anechoic configuration, is retrained to create ratio mask ERMVR7-2 for use in this reverberant configuration. The new system is labeled as VR7-2. V7-3 is also retrained in the same reverberant conditions using a different strategy to create ratio mask estimator ERMVR7-3. The new system is labeled as VR7-3. In detail, VR7-2 is trained using the strategy, denoted as TSVR7-2, for six of the 36 possible combinations of interferer direction pairs available using the AIR database BRIRs. Specifically, TSVR7-2 places the target at 0° azimuth with interferers at $[-90, 90]$ to $[-15, 15]$ in 15° steps for each interferer direction. System VR7-3 is trained with the interferers organised as the three direction pairs $[-15, 15]$, $[-45, 45]$ and $[-75, 75]$ only. This training strategy is denoted as TSVR7-3. Therefore, as for the systems trained in anechoic conditions in section 7.1.1, the 36 test mixtures can be split into three categories: matched, semi-matched and unmatched. The nomenclature for the reverberation simulations is presented in table 7.3.

Table 7.3: System definitions and training strategies for each mask set in reverberation

Training strategy	HRIR	System	Mask set
TSVR7-1	-	-	-
TSVR7-2	AIR BRIR database	VR7-2	EBMVR7-2 ERMVR7-2
TSVR7-3	AIR BRIR database	VR7-3	EBMVR7-3 ERMVR7-3

To create the training set, we generate 5,000 items of training data for each interferer direction pair. The appropriate subsets of this training data are used to train VR7-2 (using training strategy TSVR7-2) and VR7-3 (using training strategy TSVR7-3) resulting in ratio masks ERMVR7-2 and ERMVR7-3, respectively.

It is important to note that, in this initial work, dereverberation is not considered here. We aim to segregate the target speech together with its associated reverberation. No attempt is made to remove the reverberation. Hence the training target is the ratio mask for the reverberant target speech in the mixture.

The evaluation is conducted in a three-source configuration. The target speech is located at 0° azimuth, and two speech interferers are placed one on each side. To the left of the listener there are six possible directions from $-90°$ to $-15°$ in 15° steps. There is a symmetrical arrangement for the interferer on the right-hand side. This results in 36 different combinations of interferer direction pairs for testing.

7.2.2 Relative importance

The relative importance calculated using Garson's method is shown in figure 7.16. The results for the connection weights method are presented in figure 7.17. In both sets of results, compared with the anechoic scenario (figure 7.3), the importance of ILD remains high, but at all frequencies IPD is now much less important than ILD. ΔIPD has become the least important input. We hypothesise that the phase relationship between the left and right channels is very sensitive to the effect of adding even relatively low levels of reverberation. The temporal smearing caused by reverberation also has a detrimental effect on ILD, but its impact remains relatively low until the reverberation is of a similar energy to the direct sound. This may account for the reduced importance in reverberation of IPD and ΔIPD compared with ILD and ΔILD.

The importance of IC has risen and exceeded the importance of ΔIPD. This increase in importance is consistent with the results from Jeub et al. (2010a) and Alinaghi (2013), where the importance of IC has been shown to increase in reverberant environments.

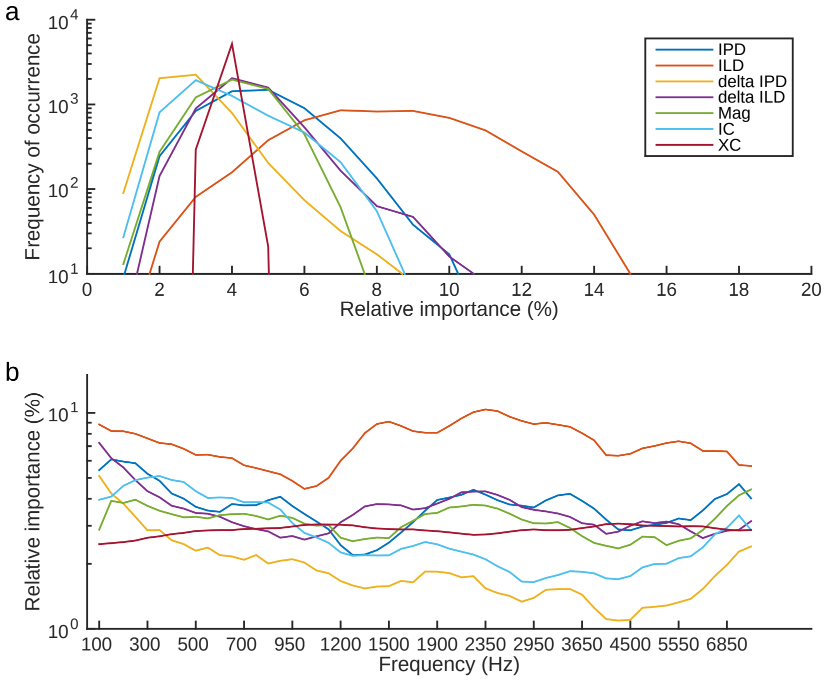


Figure 7.16: Relative importance of the inputs IPD, ΔIPD, ILD, ΔILD, magnitude (Mag), interaural coherence (IC) and cross-correlation coefficients (XC) using Garson's method in the variable three-source configuration in reverberant conditions. (a) Relative importance histogram envelopes for system VR7-2. (b) Smoothed relative importance of each type of input from 100 Hz to 8 kHz on an ERB scale.

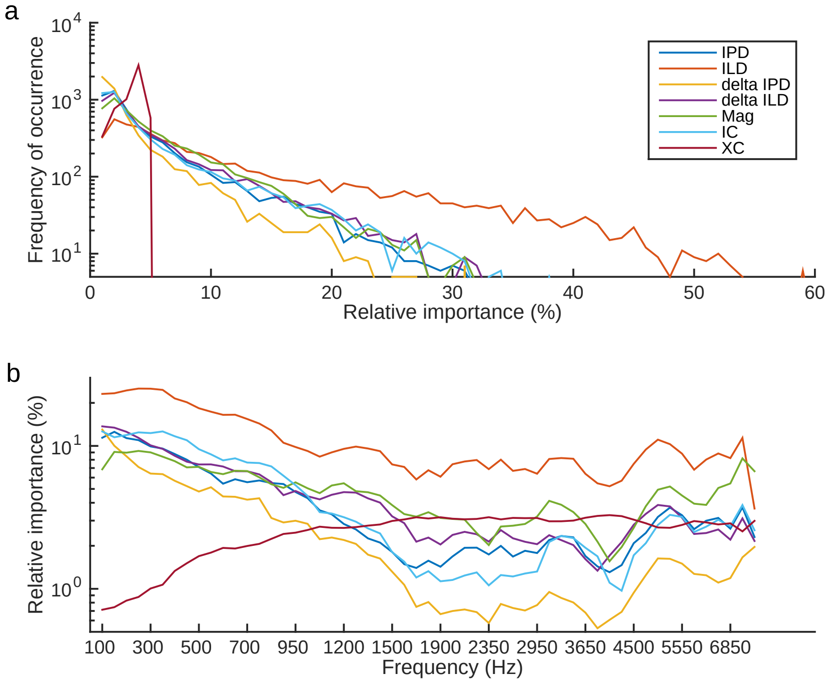


Figure 7.17: Relative importance of the inputs IPD, ΔIPD, ILD, ΔILD, magnitude (Mag), interaural coherence (IC) and cross-correlation coefficients (XC) using the connection weights method in the variable three-source configuration in reverberant conditions. (a) Relative importance histogram envelopes for system VR7-2. (b) Smoothed relative importance of each type of input from 100 Hz to 8 kHz on an ERB scale.

7.2.3 Target speech segregation STOI performance

The intelligibility of the original reverberant mixture before processing and of the segregated target speech after applying each ratio mask estimator are assessed using the STOI performance metric. Individual sources are set to 0 dB local SNR, filtered and mixed. For both systems, the STOI score is computed for each of the 36 direction pairs using 50 binaural test mixtures.

Figure 7.18 shows the STOI scores computed for the left channel of the original mixture and for the outputs from ratio masks ERMVR7-2 and ERMVR7-3, trained using training strategies TSVR7-2 and TSVR7-3, respectively. These were evaluated by treating the reverberant target speech as the desired segregated target.

The original mixtures display STOI scores of between 0.55 and 0.65 with a mean value of 0.5963 and a standard deviation of 0.0264. ERMVR7-2 and ERMVR7-3 both improve the speech intelligibility. The improvement achieved by mask ERMVR7-2 is greatest. It yields a STOI score with a mean value of 0.7710 and a standard deviation of 0.0165. The minimum STOI score for ERMVR7-2 is approximately 0.73 and the maximum is approximately 0.8. ERMVR7-3 performs less well than ERMVR7-2 (mean STOI value is 0.7580), especially for the unmatched interferer direction pairs. Hence, the STOI score surface in the 3D plot of figure 7.18 (c) is not as flat as its counterpart in figure 7.18 (b). This is reflected in the standard deviation value of 0.0193 which is greater than in the value for ERMVR7-2. Nevertheless, there is still more than a 10% improvement in the STOI scores for all the test interferer direction pairs using ERMVR7-3.

To compare the performance of the two masks more closely, the STOI scores for interferer direction pairs along the symmetric and antisymmetric diagonals are shown in figure 7.19.

Figure 7.19 (a) shows the STOI scores for all symmetric interferer direction pairs. It can be seen that when interferers are close to the target, the

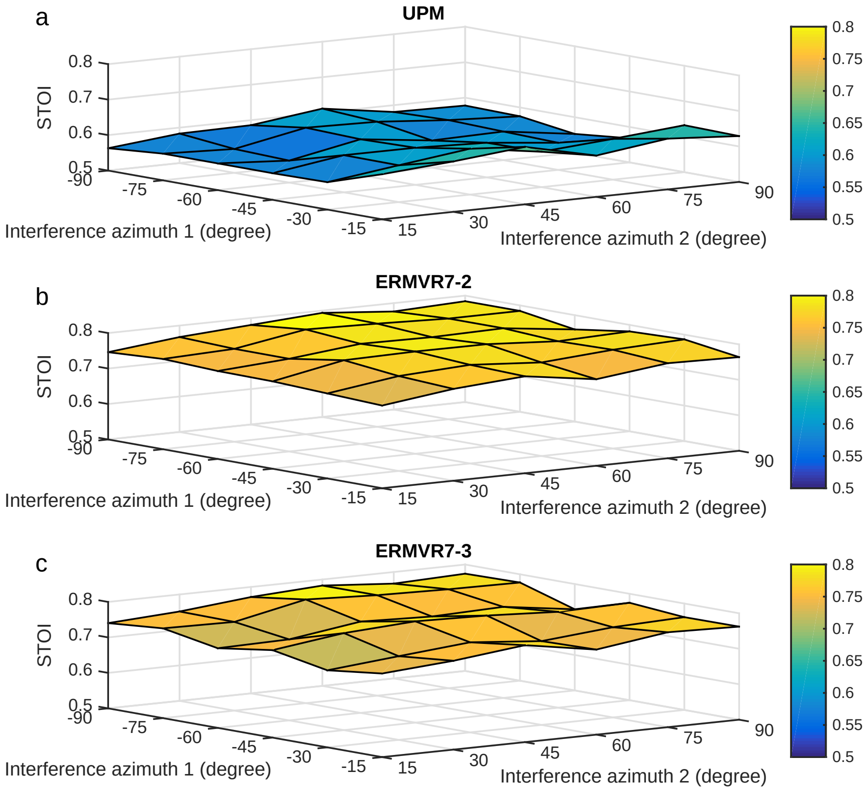


Figure 7.18: STOI scores for the unprocessed mixtures and for the outputs from the two mask estimators, ERMVR7-2 and ERMVR7-3, for all direction pair combinations in the three-source reverberant configuration. The mixtures are 0 dB local SNR before spatialisation. The target is located at 0° azimuth and the two interferers are placed at $-90°$ to 0° and 0° to 90° azimuth, respectively, with 15° steps. (a) The STOI scores for the unprocessed mixture. The mean value is 0.5963 and standard deviation is 0.0264. (b) The STOI scores for ERMVR7-2, trained using TSVR7-2. The mean value is 0.7710 and standard deviation is 0.0165. (c) The STOI scores for ERMVR7-3, trained using TSVR7-3. The mean value is 0.7580 and standard deviation is 0.0193.

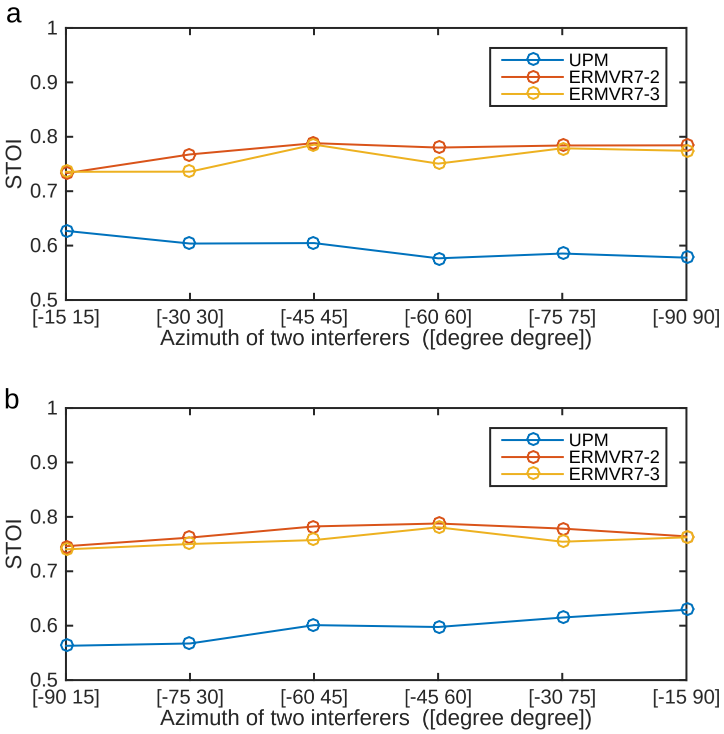


Figure 7.19: STOI scores for the unprocessed mixture with 0 dB local SNR before spatialisation and the masks ERMVR7-2 and ERMVR7-3 in the three-source reverberant configuration for the test cases (a) on the symmetric diagonal and (b) on the antisymmetric diagonal (where sources are 105° apart).

improvement in the STOI score for both estimators is less than when both interferers are further away. The improvements are 11% and 20% at $[-15, 15]$ and $[-90, 90]$, respectively, for ERMVR7-2. For ERMVR7-3, the interferer direction pair test cases in figure 7.19 (a) consist of both matched and unmatched cases. The STOI score from both systems are very similar, while there is about a 3% difference at $[-30, 30]$ and $[-60, 60]$, which are matched direction pair cases for ERMVR7-2, but unmatched cases for ERMVR7-3. The third unmatched case, $[-90, 90]$, for ERMVR7-3 produces a STOI score 1% lower than the corresponding case for ERMVR7-2.

Figure 7.19 (b) shows the STOI performance on the antisymmetric diagonal, where the interferer pairs are always separated by 105°. On this diagonal, all the test cases for both ERMVR7-2 and ERMVR7-3 are semi-matched. However, ERMVR7-2 performs slightly better than ERMVR7-3, because both of the interferer direction pairs in the test cases are used to train system VR7-2 and only one of the pair is used train ERMVR7-3. As a consequence of this difference in training, there is a drop in STOI score for ERMVR7-3, which is a maximum of approximately 2.5% for the interferer direction pairs [−60, 45] and [−30, 75].

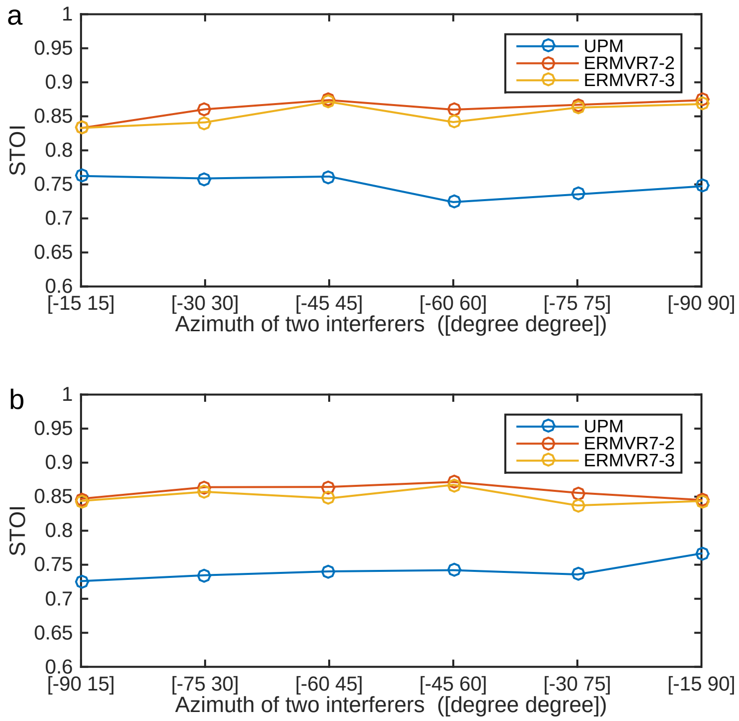


Figure 7.20: STOI scores for the unprocessed mixture with 5 dB local SNR before spatialisation and the mask estimators ERMVR7-2 and ERMVR7-3 in the three-source reverberant configuration for the test cases (a) on the symmetric diagonal and (b) on the antisymmetric diagonal (where sources are 105° apart).

Since both mask estimators are trained on a single SNR condition (0 dB), it is once again interesting to examine their ability to generalise under dif-

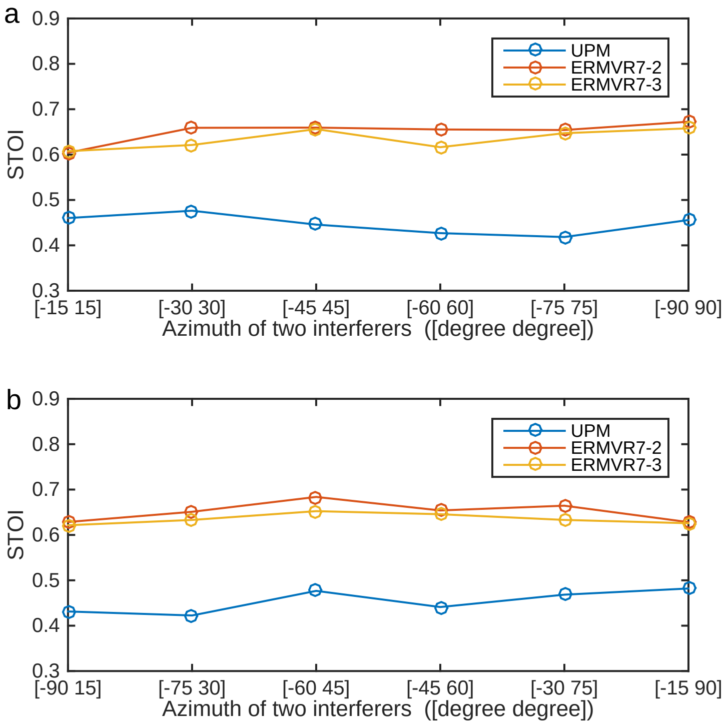


Figure 7.21: STOI scores for the unprocessed mixture with -5 dB local SNR before spatialisation and the mask estimators ERMVR7-2 and ERMVR7-3 in the three-source reverberant configuration for the test cases (a) on the symmetric diagonal and (b) on the antisymmetric diagonal (where sources are 105° apart).

ferent SNR conditions. Specifically, we test the systems using reverberant binaural input mixtures with local SNRs of 5 dB and -5 dB where, as usual, the SNR is measured before spatialisation. The STOI score surface plots for all 36 test cases may be viewed in figures A.5 and A.6 in Appendix A. We focus here on the detailed results on the symmetric and antisymmetric diagonals, shown for the 5 dB SNR and -5 dB SNR situations in figures 7.20 and 7.21, respectively.

For ERMVR7-2 in figure 7.20 there is an average improvement in STOI score of 10%. When one or both interferers are close to the target source, however, the improvement decreases to 7%. The maximum improvement in terms of STOI score is 12% for the 5 dB SNR condition. These gains are universally smaller than those at the lower SNR of 0 dB. ERMVR7-3 performs as well as ERMVR7-2 for the matched cases, while it is approximately 2% worse than ERMVR7-2 for the unmatched cases and less than 2% for the semi-matched cases. A similar trend can be observed in the -5 dB SNR test condition, shown in figure 7.21. ERMVR7-2 improves the STOI score from approximately 0.45 up to 0.63, averaging an 18% improvement in the reverberant condition. The performance gap between ERMVR7-2 and ERMVR7-3 is approximately 4% for unmatched cases and less than 4% for semi-matched cases.

7.2.4 Target speech segregation PESQ performance

In this section we analyse the impact of ratio masks ERMVR7-2 and ERMVR7-3 on target speech quality in reverberation. The PESQ score for the unprocessed mixtures and after applying both masks are compared in figure 7.22. The unprocessed (reverberant) interferer direction pairs have a PESQ score of around 2 and the standard deviation is 0.0719. For the processed signals from both estimators the PESQ scores are lifted above 2.5 in most cases. Following a similar pattern to previous results (figure 7.18), at the two edges where there is one interferer close to the target, the PESQ scores are less than this. For ERMVR7-3, the performance is not as good as ERMVR7-2 for some test cases.

A more detailed analysis of the PESQ scores is again considered for the symmetric diagonal (figure 7.23 (a)) and the antisymmetric diagonal (figure 7.23 (b)). For ERMVR7-2, the test interferer direction pairs are all matched and the results follow a similar pattern to the STOI scores. When interferers are close to the target, the PESQ score improves less than when the separation between the interferers and the target is greater. The improvements are 0.3 (6%) and 0.5 (10%) at [−15, 15] and [−90, 90], respectively. For ERMVR7-3, the PESQ performance is 0.1 (2%) and 0.05 (1%) worse than ERMVR7-2 in the unmatched direction pairs, [−30, 30] and [−90, 90], respectively. For the matched cases, both masks improve the PESQ score by similar amounts.

Figure 7.23 (b) shows the PESQ scores along the antisymmetric diagonal. Again, all the test cases along this diagonal are semi-matched for both ERMVR7-2 and ERMVR7-3, but only alternate direction pairs are used to train ERMVR7-3. However, both of the interferer directions in the test cases have been used to train the ERMVR7-2, while only one of the interferer directions is used train the ERMVR7-3, with the result that ERMVR7-2 performs better than ERMVR7-3. The two greatest PESQ score differences are approximately 0.12 (2.3%) and 0.1 (2%) in for the interferer direction pairs

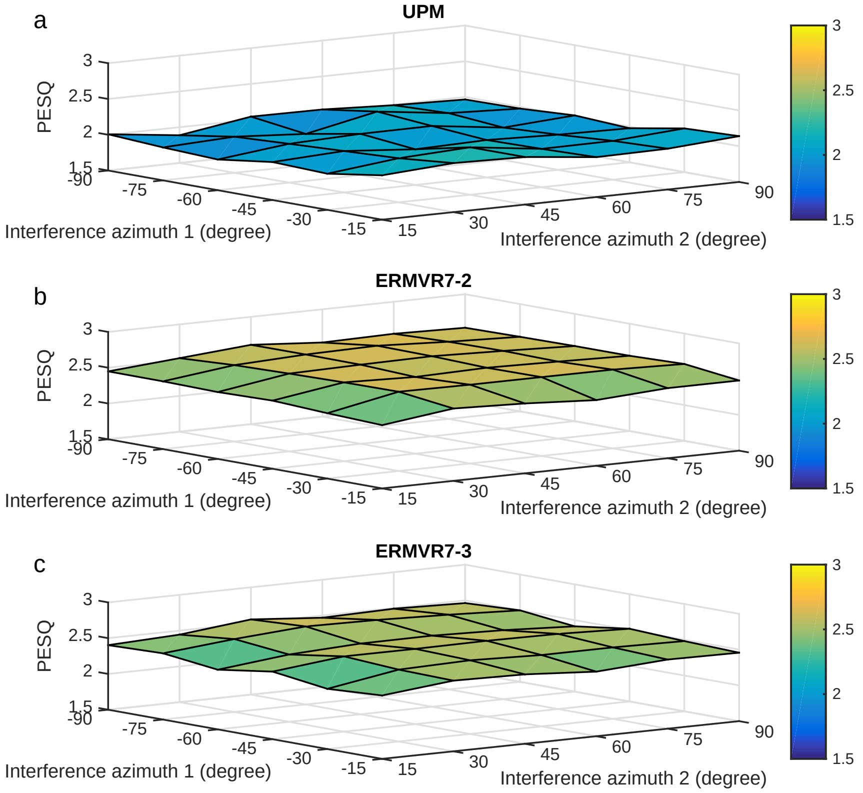


Figure 7.22: PESQ scores for the unprocessed mixtures and for the outputs from the two masks, ERMVR7-2 and ERMVR7-3, for all direction pair combinations in the three-source reverberant configuration. The mixtures are 0 dB local SNR before spatialisation. The target is located at 0° azimuth and the two interferers are placed at −90° to 0° and 0° to 90° azimuth, respectively, with 15° steps. (a) The STOI score of the original mixture. The mean value is 2.0386 and standard deviation is 0.0719. (b) The STOI score for mask ERMVR7-2, trained using TSVR-2. The mean value is 2.5408 and standard deviation is 0.0687. (c) The STOI score for mask ERMVR7-3, trained using TSVR-3. The mean value is 2.4808 and standard deviation is 0.0618.

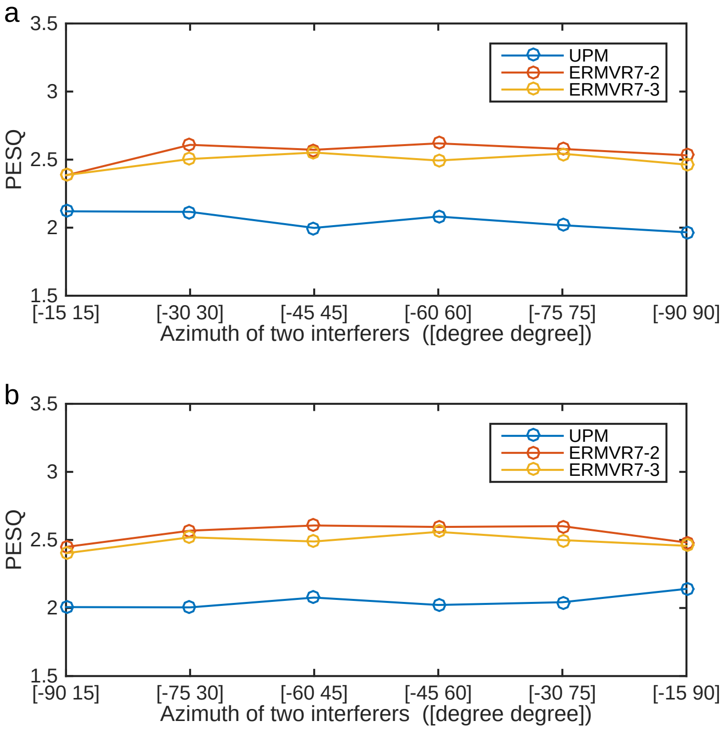


Figure 7.23: PESQ scores for the unprocessed mixture with 0 dB local SNR before spatialisation and masks ERMVR7-2 and ERMVR7-3 in the three-source reverberant configuration for the test cases (a) on the symmetric diagonal and (b) on the antisymmetric diagonal (where sources are 105° apart).

[−60, 45] and [−30, 75].

In a similar manner to the STOI analysis , we test the PESQ performance of the masks for mixture local SNRs 5 dB above and 5 dB below the 0 dB level at which they were trained, where the SNR is measured before spatialisation. The surface plot of the PESQ scores for all 36 test cases is available in figures A.7 and A.8 in Appendix A. The detailed results along the two diagonals are shown in figures 7.24 and 7.25, respectively. In a similar trend to that seen previously, there is an approximately 0.5 (10%) and 0.6 (12%) PESQ score improvement for ERMVR7-2 in the 5 dB and -5 dB SNR conditions, respectively. For ERMVR7-3 the PESQ score improvement

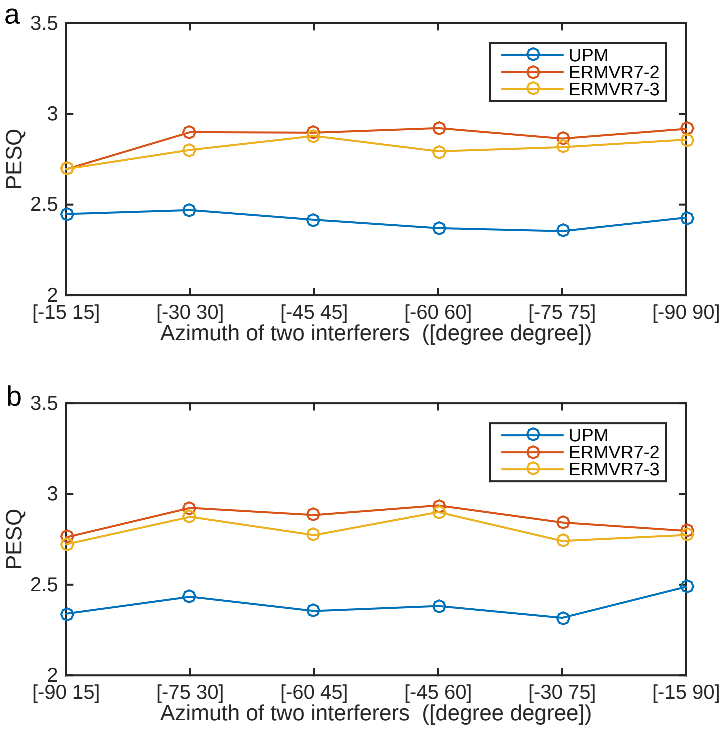


Figure 7.24: PESQ scores for the unprocessed mixture with 5 dB local SNR before spatialisation and masks ERMVR7-2 and ERMVR7-3 in the three-source reverberant configuration for the test cases (a) on the symmetric diagonal and (b) on the antisymmetric diagonal (where sources are 105° apart).

compared to ERMVR7-2 for matched test cases is very similar. ERMVR7-3 scores are approximately 0.1 (2%) lower than those for ERMVR7-2, where the interferer direction pairs for system VR7-3 are semi-matched and matched for ERMVR7-2. The difference between the matched and unmatched cases is slightly greater than 0.1 (2%).

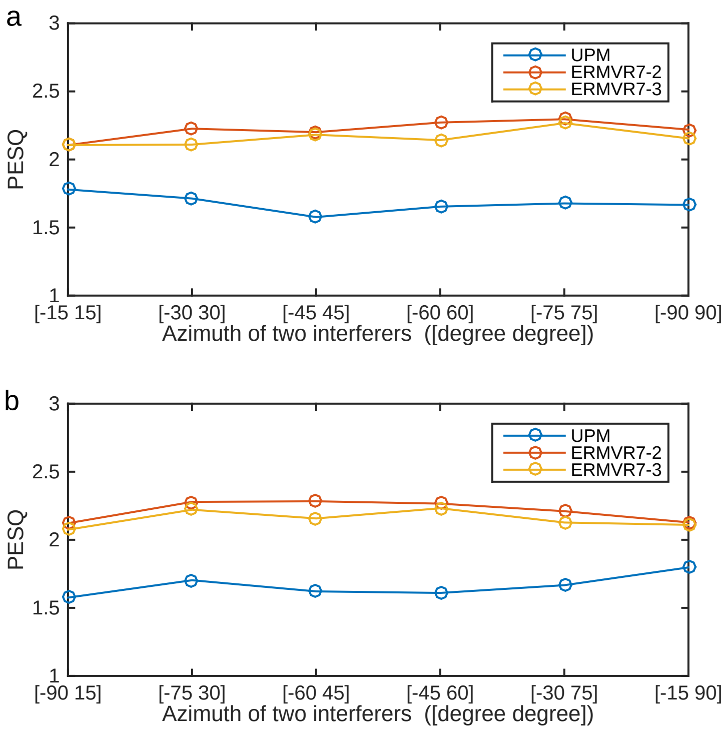


Figure 7.25: PESQ scores for the unprocessed mixture with -5 dB local SNR before spatialisation and masks ERMVR7-2 and ERMVR7-3 in the three-source reverberant configuration for the test cases (a) on the symmetric diagonal and (b) on the antisymmetric diagonal (where sources are 105° apart).

7.3 Summary

The pilot study presented in Chapter 6 shows that an ideal ratio mask outperforms a binary mask in terms of its STOI (speech intelligibility) and PESQ (speech quality) scores. However, those systems use two interferer sources whose directions are restricted to $-30°$ and $30°$ azimuth, respectively. In this chapter we have demonstrated that it is possible to relax these constraints on interferer direction with very little impact on mask performance. System F7 (with seven input types: IPD, ΔIPD, ILD, ΔILD, magnitude, interaural coherence and cross-correlation coefficients) was retrained to handle two interfering sources with varying directions in anechoic conditions. Three different training strategies were used in this variable direction scenario to produce systems V7-1, V7-2 and V7-3. By analysing the relative importance of each type of cue, we found that the importance of cross-correlation coefficients increases when the interferers are allowed to move, compared with the fixed direction scenario. This supports the notion that the 27 cross-correlation coefficients contain important directional cues which can be used by the ANN and are still useful in the reverberant condition.

Evaluating the performance of the three ratio mask estimators shows that there is an improvement in the intelligibility and in the quality of a target speech source segregated from a mixture of two directionally distinct and variable interfering speech sources. V7-1, trained using strategy TSV7-1, which covers all the tested interferer directions, produced the highest performance in terms of both intelligibility and quality.

System V7-2 was trained using strategy TSV7-2, for which interferer direction pairs always had lateral symmetry (i.e. the two sources had equal positive and negative azimuth angles, respectively). This restricted the number of distinct interferer direction pairs presented during training to only nine of the 81 combinations ultimately tested in the anechoic condition. The purpose of this restriction was to reduce the training data size compared with the training set in which all the interferer directions are exhaustively covered. An additional benefit is that this reduces the ANN training time without excessively compromising ratio mask performance. It also allows us to reduce the size of the training data and to reduce the number of hidden neurons in the ANN from 50 down to 35.

We find that system V7-2 performs as well as the fully trained system, V7-1, for all interferer direction pairs seen during training (matched test cases) and where only one of the two interferer directions had been presented during training (semi-matched test cases). This result indicates that the system V7-2 has the ability to generalise and appropriately handle the partially unseen interferer direction combinations in semi-matched cases.

A third training strategy, TSV7-3, produced system V7-3. This system was trained using half the training data used in strategy TSV7-2. The performance gap between V7-1 and V7-3 shows, however, that this further reduction in training data leads to defective performance of V7-3 for unmatched test cases, i.e. where neither interferer direction in a training pair has been presented during training.

Estimators trained in the 0 dB SNR condition also have the ability to generalise to SNR conditions 5 dB above and below this level. The fully trained system ERMV7-1 improves intelligibility by approximately 11.8%, 18.6% and 24.3% for the 5 dB, 0 dB and -5 dB SNR conditions. The speech quality using the PESQ metric is also improved by more than 0.9 (18%) in all the tested cases.

Due to the advantage of the training strategy TSV7-2 when applied in anechoic conditions, we further retrained system V7-2 using training strategy TSVR7-2, which adds reverberation to the binaural mixture.

The relative importance of each type of cue was again measured, this time in these reverberant conditions. Results show that the importance of interaural coherence increases for frequencies above 2.2 kHz compared with anechoic conditions. This provides confirmation that interaural coherence contributes to the ratio mask estimation process and also indicates the range of the spectrum over which it chiefly operates. Moreover, the introduction of reverberation leads to a greater reduction in the relative importance of IPD compared with ILD. The mechanisms behind this finding are worthy of further investigation.

The results in this chapter indicate that there is potential for successful source segregation by these ANN-based mask estimators in reverberation. The results show that system VR7-2 improves the STOI score for a binaural input mixture by 11.5%, 17.5% and 20.1% for 5 dB, 0 dB and -5 dB SNR conditions, respectively. Compared with similar scenarios in anechoic conditions, the improvement is smaller due to the additional segregation challenges caused by reverberation. Reverberation also reduces mask performance in terms of speech quality. System VR7-2 improves the PESQ score compared with the input mixture by 0.46 (9%), 0.5 (10%) and 0.55 (11%) at 5 dB, 0 dB and -5 dB condition, respectively.

In general, we have demonstrated that the optimal cue mapping (OCM) approach in binaural signal processing can reveal and allow us to probe the relative importance of different cues when tackling the general problem of speech segregation. We have shown that it improves the intelligibility and quality of a target speech source in specific simulated acoustic conditions. In the next chapter we investigate the performance of the OCM approach further. In particular, we directly compare its ability to estimate binary and ratio masks with a leading alternative method based on state-of-the-art deep neural networks.

Chapter 8

A Comparative Evaluation of OCM

In this chapter, the performance of the optimal cue mapping (OCM) algorithm is compared with one of the latest representative neural network-based binaural segregation methods.

In December 2014 Jiang and Wang described a deep neural network (DNN)-based classification algorithm which performs binaural segregation in multi-source anechoic and reverberant environments (Jiang et al., 2014). The system employs a set of DNNs as a binary classifier. Each DNN corresponds to one particular frequency band. Binaural features such as ITD and ILD, and monaurally extracted gammatone frequency cepstral coefficients (GFCCs), are used to train the DNNs. The trained DNNs yield the binary decision for switching on or off each T-F unit in a binary mask for extracting the target speech from a binaural mixture. For ease of reference, we denote the DNN classification approach in Jiang et al. (2014) as the CLAS algorithm.

Jiang et al. (2014) compared their algorithm with the following four binaural separation algorithms:

TARG algorithm

Roman et al. (2006) proposed a binaural segregation method in multi- source reverberant environments by utilising the location information of the target source only. The first step is to perform target cancellation through adaptive filtering. Based on the observation that there is a correlation between the amount of cancellation and the relative strength of target to mixture, a binary decision is made to estimate the ideal binary mask (IBM). For convenience, we label this target-based approach as the TARG algorithm.

DUET algorithm

DUET (the Degenerate Unmixing Estimation Technique) is a blind source separation method which can, in principle, separate any number of sources using only two microphones (Rickard, 2007). It assumes that the signals in the time-frequency domain are sparse and exhibit W-disjoint orthogonality. Each source is segregated using a binary mask.

STATE algorithm

Woodruff (2013) proposed a binaural segregation approach which is based on pitch and azimuth cues. He formulates the IBM estimation as a multi-source state space searching across time. Each multi-source state encodes the number of active sources and the azimuth and the pitch of each active source. A set of multilayer perceptrons are trained to assign the time-frequency units to one active source in each multi-source state. A hidden Markov model framework is used to estimate the most probable path through the state space. Then segregation is achieved with an azimuth-based sequential organisation stage. For ease of reference, we label this state-space approach as the STATE algorithm.

MESSL algorithm

Mandel et al. (2010a) describe the Model-based Expectation Maximisation Source Separation and Localisation (MESSL) method for separating multiple sound sources by clustering for source localisation. Based on the interaural phase and level difference of each time-frequency unit, a probabilistic mask is estimated by computing the maximum-likelihood parameters from the expectation maximisation algorithm. Note: Jiang et al. (2014) quantise the MESSL output into a binary mask with a threshold of 0.5 in their comparison.

Even in low SNR test conditions and strong reverberation, the results obtained by Jiang et al. indicate that the joint binaural and monaural features they employed enable their DNN-based segregation algorithm to outperform the four representative binaural separation algorithms which have been summarised above. In addition, theirs was the first approach to apply DNNs for binaural segregation. The DNN-based algorithm is similar to our OCM method in that both are based on supervised machine learning techniques to estimate the T-F mask for a target source in adverse acoustic environments.

Since Jiang et al. (2014) have shown their DNN-based approach to be superior to several other competing segregation methods it is of great interest to compare it to our approach. In this chapter we first describe and validate our implementation of the DNN-based method. We then go on to compare the performance of the two algorithms in a variety of simulated, increasingly challenging acoustic environments.

8.1 DNN model verification

Before comparing our OCM approach with the CLAS algorithm, we first describe the CLAS algorithm and then verify that our implementation of it produces results in line with those reported by Jiang et al. (2014). Because the CLAS algorithm generally outperforms the TARG, DUET, STATE and MESSL algorithms, we compare the performance of our OCM algorithm with that of the CLAS algorithm and we do not present results for the other algorithms.

8.1.1 Description of the CLAS algorithm

The binaural segregation problem is considered by Jiang et al. (2014) as a classification task in their DNN-based approach. For signal decomposition, they firstly use a 64-channel fourth-order order gammatone filterbank for auditory peripheral processing. The filtered signal is then half-wave rectified to simulate the auditory nerve. Finally, each channel is divided into 20 ms frames with 50% overlap. The resulting time-frequency (T-F) representation of the signal is known as a cochleagram. Based on left and right channel T-F units, they extract the ILD in each 20 ms frame. They further break down each 20 ms frame into two 10 ms time durations. In this way they obtain extra two-dimensional (2D) ILD features. For the estimation of ITD, they calculate a 32-coefficient normalised cross-correlation function (CCF). The ITD is derived from this as the lag corresponding to the maximum in the CCF. For comparison reasons, they additionally employ all 32 CCF coefficients to determine the time delay between the two ears. Besides utilising binaural features, Jiang et al. (2014) also extract monaural features. Specifically, they calculate 36 gammatone frequency cepstral coefficients (GFCCs). One DNN is trained for each frequency subband using all of the above features, resulting in 64 DNNs in total. Each subband DNN consists of two hidden layers with 200 hidden neurons in each layer.

Jiang et al. (2014) use the HIT-FA evaluation criterion to assess the classification accuracy of their CLAS algorithm. HIT-FA is the difference between the percent of correctly labelled target-dominant T-F units (HIT rate) and the percent of incorrectly estimated interference-dominant T-F units (FA or false-alarm rate). The HIT-FA scores for the CLAS algorithm are compared under three different binaural feature conditions and within each of these for three different reverberation times. The results for the CLAS algorithm are shown in figure 8.1. The figure shows that the performance of the system with 34D (32 CCF coefficients and 2 ILDs) features yields the best HIT-FA score compared to the other two systems, although the performance of all three systems decreases as the reverberation time increases. The 32D CCFs feature provides more detailed information about relative time delays between the left and the right channels compared with only one ITD and improves the HIT-FA score. Furthermore, using the two-ILD vector results in slightly better scores than using only one such feature. Therefore, the binaural features they adopted for their subsequent studies included 32 CCFs and 2 ILDs, giving a feature vector comprising 34 inputs.

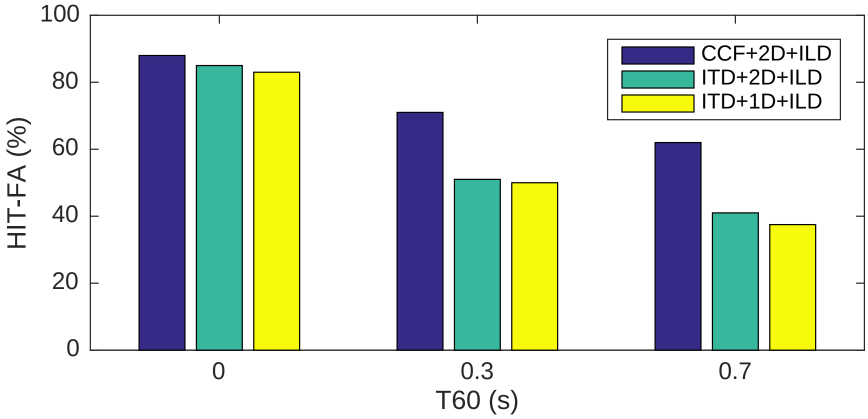


Figure 8.1: HIT-FA scores for the two-source setup used by Jiang et al. (2014) for the trained azimuths using a 0 dB SNR input mixture with target and interference located at 0° and −45° azimuth, respectively. Figure is redrawn from Jiang et al. (2014).

Jiang et al. (2014) next integrated monaural feature GFCCs in an attempt to improve the classification performance further. The combination of 34D binaural feature input data with 36D GFCC feature input data results in a 70D feature vector input for the DNNs. Figure 8.2 shows the comparisons between DNNs with and without GFCC inputs in anechoic conditions, where the interference azimuth is in the range of −180° to 180° in 10° steps. They achieved 1% better performance by using GFCC features at most azimuths.

However, the performance improved by 10% with the help of the GFCC feature data at 0° and 180° , which are in front and behind the head.

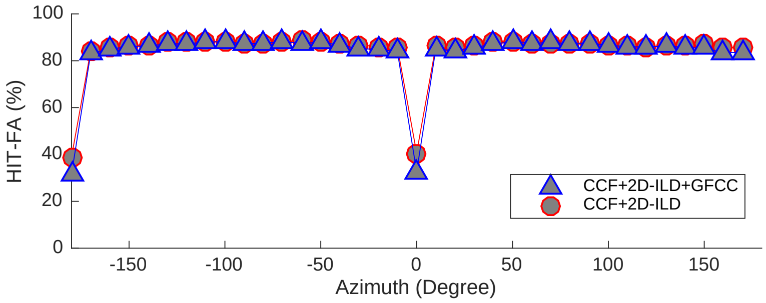


Figure 8.2: HIT-FA scores for two-source anechoic setup at 0 dB SNR with target at 0° and interference located at all azimuths in 10° steps. Figure redrawn from Jiang et al. (2014).

Because of the superior performance of their system with the 70D feature set, we reproduced this DNN-based binaural segregation algorithm with the demonstration code provided by the authors. In order to implement it accurately, we train the DNNs using exactly the procedure described by Jiang et al. (2014), and also confirmed the implementation with the first author.

8.1.2 Initial validation of our CLAS algorithm implementation

In this section we describe two small validation experiments which we implement to confirm that our replication of the DNN approach by Jiang et al. (2014) produces results in line with their reported findings. The system is tested with 50 sentences randomly picked from the TIMIT database and spatialised at each azimuth angle. It is important to note that the exact training and test data used by Jiang et al. (2014) is not available, nor are there sufficiently detailed records of the data they used to be able to recreate it exactly. Although our training and test data are drawn from the same database, the training and test sentence lists will not be the same, leading to minor differences in the detailed properties of our training and test data

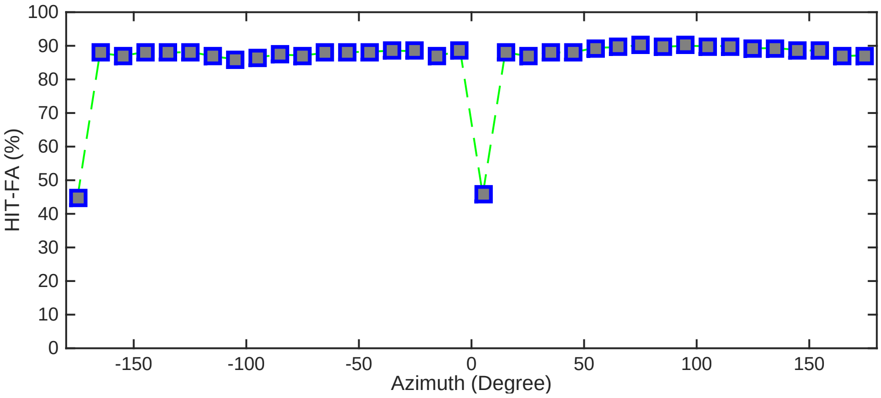


Figure 8.3: Verification of reproduced DNN-based algorithm for the two-source anechoic setup using the 70-feature input at 0 dB SNR. The target is at 0° azimuth and the interferer is stepped in 10° intervals around the virtual listener.

compared with the data created and used by Jiang et al. (2014). Therefore, small differences can be expected between the results originally obtained by Jiang et al. (2014) and the results we obtain from our replication of their CLAS algorithm.

Firstly, we evaluate the performance of the replicated CLAS algorithm for a two-source setup in anechoic conditions. The results are presented in figure 8.3. The azimuth angle $-180°$ and 0° in our axis system is equivalent to 0° and 180° in their axis system. At these two azimuth angles, the HIT-FA score is above 40%. For the other azimuth angles, the performance always exceeds 80% and is generally very close to 90%. Compared to the results shown in figure 8.2, the reproduced DNN algorithm achieves a similar performance.

We also validate the performance of the replicated CLAS system in reverberant conditions, using the same binaural room impulse responses as Jiang et al. (2014), specifically the ROOMSIM package by Campbell et al. (2005). The reverberation time constant (T60) is 300 ms. During training, we place the target at an azimuth of 0° and vary the direction of the interference over the full 360° range around the virtual listener (between $-180°$ and 180°), in 10° intervals. The system is tested with interference spaced at

5° intervals over the full 360° range. Therefore, the system is tested using interference both in trained azimuth directions and in untrained directions. At each azimuth angle, 50 test sentences are used.

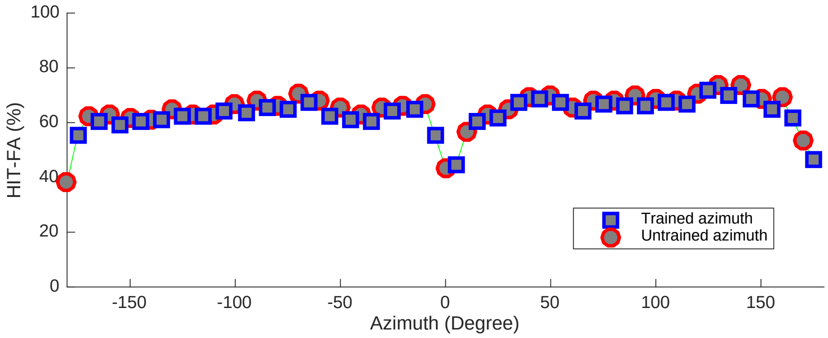


Figure 8.4: Performance of two-source reverberation setup at trained and untrained azimuths with reverberation and using input mixtures with 0 dB SNR. Figure redrawn from Jiang et al. (2014).

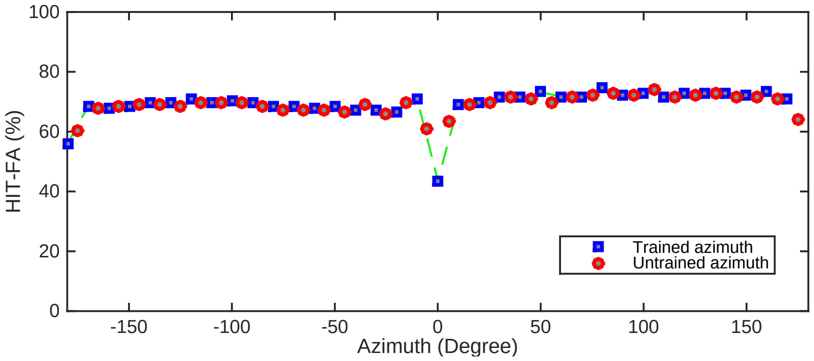


Figure 8.5: Verification of reproduced DNN-based algorithm for the two-source reverberation setup at trained and untrained azimuths with reverberation and using input mixtures with 0 dB SNR.

The original results reported in Jiang et al. (2014) are shown in figure 8.4 and the replicated results are shown in figure 8.5. Again, the azimuth angles presented in figure 8.4 are defined differently from those used in figure 8.5. For example, azimuth angle 180° used by Jiang et al. (2014) in figure 8.4 lies directly in front of the listener, which is 0° in our coordinate system in figure 8.5. By comparing the two figures, it can be seen that the replicated DNNs output HIT-FA scores of approximately 70% for most interference directions are approximately 3-5% higher than the reported results in Jiang et al. (2014). When the interference lies directly in front of the head (i.e., in the same direction as the target speech), both implementations produce approximately the same HIT-FA scores of around 45%. As mentioned above, precise details of the training and test data used by Jiang et al. (2014) were unavailable and we therefore emulated the training and test data as closely as possible from the information which was available, such as using the same BRIRs and the TIMIT database. It is likely that small differences between the properties of these two sets of data are the cause of the small but consistent performance differences observed between figures 8.2 and 8.3, and also between figures 8.4 and 8.5 in terms of their HIT-FA scores.

Having confirmed visually and by using HIT-FA scores that our implementation of the algorithm by Jiang et al. (2014) closely matches their reported results, we continue with a more thorough comparison of the CLAS and OCM approaches under a variety of simulated acoustic conditions.

8.2 Universally applied configuration details

In order to achieve a fair comparison of the CLAS and OCM methods, we apply the same configurations to both training and testing setups. For convenience, we use the configurations which have been previously reported by Jiang et al. (2014).

In all the training processes described in this chapter, only multiples of 10° in azimuth are used. The sources of all the audio material used here are the same as described in Jiang et al. (2014). Specifically, the training and test sentences are randomly selected from the training and test set of the TIMIT corpus (John et al., 1993). The interference is speech babble obtained from the NOISEX corpus (Varga and Steeneken, 1993). The corpus is of about 4 minutes' duration and is divided into two parts. The first part, of duration 106 s, is used in the training phase, and the other part, of duration 128 s, is used in the testing phase. The data were provided by Jiang et al. (2014). For each configuration, we generate 500 different binaural mixtures to train the DNN algorithm, which is the same number used by Jiang et al. (2014). All the training mixtures have 0 dB global SNR, as measured in the left channel. Each test case consists of 50 test sentences.

As explained in section 8.1.1, Jiang et al. (2014) use a feature space comprising 70 inputs for training the DNNs. The inputs are composed of 32 normalised cross-correlation function coefficients (CCF) for each pair of T-F units, two ILD features which are measured every 10 ms in 20 ms duration, and 36 monaural GFCC features. They found that this combination achieved the best performance. Therefore, we inherit this setup to train the replicated DNNs.

Each subband DNN consists of an input layer with 70 neurons, two hidden layers with 200 binary neurons in each layer and an output layer with only one neuron with a binary label. Hence, the DNN is used to estimate the ideal binary mask. DNN training follows exactly the same process described in Jiang et al. (2014).

In the evaluation phase, we randomly choose 50 sentences from the TIMIT database to compare the two systems for each test case. For the comparison we generate STOI and PESQ scores. We adopt the clean-target-modulated SER metric to measure the similarity between the estimated target signal and the original clean target signal. The SER metric is defined as:

$$SER = 10 \cdot \log 10 \frac{\sum\limits_{t} s_I(t)^2}{\sum\limits_{t} (s_I(t) - s_E(t))^2} \tag{8.1}$$

where s_I and s_E denote the ideal clean target signal and the signal resynthesized from the estimated binary mask or estimated ratio mask, respectively.

Note, in order to be consistent with the test conditions in Jiang et al. (2014), all the test SNRs in this chapter are measured in the left channel

Table 8.1: System definitions and training details for each group of comparisons. Part A.

Interferer configuration	BRIR	Algorithm	System	Mask set
			Ideal	IBM IRM
2-source one interferer 10° step	BRIR-A Anechoic	OCM	O2-A	EBMO2-A ERMO2-A
		CLAS	C2-A	EBMC2-A
3-source one interferer on either side 10° step	BRIR-A Anechoic	OCM	O3-A	EBMO3-A ERMO3-A
		CLAS	C3-A	EBMC3-A
5-source one interferer in each quadrant 10° step	BRIR-A Anechoic	OCM	O5-A	EBMO5-A ERMO5-A
		CLAS	C5-A	EBMC5-A
2-source one interferer 10° step	BRIR S300 simulated	OCM	O2-S300	EBMO2-S300 ERMO2-S300
		CLAS	C2-S300	EBMC2-S300
2-source one interferer 10° step	BRIR S700 simulated	OCM	O2-S700	EBMO2-S700 ERMO2-S700
		CLAS	C2-S700	EBMC2-S700
3-source one interferer on either side 10° step	BRIR S300 simulated	OCM	O3-S300	EBMO3-S300 ERMO3-S300
		CLAS	C3-S300	EBMC3-S300
3-source one interferer on either side 10° step	BRIR S700 simulated	OCM	O3-S700	EBMO3-S700 ERMO3-S700
		CLAS	C3-S700	EBMC3-S700
5-source one interferer in each quadrant 10° step	BRIR S300 simulated	OCM	O5-S300	EBMO5-S300 ERMO5-S300
		CLAS	C5-S300	EBMC5-S300
5-source one interferer in each quadrant 10° step	BRIR S700 simulated	OCM	O5-S700	EBMO5-S700 ERMO5-S700
		CLAS	C5-S700	EBMC5-S700

Table 8.2: System definitions and training details for each group of comparisons. Part B.

Interferer configuration	BRIR	Algorithm	System	Mask set
2-source one interferer	BRIR R320	OCM	O2-R320	EBMO2-R320 ERMO2-R320
10° step	Room A	CLAS	C2-R320	EBMC2-R320
2-source one interferer	BRIR R470	OCM	O2-R470	EBMO2-R470 ERMO2-R470
10° step	Room B	CLAS	C2-R470	EBMC2-R470
2-source one interferer	BRIR R680	OCM	O2-R680	EBMO2-R680 ERMO2-R680
10° step	Room C	CLAS	C2-R680	EBMC2-R680
2-source one interferer	BRIR R890	OCM	O2-R890	EBMO2-R890 ERMO2-R890
10° step	Room D	CLAS	C2-R890	EBMC2-R890

after spatialisation. The rigorous comparison of the OCM and CLAS systems involves 13 training configurations and generating 41 masks. These are summarised in tables 8.1 and 8.2. The labels in the table will be referred to frequently in subsequent sections.

8.3 Comparison in anechoic configuration

The anechoic BRIR, BRIR-A, is employed as a baseline in this evaluation. In the training process for all simulated BRIRs, two-source configurations, the target talker is located at 0° and the location of the interference is systematically varied in 36 steps from $-180°$ to 170° in 10° steps. Training follows the process described in section 8.2 to produce the two-source, anechoic, OCM and CLAS systems, O2-A and C2-A, respectively.

8.3.1 Preliminary comparison for one test condition

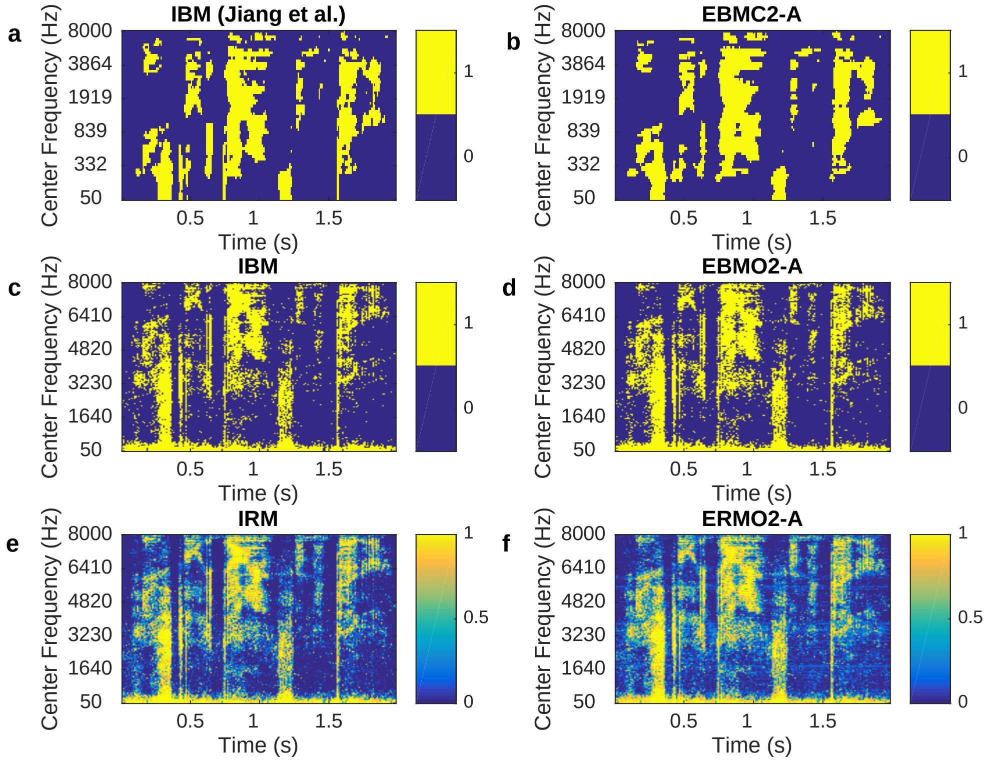


Figure 8.6: Ideal and estimated masks for systems O2-A and C2-A in the two-source configuration using input mixtures at 0 dB SNR with target speech at 0° and interferer at −45° azimuth.

With the target speech always located at an azimuth of 0° in the horizontal plane, we begin by analysing the performance of O2-A and C2-A for a single speech interferer at an azimuth of −45°. When comparing the two systems, it is important to keep in mind that the CLAS method creates a binary mask estimate only (EBMC2-A), whereas the OCM method produces both a binary and a ratio mask estimate (EBMO2-A and ERMO2-A, respectively).

As an informal visual indication of the relative performance of the two systems, figure 8.6 compares the ideal binary mask, IBM2 with the estimated ideal binary mask, EBMC2-A for system C2-A and the estimated binary and

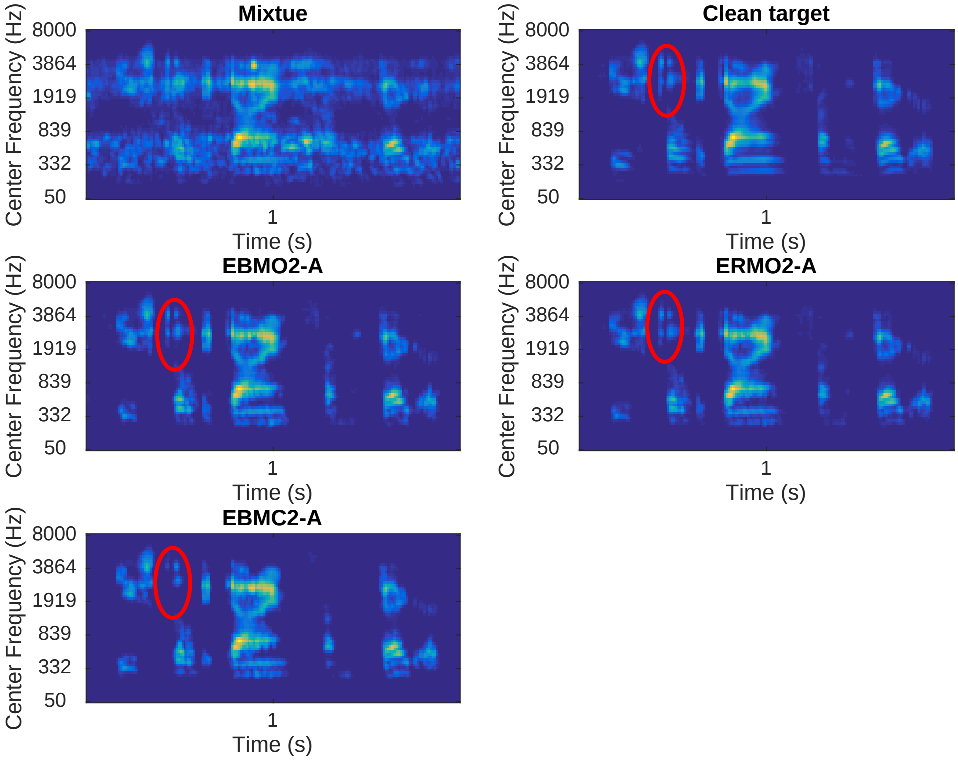


Figure 8.7: Cochleagrams for mixture and the segregated signals for two-source configuration at 0 dB SNR with target speech at 0° and interferer at $-45°$.

ratio masks, EBMO2-A and ERMO2-A, respectively, for system O2-A. The figure shows the masks for one of the 50 test sentences.

Note that because the single speech interferer is at $-45°$ azimuth, this is an unmatched test case, i.e. this interferer angle has not been seen by the systems during their training phases. The visual impression from figure 8.6 is that all the estimated masks closely match their ideal mask counterparts.

More spectral detail is observable in the masks relating to system O2-A due to the finer frequency resolution afforded by the short-time Fourier transform approach compared with the simulated auditory filterbank used in system C2-A. The greater spectral detail, however, is unlikely to be perceptually significant since, by definition, the auditory filterbank bandwidths are designed to match the spectral resolution of the human hearing system.

Figure 8.7 shows cochleagrams for the original speech, the target and interferer mixture, the clean target speech and the segregated target speech. The cochleagrams of the estimated target speech are very similar for all three estimates. However, an example of an observable difference has been circled in red. In figures 8.7 (c) and (d), showing the results for the OCM method, the detail inside the red circle matches the target source closely, whereas some features are partially or completely missing in figure 8.7 (e) for the CLAS system. Again, this is likely to be due to the use of an auditory filterbank model in this system. The short-time Fourier transform approach used here has a finer frequency resolution than the auditory filterbank model and so preserves more spectral detail.

Figure 8.8 (a), (c) and (e) show waveforms of the segregated target speech in the time domain and the output clean-target-modulated SER for each of the three estimated masks, EBMC2-A, EBMO2-A and ERMO2-A, respectively. Although all the estimated waveforms appear very similar after correction of the processing delay, the residual errors for the OCM method (figures 8.8 (d) and (f)) are slightly superior to that of the CLAS method (figure 8.8 (b)) in terms of their SERs. The error signal level using mask EBMO2-A is 0.19 dB smaller than that produced by mask EBMC2-A using the approach by Jiang et al. (2014). There is a further improvement of 1.39 dB in terms of SER using the OCM ratio mask ERMO2-A.

In this typical example, both the OCM and CLAS algorithms visually do a good job of estimating masks for segregating the target speech. There are only small differences between the ideal and estimated masks in each case. The clean-target-modulated SERs indicate that the optimal cue mapping binary mask performs slightly better than the CLAS system and that the OCM ratio mask performs better still. In the next section, the comparison between the two approaches is put on a more rigorous footing.

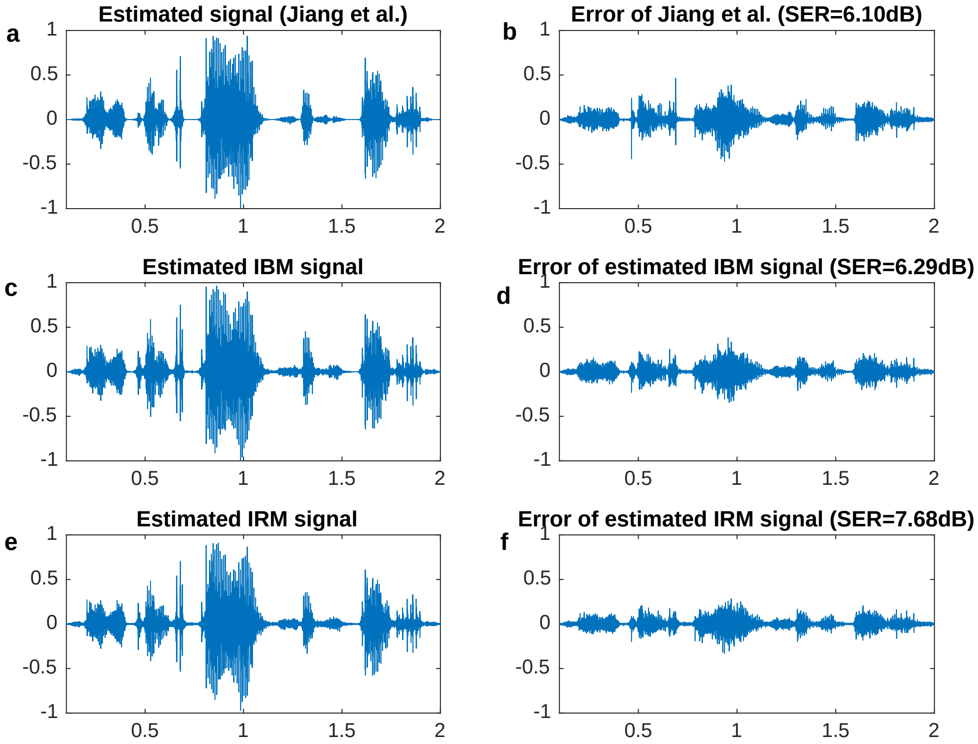


Figure 8.8: Segregated target waveforms for the two-source configuration at 0 dB SNR with interference at $-45°$ produced by masks EBMC2-A, EBMO2-A and ERMO2-A in (a), (c) and (e), respectively. The corresponding target waveform estimation errors are shown in (b), (d) and (f), respectively.

8.3.2 Detailed comparison of two-source configurations

We extend our analysis of systems O2-A and C2-A to include the full set of interferer directions and evaluate their STOI intelligibility scores, PESQ quality scores and signal-to-error ratios (SERs). We determine and compare the performance of the two systems to create a baseline performance in all subsequent test configurations. The evaluation can be split into two parts, the matched and unmatched cases. Training is carried out using a single interferer at multiples of 10° azimuth. Therefore, azimuth angles between

$-170°$ and $180°$ in $10°$ steps belong to the matched case and the azimuth angles between $-175°$ and $175°$ in $10°$ steps are classified as the unmatched case. This means that only half of the 72 interferer locations encircling the virtual listener are used for training.

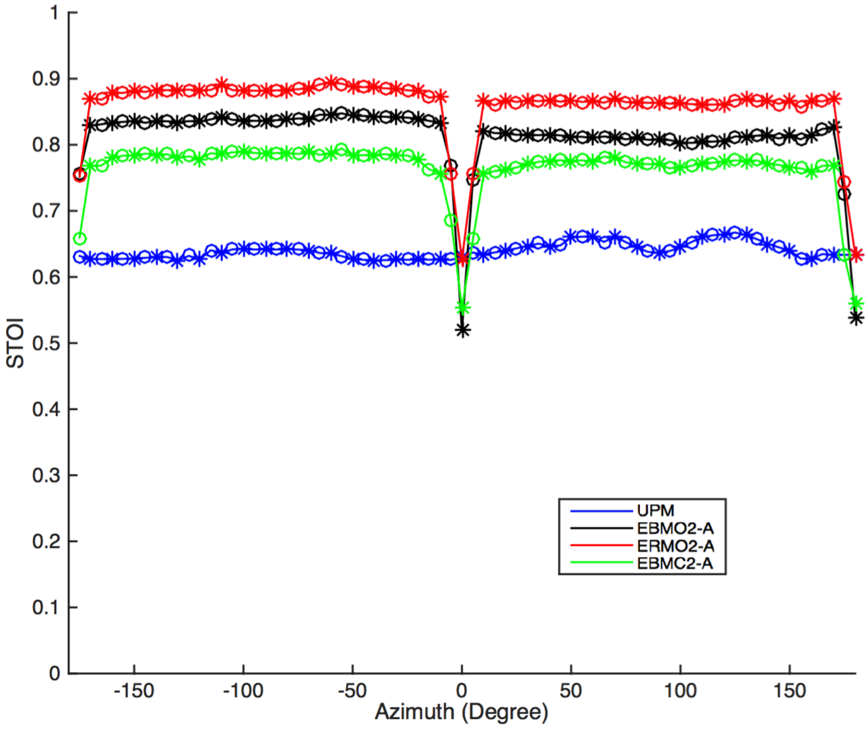


Figure 8.9: STOI scores for anechoic, two-source systems O2-A and C2-A for interference locations in 5 degree steps using 0 dB input mixture SNR. ∘ and $*$ denote scores for the matched and unmatched case, respectively.

Figures 8.9 and 8.10 show the STOI and PESQ scores in both the matched and the unmatched cases for all 72 directions in anechoic conditions. In figure 8.9, the OCM system O2-A, producing the ratio mask ERMO2-A, performs best. The STOI scores using the ratio mask system are above 0.85 for most interference azimuths and they are approximately 5% better than those produced by the OCM estimated binary mask EBMO2-A. Furthermore, compared with the CLAS-estimated binary mask, EBMC2-A, the OCM ratio mask yields approximately 10% further improvement. Compared with the

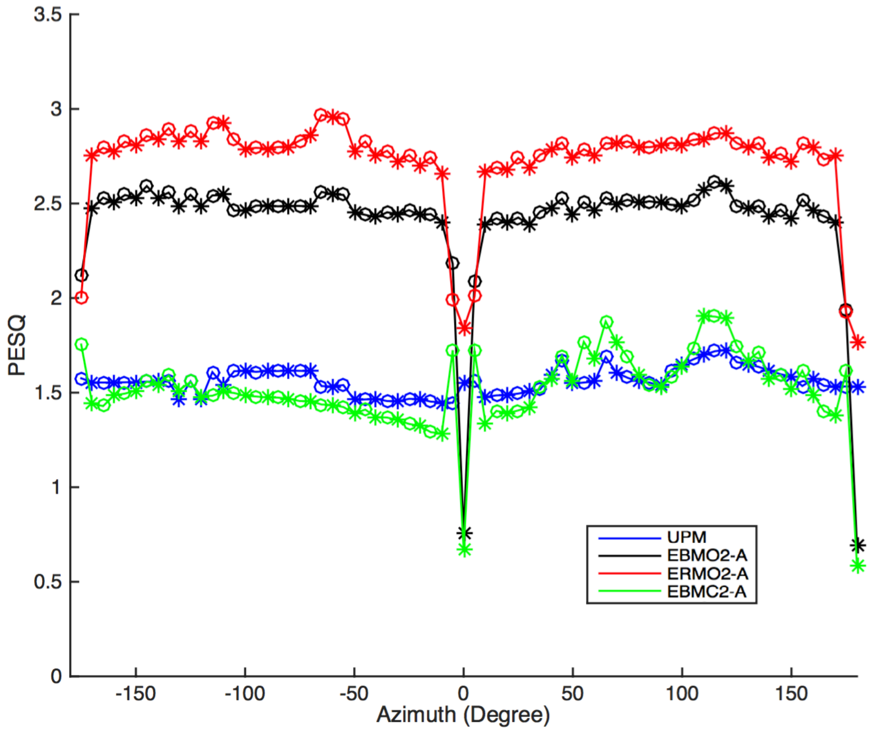


Figure 8.10: PESQ scores for anechoic, two-source systems O2-A and C2-A for interference locations in 5 degree steps at 0 dB SNR mixture levels. ∘ and ∗ denote scores for the matched and unmatched case, respectively.

STOI score for the unprocessed mixture (UPM), applying the OCM ratio mask, ERMO2-A, provides an improvement of almost 25%, from 0.62 up to 0.87 for most interferer directions. However, the scores obtained using both binary masks (OCM and CLAS) fall below the original mixture score when the interference is located in front or behind the listener at 0° or −180°, whereas the ratio mask score is similar to that of the original. The matched cases yield similar STOI scores to their adjacent unmatched cases.

The relative scores for interference located to the left side of the virtual listener are predominantly mirrored on the right. The small asymmetry is because the test SNRs are measured in the left channel, causing the SNRs to rise above the nominally defined level of 0 dB when the interferer is placed on the left side and below 0 dB when the interferer is placed on the right side.

Figure 8.10 shows the PESQ scores for both systems and the original mixture. Again, the ratio mask, ERMO2-A, performs best out of the three estimated masks, with its PESQ scores above 2.7 for most interference locations. The scores for the OCM estimated binary masks, EBMO2-A, drop to approximately 2.5. Furthermore, the CLAS binary mask, EBMC2-A, yields scores close to those of the original mixture, which is about 1.5. Overall, the PESQ scores suggest that the OCM estimated binary and ratio masks both generally improve target speech quality.

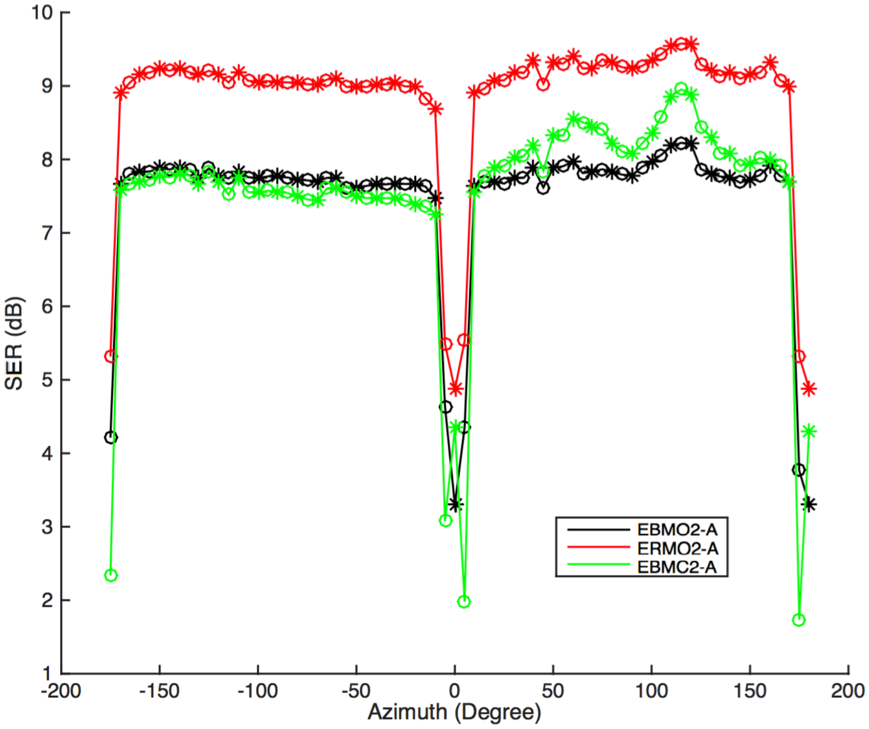


Figure 8.11: Output clean-target-modulated SER performance for anechoic, two-source systems O2-A and C2-A for interference locations in 5 degree steps at 0 dB SNR input mixture levels in anechoic condition. ∘ and ∗ denote scores for the matched and unmatched case, respectively.

To assess the numerical similarity of the segregated target signal to its original clean form, figure 8.11 shows the clean-target modulated SER output results. The OCM ratio mask's superior performance is clearly demonstrated, with an SER close to 9 dB for most interferer directions, which is over 1 dB greater than the estimated binary mask performance of the OCM system. The CLAS systems's estimated binary mask, EBMC2-A, performs similarly for interferers on the left side, but up to 0.6 dB better than EBMO2-A on the right. These results lend support to the statement that improvements in SER do not necessarily correlate with either speech intelligibility or speech quality improvements, since the OCM binary masks outperform the CLAS binary masks in both these aspects.

8.3.3 Detailed comparison with multiple interferers

We next extend our evaluation to OCM and CLAS-based systems in three-source and five-source configurations. We again use the same interferer configurations and training procedures used by Jiang et al. (2014). O3-A and C3-A are the trained three-source OCM and CLAS systems, respectively. During training, one noise source is always randomly chosen on the left side of the virtual listener and the direction of the other source is randomly chosen on the right side. For the five-source configuration systems, O-A5 and C-A5, training proceeds with the four interferers in randomly selected directions, each within a different quadrant so that, once again, the entire horizontal plane is covered. In initial testing of the three-source configuration, two interferers are placed at $-45°$ and $45°$ azimuth and in the five-source test configuration, four interferers are located symmetrically within each quadrant at azimuths of $-135°$, $-45°$, $45°$ and $135°$, respectively.

The resulting test scores for STOI, PESQ and SER analyses are shown in figure 8.12. As expected, there is a visible decline in scores for the 5-source configuration compared with the 3-source one. However, the OCM ratio masks, ERMO3-A and ERMO5-A, produce the best results in all test conditions. The performance of the OCM estimated binary masks, EBMO3-A/5-A, ranks second, followed by the CLAS masks EBMC3-A/5-A. The STOI score for the three-source configuration OCM binary mask shows an improvement of 6% over the original mixture, from 0.63 up to 0.69. The OCM ratio mask, ERMO3-A, produces a 10% STOI improvement. There is a 1% reduction in STOI score using the CLAS binary mask, EBMC3-A. In the five-source configuration, the STOI improvement is 8%, 11% and 5% for the estimated masks EBMO5-A, ERMO5-A and EBMC5-A, respectively, with the OCM approach again showing greatest improvement.

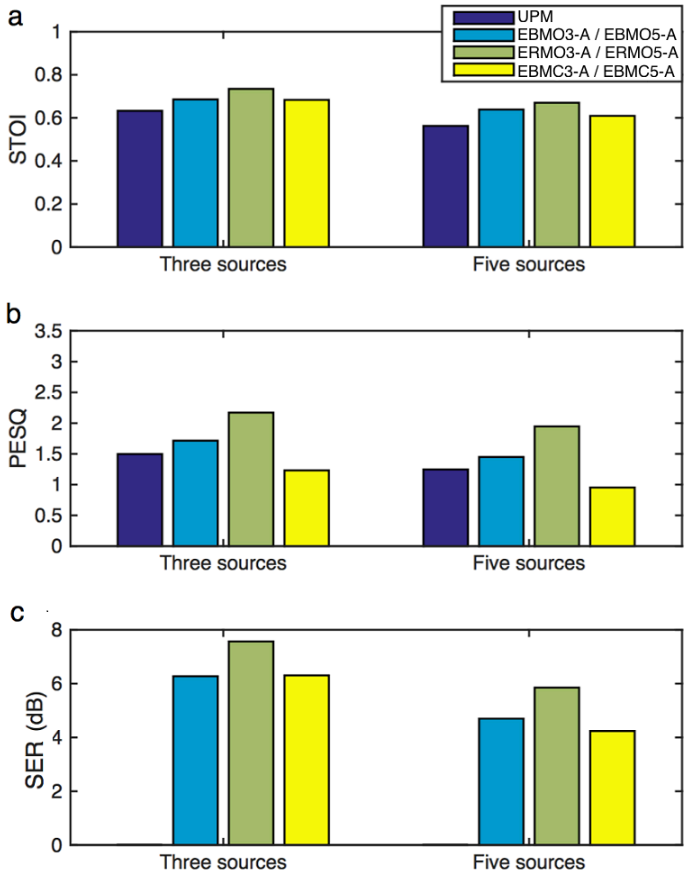


Figure 8.12: STOI, PESQ and output clean-target-modulated SER performance for anechoic three-source and five-source configurations for systems O3-A/O5-A and C3-A/C3-A for 0 dB SNR input mixture levels.

The resynthesised speech quality indicated by the PESQ score in figure 8.12 actually declines for the CLAS system. By contrast, the PESQ scores for the OCM estimated ratio masks, ERMO3-A/5-A, improve by 0.7 in both the three- and the five-source configurations, which is a rise of 14% compared with the score for the original mixture.

The clean-target-modulated SER indicates that the segregated target signal using the OCM ratio mask is closest to the original signal in terms of signal power. The two binary masks create similar SERs in the three-source configuration. In the five-source case, the OCM binary mask outperforms the CLAS binary mask by about 0.5 dB.

8.3.4 System performance for different SNRs

Systems O2-A and C2-A are trained using signal-to-noise ratios (SNRs) nominally set to 0 dB (see section 8.2). In this section we investigate how well the masks trained under these conditions perform for two different SNRs: 5 dB and -5 dB. Since the SNRs of the binaural mixture are measured in the left channel, figure 8.14 shows the results for the left channel only. The speech intelligibility of the target signal after source segregation is shown in figure 8.13 (a) and (b). In the 5 dB SNR test condition, the OCM ratio mask, ERMO2-A, trained at 0 dB SNR, improves the STOI score for the original mixture by approximately 20% for most interferer directions and performs best out of the three masks. The OCM binary mask, EBMO2-A, performs slightly (2%) better than the CLAS system binary mask, EBMC2-A. When the interferer is close to the target signal, at -5° azimuth, the improvement is limited and there is only an 8% increase in STOI score as a result of applying the OCM ratio mask. In the -5 dB input SNR condition, the differences in performance are magnified. There is approximately a 30%, 25% and 18% improvement in the STOI scores for the masks ERMO2-A, EBMO2-A and EBMC2-A, respectively. When the interferer lies at -5° azimuth, it is interesting to note that the OCM binary mask improves the STOI score by 20% and outperforms the score produced using the OCM ratio mask by 4%.

In terms of speech quality in the 5 dB SNR test condition, as shown in figure 8.13 (c), the OCM ratio and binary masks improve the PESQ score for the original mixture by approximately 1 (20%) and 0.8 (16%), respectively, except when the interferer is located at -5° azimuth. The CLAS binary mask,

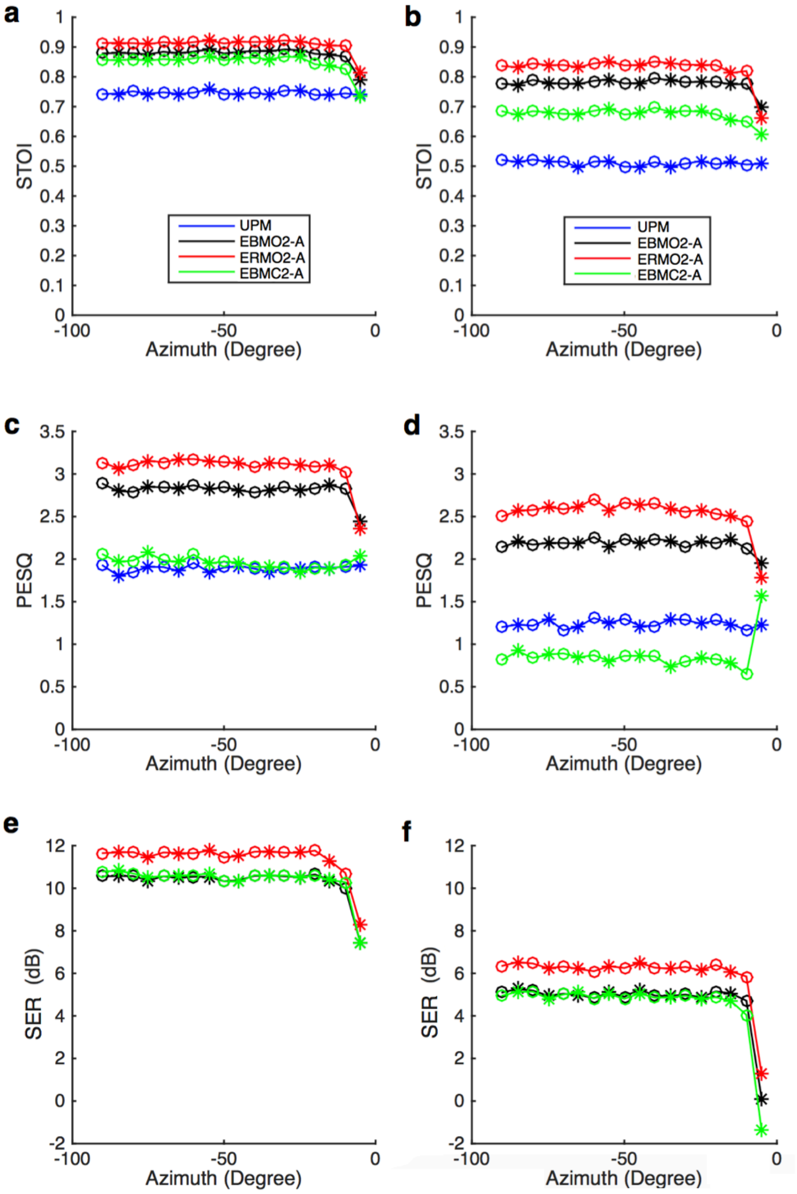


Figure 8.13: STOI, PESQ and output clean-target-modulated SER performance for the anechoic two-source configuration as a function of interference location for 5 dB and -5 dB SNR input mixtures. ∘ and ∗ denote matched case and unmatched case, respectively. (a), (c) and (e) are the results obtained using the 5 dB mixture. (b), (d) and (f) are the results obtained using the -5 dB mixture.

however, yields a similar PESQ score to that of the original mixture at this SNR. In the -5 dB SNR test condition (figure 8.13 (d)), the improvement in PESQ score is 1.3 (26%) and 1 (20%) for the OCM ratio and binary masks for most interferer directions. Performance of the CLAS binary mask is worse and it generally reduces speech quality compared with the PESQ score for the original mixture.

The SER comparison between the segregated target signals and the ideal clean target signals is shown in figure 8.13 (e) (the 5 dB SNR case) and (f) (the -5 dB SNR case). The OCM ratio mask, ERMO-A2, performs best for all interferer directions in both the 5 dB and -5 dB SNR conditions. The two binary masks produce similar clean-target-modulated SERs.

8.3.5 Performance of three- and five-source configurations under varying SNR

The three-source and five-source configurations described in section 8.3.3, are trained using 0 dB binaural mixtures. In this section, the masks trained using 0 dB SNR input mixtures are presented with input mixtures which have SNRs of 5 and -5 dB. The test results are shown in figure 8.14. It can be seen in figure 8.14 (a) that the set of OCM ratio masks, ERMO3-A, for the three-source configurations have the highest mean STOI scores compared with the two binary masks, EBMO3-A and EBMC3-A. In the 5 dB SNR test condition, the OCM binary mask, EBMO3-A, performs worse than the CLAS binary mask, EBMC3-A, in the three-source setup. The situation is more straightforward for the 5 dB SNR three- and five-source setups (figure 8.14 (b)), where the performance of the OCM ratio mask, ERMO5-A, produces the highest mean STOI for both the three- and five-source configurations and the OCM binary mask is in second place for both source configurations.

Figures 8.14 (c) and (d) show the mean PESQ scores for the same test conditions as the STOI analysis, above. The results for the OCM masks

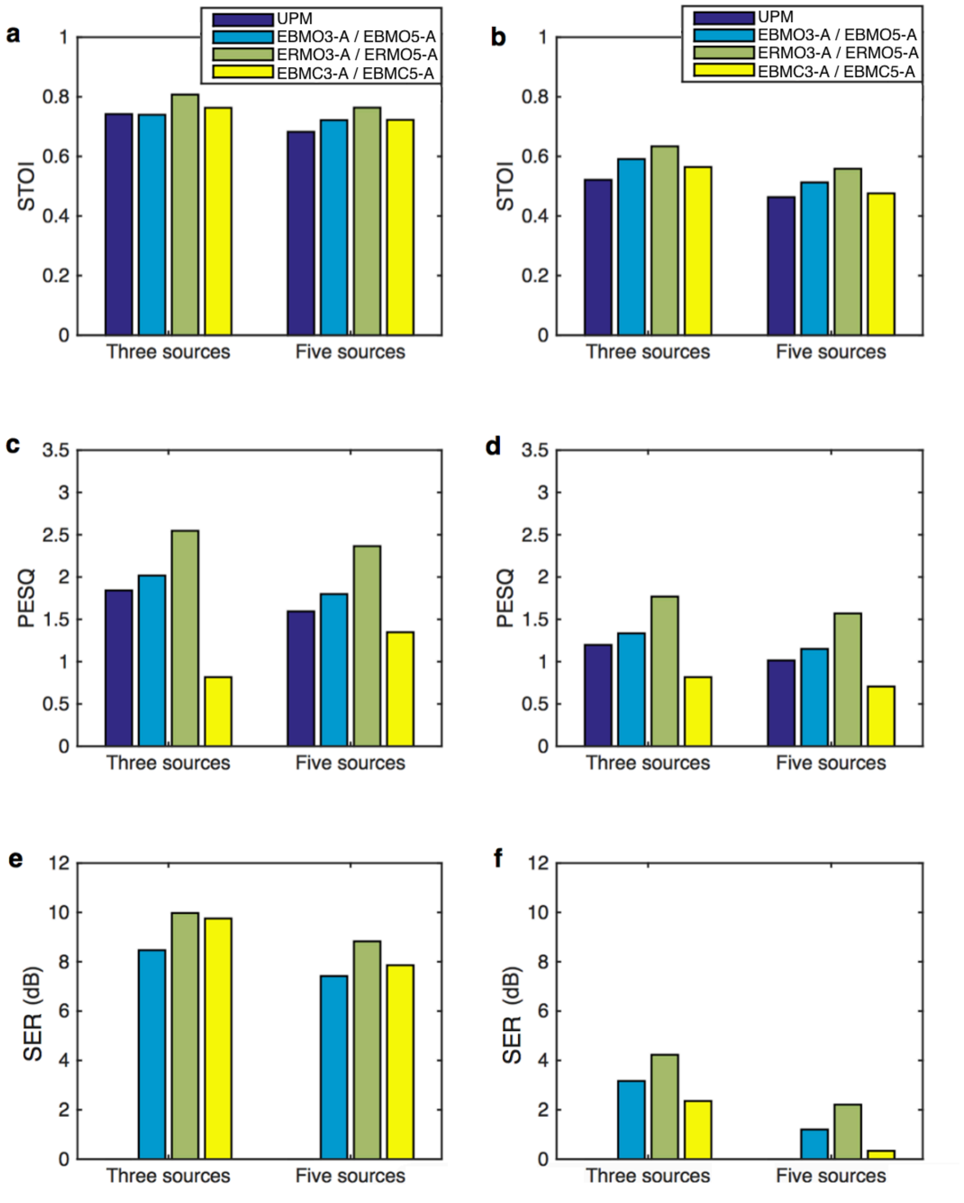


Figure 8.14: Mean STOI, PESQ and output clean-target-modulated SER performance for anechoic three-source and five-source configurations for different interferer directions at 5 dB and -5 dB input mixture SNRs. (a), (c) and (e) are the STOI, PESQ and SER results, respectively, using 5 dB input mixture SNR. (b), (d) and (f) are the corresponding results using -5 dB input mixture SNR.

display a familiar pattern, with the ratio masks, ERMO3-A and ERMO5-A producing the highest quality of segregated target speech. The OCM binary mask also shows an improved mean PESQ score compared with the score for the original binaural mixture. The CLAS binary mask produces segregated target speech with a lower speech quality than the original mixture in all test conditions.

The similarity between the segregated target speech and the clean target speech in terms of their SER mean values is shown for the 5 dB input mixture SNR test condition in figure 8.14 (e). Here, the ratio mask delivers the highest mean, followed by the CLAS binary mask for both the three- and five-source setups. For the -5 dB SER test condition, the OCM binary mask outperforms the CLAS binary mask, which is a reversal in the trend and which was in transition in the 0 dB SER case shown in figure 8.12 (c).

8.4 Comparison in reverberant configuration

So far in this chapter, we have compared our optimal cue mapping (OCM) algorithm with the DNN-based binary classification (CLAS) approach by Jiang et al. (2014) in simulated anechoic conditions. The anechoic comparisons involved two-source, three-source and five-source setups using 5 dB, 0 dB and -5 dB SNR binaural test mixtures. In this section, we compare the OCM and CLAS systems in increasingly realistic conditions which include reverberation. Conform to the analysis procedure presented in Jiang et al. (2014), we use both simulated binaural room impulse responses (BRIRs) and BRIRs measured in real rooms, using two, three and five sources and binaural input mixtures with SNRs of 5 dB, 0 dB and -5 dB.

8.4.1 OCM and CLAS system performance in simulated reverberation

We begin by analysing the target segregation performance of the OCM and CLAS systems using two sets of simulated BRIRs. The simulated BRIRs, referred to as BRIR-S, are generated using the ROOMSIM package described by Campbell et al. (2005). BRIR-S contains two sets of BRIRs, the reverberation time (T60) of BRIR-S300 is 300 ms and that of BRIR-S700 is 700 ms. The direct-to-reverberant ratio (DRR) of each BRIR is -1.97 dB and -7.74 dB respectively.

With reference to figure 8.15, the reflection coefficients of all the surfaces in the simulated room are uniform. The room is 6 meters long, 4 meters wide and 3 meters high. The virtual listener is fixed in location, 2.5 meters from two adjacent walls and 2 meters above floor level. 72 BRIR measurements are simulated in the horizontal plane with an azimuth resolution of 5°, which thus describe a full 360° revolution around the virtual listener. As shown in figure 8.15, all sound sources are located 1.5 meters away from the listener,

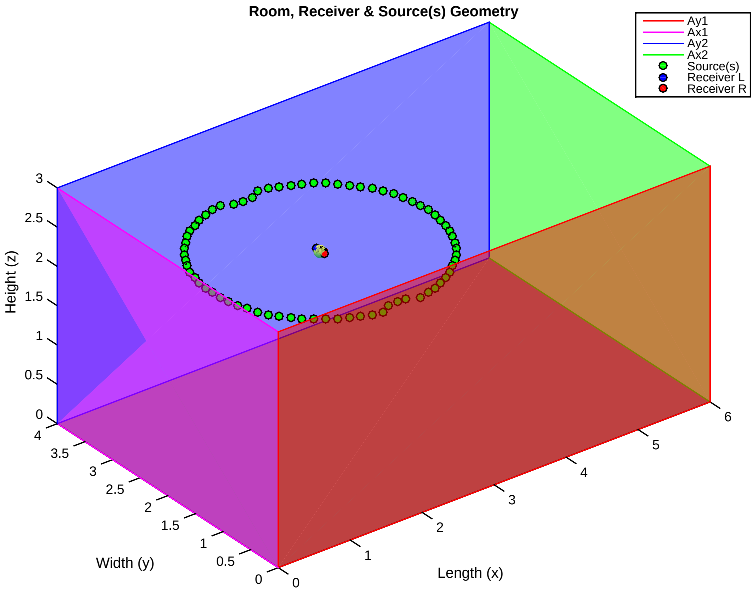


Figure 8.15: The geometrical arrangement of the virtual room used to simulate binaural room impulse response sets BRIR-S.

apart from the three sound sources at azimuth angles of −85°, −90° and −95° to the left of the listener and three more at 85°, 90° and 95° to the right. These six sources are constrained by the proximity of the walls and so they are located only 1.4 meters from the listener. This matches the setup used in Jiang et al. (2014). The virtual target talker is always located at 0° azimuth and during training the location of the interfering babble noise source is spatialised at 36 positions, from −180° to 170° at 10° intervals using the desired set of BRIRs.

As summarised in table 8.1, OCM system O2-S300 and CLAS system C2-S300 are trained using simulated BRIR set BRIR-S300 in a similar manner to that described in section 8.2. The SNR of the original mixture is set to 0 dB. Three masks are produced, EBMO2-S300, ERMO2-S300 and EBMC2-S300, which are the OCM binary mask and ratio mask and the CLAS binary mask, respectively. This training process is repeated using the BRIR set BRIR-S700.

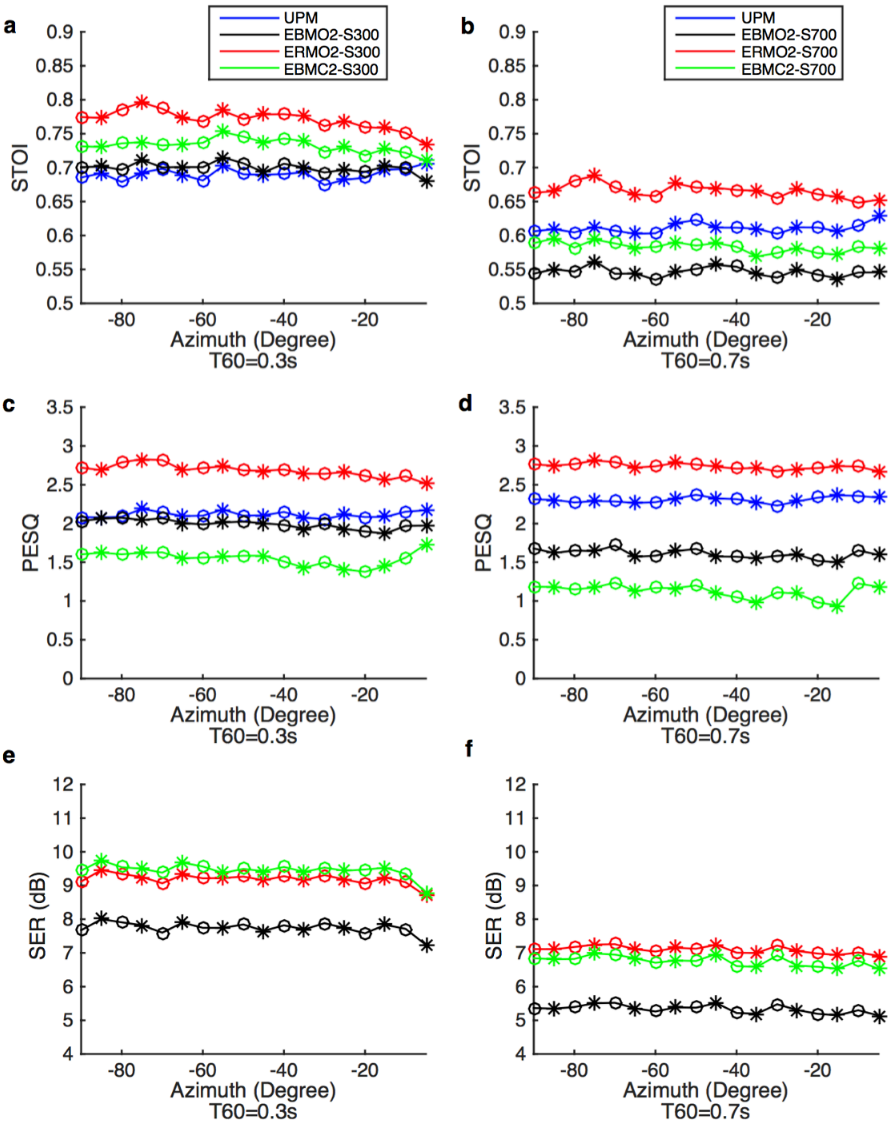


Figure 8.16: STOI and PESQ scores and output clean-target-modulated SER performance in the reverberant two-source configuration for the OCM and CLAS masks as a function of interference location. The input mixture SNR is 5 dB. Left column, T60 is 300 ms; right column, T60 is 700 ms. ○ and ∗ denote a matched case and an unmatched case, respectively.

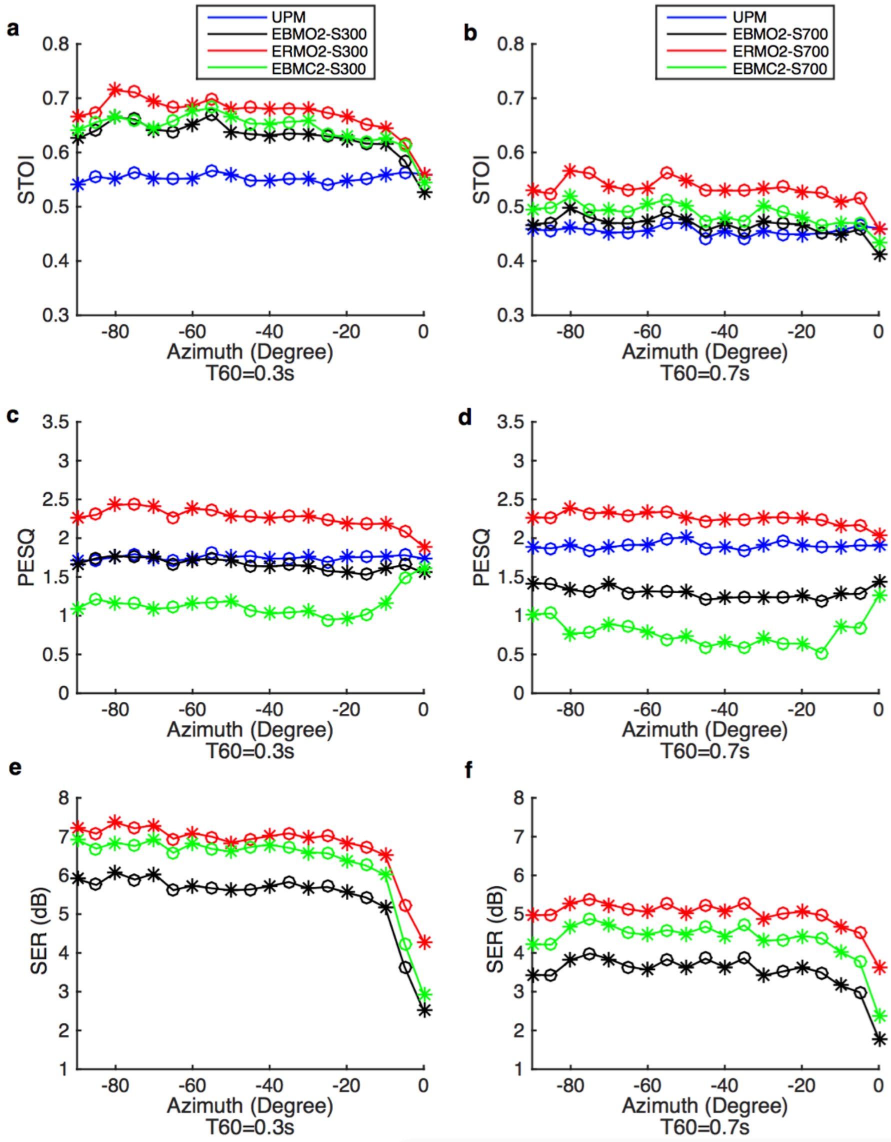


Figure 8.17: STOI and PESQ scores and output clean-target-modulated SER performance in the reverberant two-source configuration for the OCM and CLAS masks as a function of interference location. The input mixture SNR is 0 dB. Left column, T60 is 300 ms; right column, T60 is 700 ms. ∘ and ∗ denote a matched case and an unmatched case, respectively.

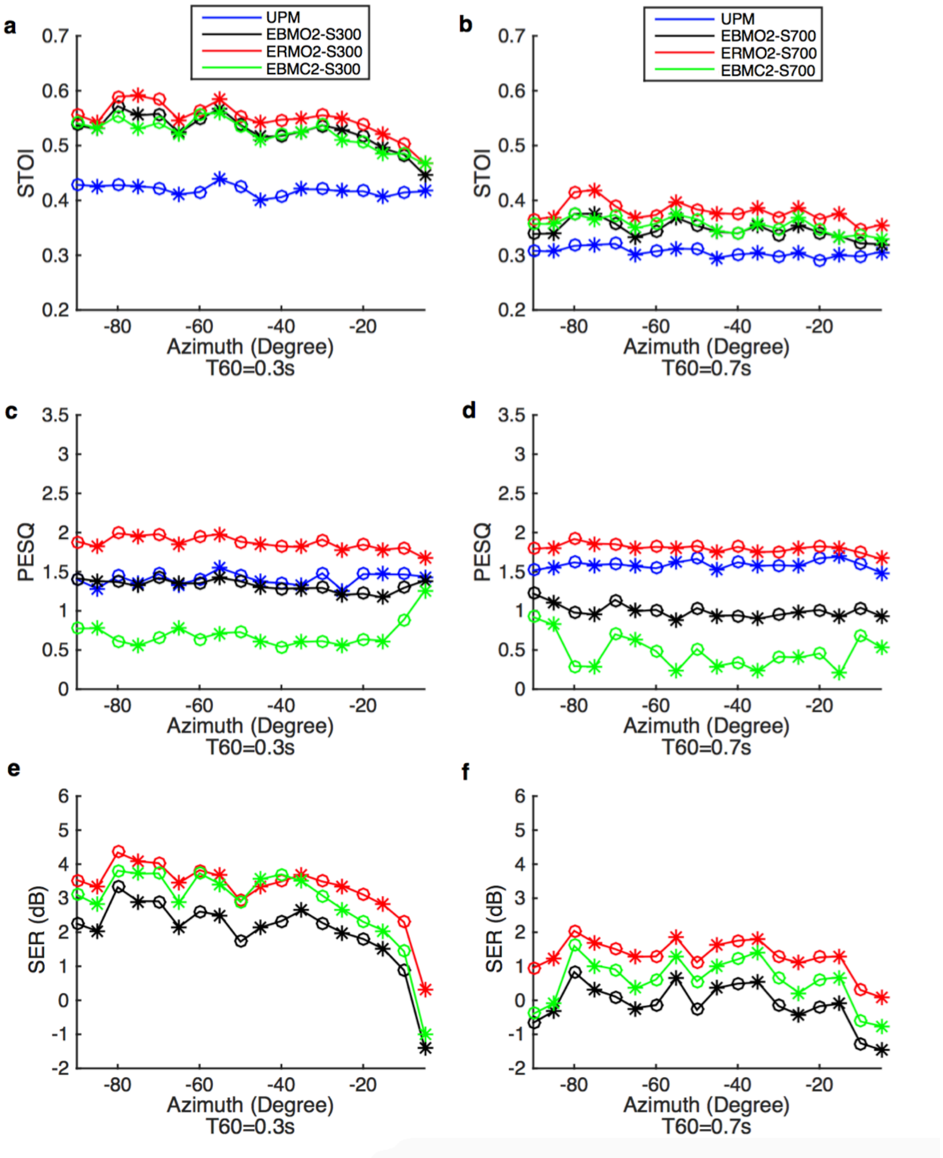


Figure 8.18: STOI and PESQ scores and output clean-target-modulated SER performance in the reverberant two-source configuration for the OCM and CLAS masks as a function of interference location. The input mixture SNR is -5 dB. Left column, T60 is 300 ms; right column, T60 is 700 ms. ○ and ∗ denote a matched case and an unmatched case, respectively.

During testing, the direction of the interferer is stepped from −90° to −5° in azimuth. The results of comparing this two-source configuration using 5 dB SNR input mixture are shown in figure 8.16. In addition, figures 8.17 and 8.18 show the corresponding results using input mixture SNRs of 0 dB and -5 dB for the same two T60s.

Figures 8.16 (a), 8.17 (a) and 8.18 (a), show the STOI score for each system when T60 equals 300 ms. It can be seen that the OCM ratio mask, ERMO2-S300, performs best in all input mixture SNR conditions. The CLAS binary mask, ERMC2-S300 yields better STOI scores than the OCM binary mask, EBMO2-S300, for the 5 dB input mixture SNR condition. However, the difference in STOI scores between the two binary mask methods becomes smaller as the SNR decreases. A similar pattern of STOI performance is evident when the reverberation time is increased to 700 ms, as shown in figures 8.16 (b), 8.17 (b) and 8.18 (b)

The PESQ scores for the target speech quality in the 300 ms T60 condition are shown in figures 8.16 (c), 8.17 (c) and 8.18 (c). Again, the OCM ratio mask performs best for all three input mixture SNRs. The OCM binary mask actually reduces the speech quality compared with the original mixture, although the reduction tails away as the angular separation of the interferer from the target increases. The CLAS binary mask also outputs reduced speech quality, with PESQ scores which are substantially lower than the OCM binary mask scores. When the input test mixtures T6O equals 700 ms, the OCM ratio mask shows an improvement in PESQ score compared with the input mixture and both binary mask methods harm the speech quality in all test SNR conditions.

In figures 8.16 (e), 8.17 (e) and 8.18 (e), which show the clean-target-modulated SER performance in the 300 ms reverberant condition, the OCM ratio mask generally performs best, with the CLAS binary mask either coming a close second or, in the case of the 5 dB input mixture SNR condition, narrowly exceeding the performance of the ratio mask. When T60 equals 700 ms, the subplot (e) in each figure demonstrates that the segregated target signals from the OCM ratio mask are consistently most similar to the clean target, followed by the CLAS binary mask, which in turn is followed by the OCM binary mask.

8.4.2 Performance in reverberation for multiple interference sources

In this section we consider more challenging source configurations. Specifically, we evaluate the systems' performance using three and five sources under the two simulated reverberant conditions applied in the previous section when analysing two sources. In the three-source configuration, the two interfering sources are located at azimuths of $-45°$ and $45°$. When there are five sources, the four interferers are located, one in each quadrant, at azimuth angles of $-45°$, $45°$, $-135°$ and $135°$. The OCM and CLAS systems are trained according to the process described in section 8.2 for the four training conditions. These training conditions are summarised in table 8.1, which also lists the 12 associated masks which are generated.

The results of segregating the target speech using each of the masks are summarised in figures 8.19, 8.20 and 8.21 using test input mixture SNRs of 5 dB, 0 dB and -5 dB, respectively. Figure 8.19 (a) shows the mean STOI scores for the directions tested. It can be seen that the OCM ratio masks, ERMO3-S300, ERMO5-S300, ERMO3-S700 and ERMO5-S700, respectively, perform best and that, as observed when using two sources in section 8.4.1, the corresponding CLAS binary masks ranks second and the OCM binary masks perform least well. As the test mixture SNR decreases (subplot (a) in figures 8.20 and 8.21), the ratio mask STOI scores remain the highest.

In all input mixture SNR conditions, the OCM ratio mask generates the highest PESQ scores (subplot (b) in figures 8.19, 8.20 and 8.21). The scores are an improvement on the PESQ score for the input mixture, except for the most challenging five-source setup using a T60 of 700 ms. Similarly, the OCM binary mask produces better speech quality than the CLAS binary mask, except in the most challenging setup. Both sets of binary masks reduce speech quality compared with the input mixture PESQ scores.

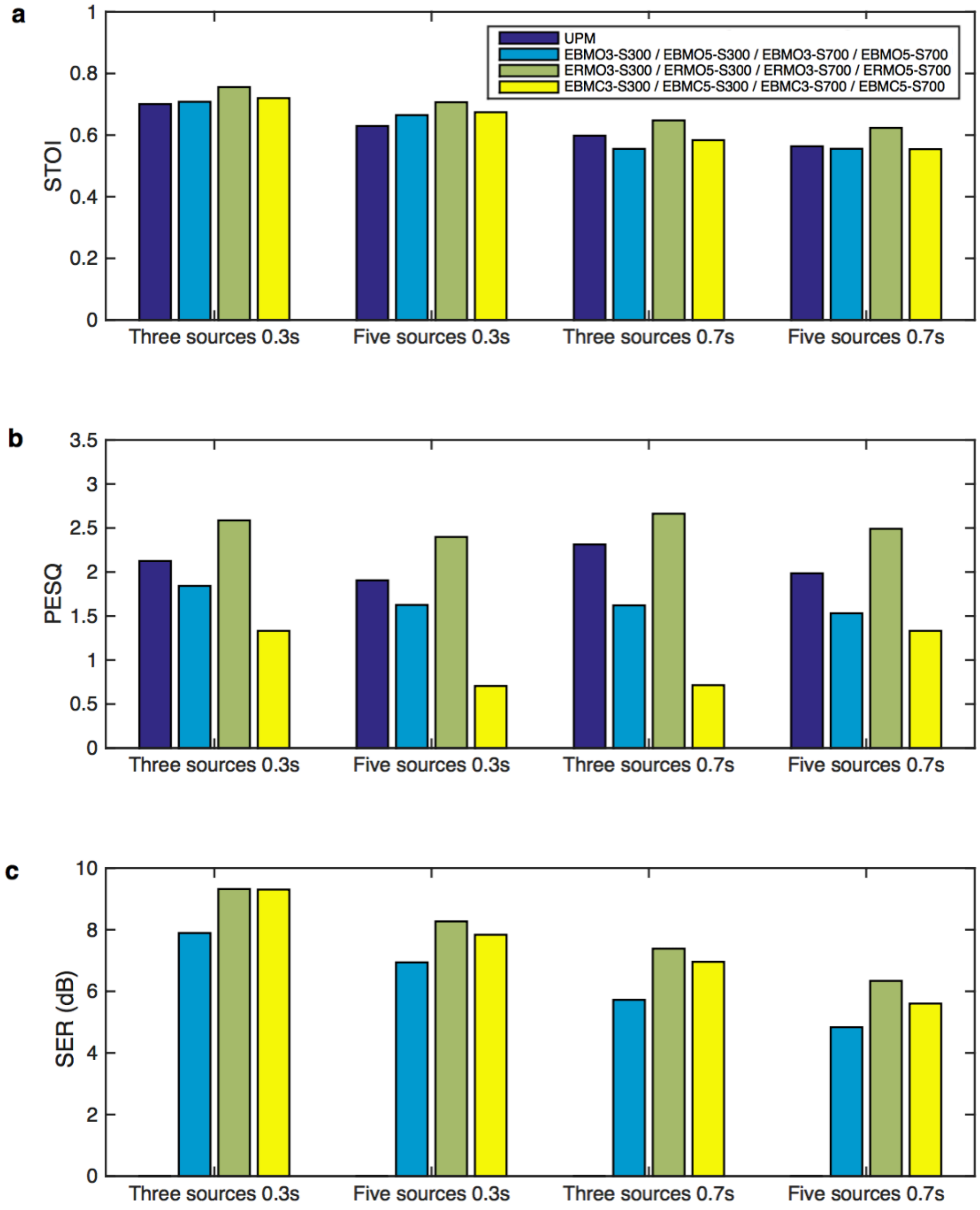


Figure 8.19: Mean STOI and PESQ scores and mean output clean-target-modulated SER for reverberant, three-source and five-source configurations. The T60 reverberation times are 300 ms and 700 ms. The input mixture SNR is 5 dB.

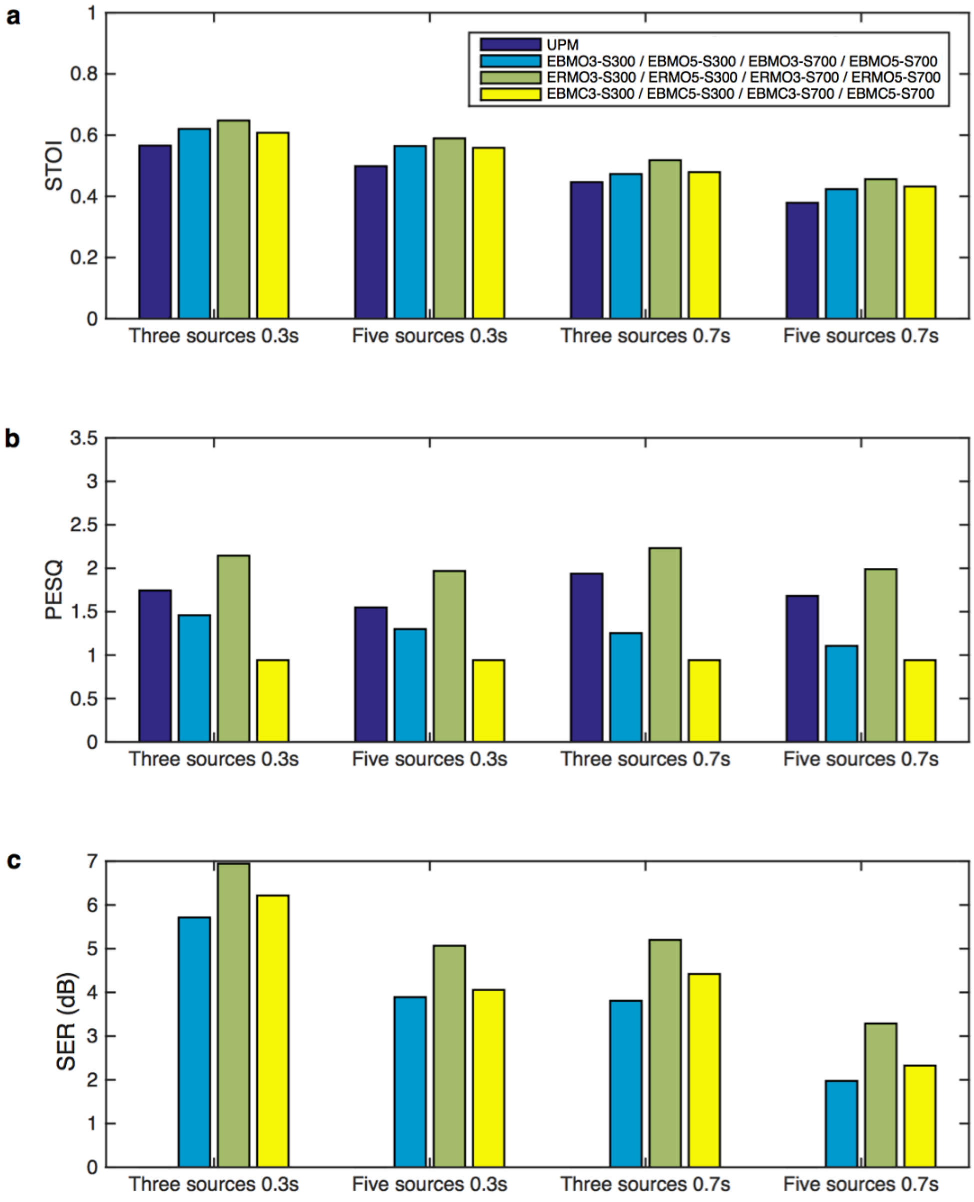


Figure 8.20: Mean STOI and PESQ scores and mean output clean-target-modulated SER for reverberant, three-source and five-source configurations. The T60 reverberation times are 300 ms and 700 ms. The input mixture SNR is 0 dB.

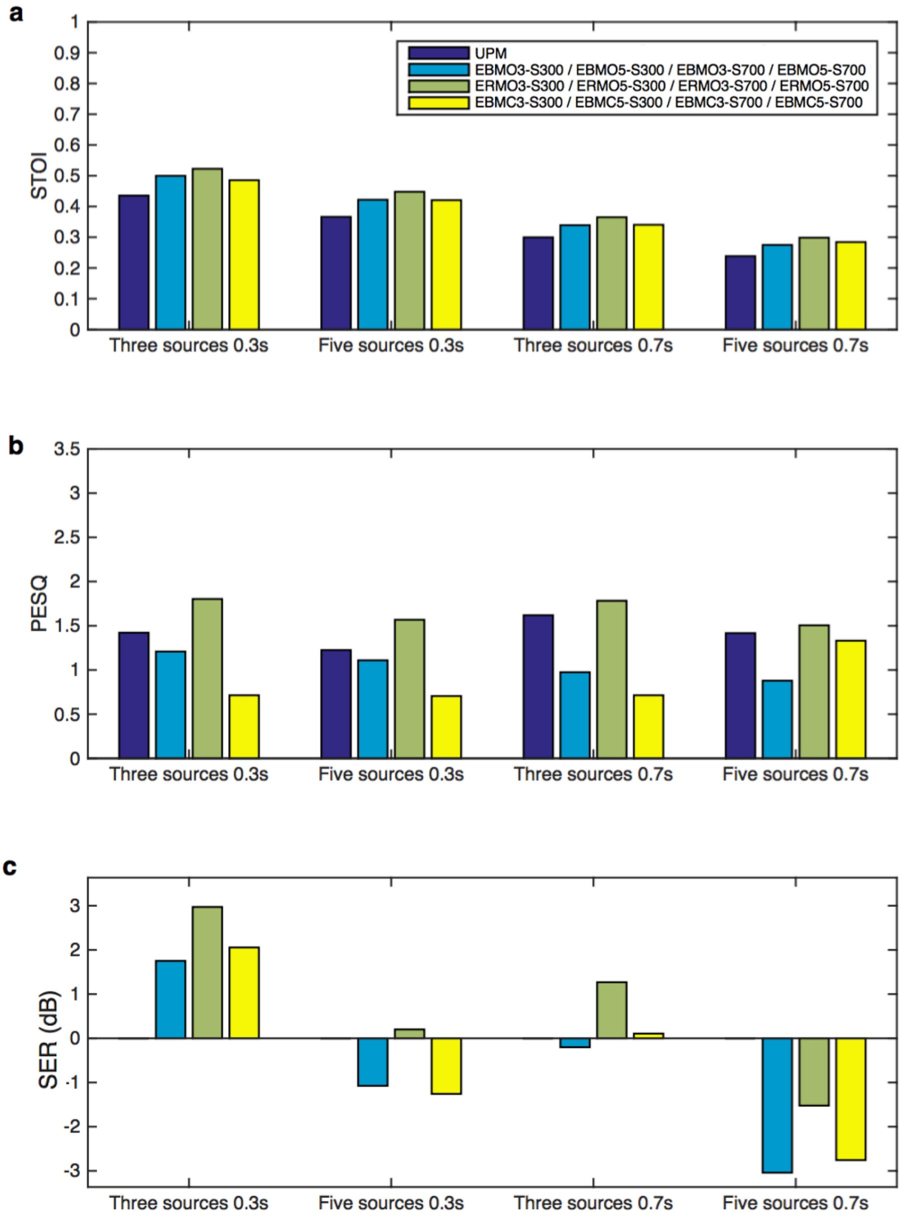


Figure 8.21: Mean STOI and PESQ scores and mean output clean-target-modulated SER for reverberant, three-source and five-source configurations. The T60 reverberation times are 300 ms and 700 ms. The input mixture SNR is -5 dB.

8.4.3 OCM and CLAS system performance in recorded reverberation

We continue to follow the analysis procedures adopted in Jiang et al. (2014) and in this section we compare the performance of our OCM approach and their CLAS approach for target speech segregation in real reverberation. Thus, we use measured binaural room impulse responses to compare the two systems in a two-source configuration.

The real BRIR sets (BRIR-Rnnn) contain four sets of binaural room impulse responses, each captured in a different real room at the University of Surrey using a Cortex (MK.2) Head and Torso Simulator (HATS) (Hummersone, 2010). The rooms, labelled A, B, C and D, have different sizes and reflection characteristics. The initial time delay gap (ITDG), direct-to-reverberant ratio (DRR) and reverberation time (T60) for each room are listed in table 8.3. In each set of measurements, the sound sources are placed from $-90°$ to $90°$ with a $5°$ spacing.

Table 8.3: Room acoustical properties of BRIR-Rnnn. Table reproduced from Hummersone (2010).

Room	ITDG (ms)	DRR (dB)	T60 (s)
A	8.72	6.09	0.32
B	9.66	5.31	0.47
C	11.9	8.82	0.68
D	21.6	6.12	0.89

The training data is prepared from the TIMIT corpus (John et al., 1993) and the NOISEX corpus (Varga and Steeneken, 1993), as described in section 8.2. In the training stage for all the systems, the babble noise is placed at $-90°$ to $90°$ by $10°$ step. The binaural mixture is set to 0 dB SNR in the left channel, as described in section 8.2. The system and mask labels for this analysis are listed in table 8.2. During testing, the babble noise is placed on the left side only, from an azimuth angle of $-90°$ down to $-5°$ in $5°$ step, which is twice the density of directions used during training. Therefore, the test includes both matched (previously seen) and unmatched (previously unseen) cases. 50 sentences are used in each test direction. Further details can be found in section 8.2.

Figure 8.22 illustrates the target segregation results at 5 dB (left plots), 0 dB (central plots) and -5 dB (right plots) in Room A (T60 equals 320 ms). As expected, the figure shows that the STOI and PESQ scores and the clean-target-modulated SER of each system decrease as the test mixture SNR decreases. The OCM ratio mask, ERMO2-R320, remains the best performer of the three masks in terms of STOI and PESQ scores in all test cases. In detail, ERMO2-R320 has a STOI score which is between 3% and 5% more than the scores for the CLAS binary mask, EBMC2-R320, in all input mixture SNR conditions. In the 5 dB SNR condition, the CLAS mask STOI score is approximately 3% higher than the OCM binary mask, EBMC2-R320. The STOI scores are very close to each other in the 0 dB input mixture test condition. In a repeating trend, the OCM binary mask STOI score at -5 dB SNR is slightly higher than the score for the CLAS mask.

The speech quality comparison is shown in figure 8.22 (d), (e) and (f). The results demonstrate that the PESQ score performance is highest for the OCM ratio mask, followed by the OCM binary mask. The CLAS binary mask performs relatively poorly, with a score considerably below the PESQ score for the input mixture at all SNRs. The clean-target-modulated SER displays the normal reversal in performance, with the CLAS mask outperforming the OCM binary mask for this metric. The results in figure 8.22 (g), (h) and (i) add further support for the notion espoused that a relatively high SER value does not necessarily correlate with higher speech intelligibility or quality. This is particularly strongly demonstrated in the 5 dB test condition (left column), where it can be seen that the CLAS mask displays higher SER values than the OCM ratio mask for some directions, whereas the STOI and PESQ scores in the corresponding directions are much worse. At lower input SNRs, the difference between the STOI and PESQ performances of the systems become smaller. However, they still follow the pattern: OCM ratio mask, then CLAS binary mask and finally the OCM binary mask.

Similar tests in Room B, C and D are also carried out. Their T60s are 470 ms, 680 ms, and 890 ms and their results are shown in figures 8.23, 8.24, and 8.25, respectively. The performance of each mask in the different rooms follows the same pattern as the results for Room A. The OCM ratio mask produces the highest STOI and PESQ scores, and, in most cases, the highest clean-target-modulated SER. The OCM binary masks with the 5 dB input mixture do not have as high a STOI score as the CLAS binary mask. However, the difference between the STOI scores for the two binary masks becomes smaller as the test SNR decreases. The observed PESQ scores demonstrate that the OCM ratio mask enhances the target speech quality in all test cases whereas the OCM binary masks do not, in general, improve it. The CLAS binary mask invariably reduces the PESQ score compared with the score for the input mixture.

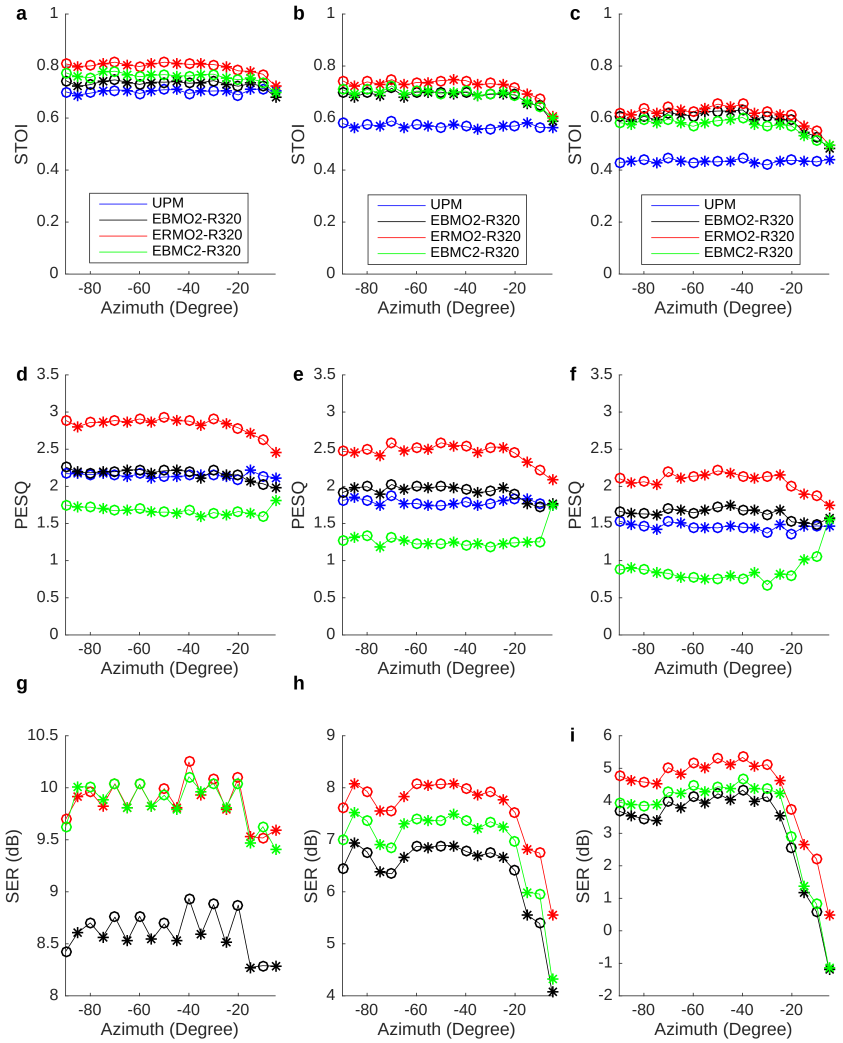


Figure 8.22: Room A (T60 = 320 ms). STOI and PESQ scores and output clean-target-modulated SER performance for reverberant, two-source configurations as a function of interferer direction. Mixture SNRs are 5 dB (left column), 0 dB (middle column) and -5 dB (right column). ∘ and ∗ denote matched and unmatched cases, respectively. The black line in (b) is obscured by the green line.

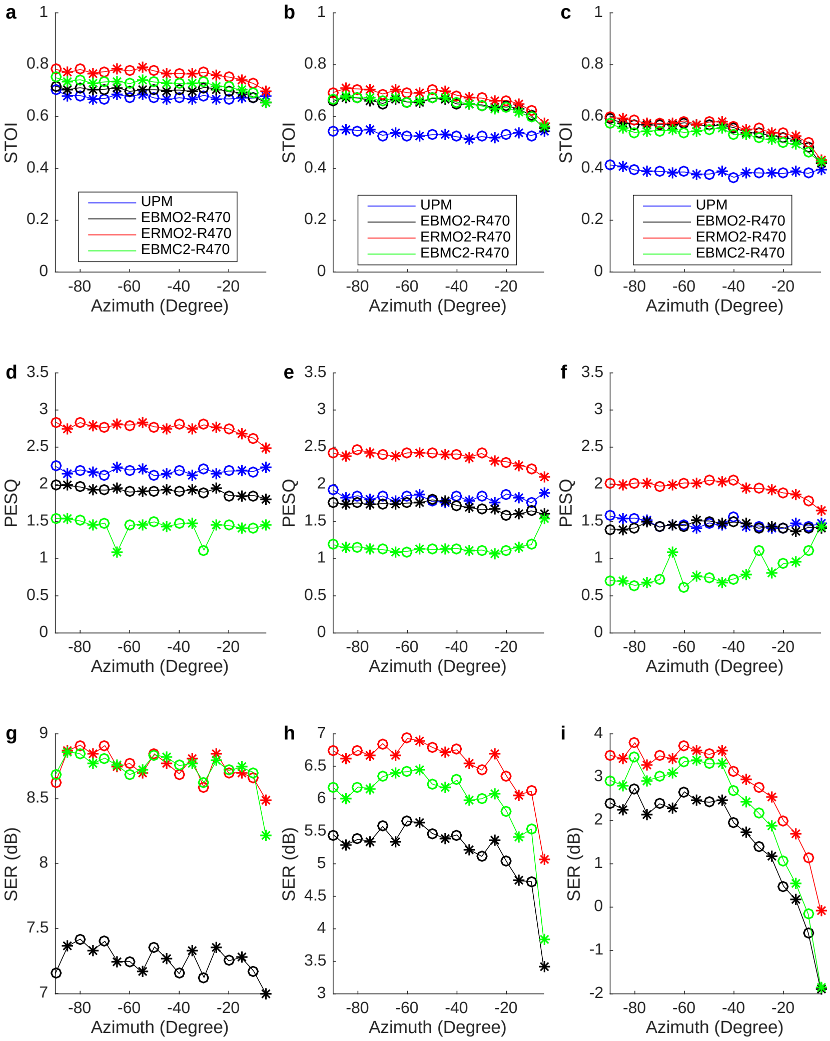


Figure 8.23: Room B (T60 = 470 ms). STOI and PESQ scores and output clean-target-modulated SER performance for reverberant, two-source configurations as a function of interferer direction. Mixture SNRs are 5 dB (left column), 0 dB (middle column) and -5 dB (right column). ○ and ∗ denote matched and unmatched cases, respectively. The black line in (b) and (c) is obscured by the green line.

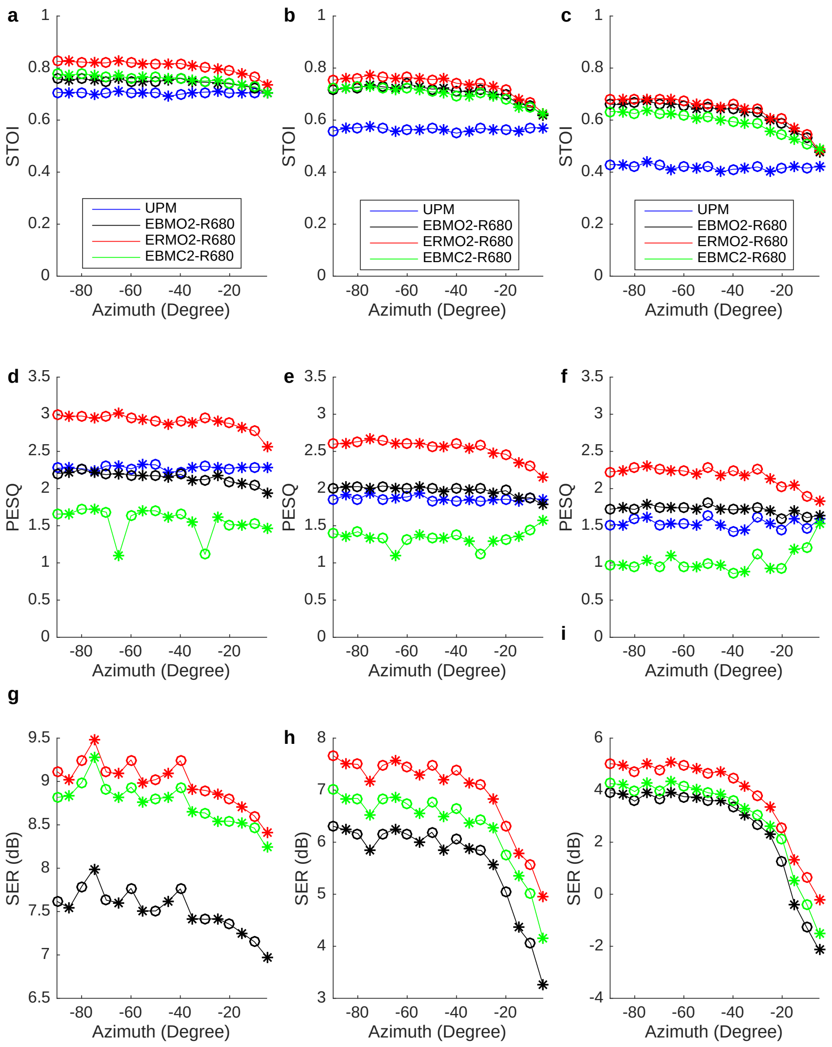


Figure 8.24: Room C (T60 = 680 ms). STOI and PESQ scores and output clean-target-modulated SER performance for reverberant, two-source configurations as a function of interferer direction. Mixture SNRs are 5 dB (left column), 0 dB (middle column) and -5 dB (right column). ○ and $*$ denote matched and unmatched cases, respectively. The black line in (b) is obscured by the green line.

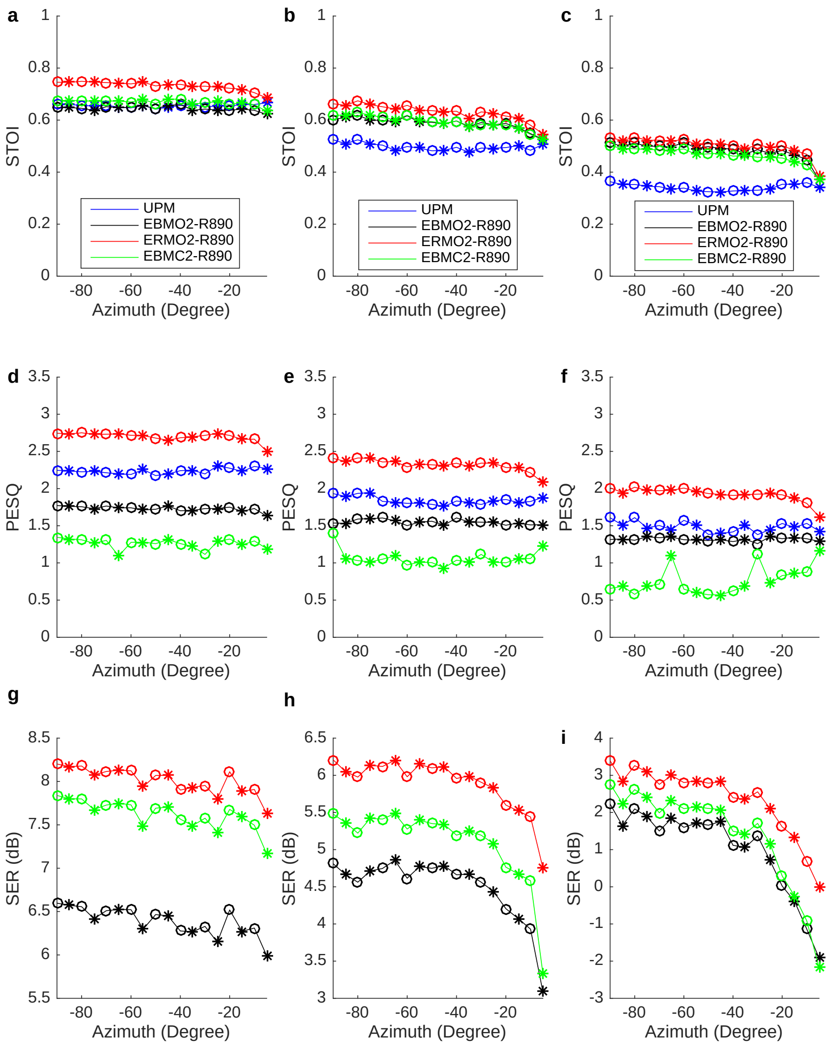


Figure 8.25: Room D (T60 = 890 ms). STOI and PESQ scores and output clean-target-modulated SER performance for reverberant, two-source configurations as a function of interferer direction. Mixture SNRs are 5 dB (left column), 0 dB (middle column) and -5 dB (right column). ∘ and ∗ denote matched and unmatched cases, respectively. The black line in (b) is obscured by the green line.

8.5 Comparison of generalisation ability

In the preceding sections, we evaluate and compare the OCM and CLAS algorithms in a variety of conditions. These include the performance of both systems in anechoic and reverberant environments. The reverberation analysis involves simulated and recorded reverberation. The number of interferers varies from one interference source up to four. Both systems trained for the 0 dB SNR condition have the ability to generalise and operate at other signal-to-noise ratios, namely at -5 dB and 5 dB. A limitation of the evaluation process so far, however, is that both the OCM and CLAS estimators have been evaluated and compared in the conditions for which they were trained. Specifically, a system trained in anechoic conditions has not been tested in reverberant conditions and a system trained for one set of interferers has not been tested with a different number and arrangement. Both of these situations will arise in practice and in this section we therefore test and compare the generalisation ability of both estimator approaches further in each of these dimensions.

The first scenario examined in this section is to train the estimators using signals that contain N sources and test them using mixtures containing M sources, where $N \neq M$. Detailed experiments are described in section 8.5.1. The second scenario is to evaluate the performance of estimators that were trained on signals generated in one room when tested using mixtures generated in a different one. Detailed experiments in this scenario are presented in section 8.5.2. The ability of an estimator to generalise in these circumstances is crucial for their intended application in a hearing aid, since the number of interferers may vary from moment to moment and a hearing aid user may move from one acoustic space to another.

8.5.1 Generalisation ability for different numbers of interferers

In this section, we extend our generalisation analysis of OCM and CLAS systems to consider test configurations which differ from the training configuration in the number and arrangement of sound sources. Thus, systems trained using many sources are tested on fewer sources and vice versa. We inherit the systems previously trained in sections 8.3 and 8.4 and test them using different source configurations. We examine OCM and CLAS estimators trained using two sources, three sources and five sources in anechoic and two reverberant conditions (T60 = 300 ms and T60 = 700 ms). Details of the training process can be found in section 8.2. For the sake of brevity we focus on one acoustic condition and investigate the effect on performance of altering the number and directions of sources in a reverberant environment with T60 = 300 ms. Results for two other acoustic environments, one anechoic and the other a different reverberant environment with T60 = 700 ms, are provided in appendix A.

Three different source configurations in three different acoustic spaces results in nine different setups. For each setup, the OCM binary and ratio mask estimators and the CLAS binary mask estimator are tested using signals containing two, three and five sources, in the same acoustic space that they were trained in. The effect of altering the acoustic space will be considered in the next section.

Throughout this work, the target source continues to be located at 0° azimuth. For the two-source configuration the single interfering source is located at $-45°$ azimuth. In the three-source case, the two interferers are placed at $\pm 45°$ azimuth. In the five-source configuration, the four interferers are sited at 75°, 135°, $-30°$ and $-120°$, respectively. These directions are chosen such that they possess substantially different ITDs and so appear strongly directionally distinct, since they each lie on a different cone of confusion (see section 3.1). Each test case contains 50 randomly selected sentences from the test set in the TIMIT corpus (John et al., 1993). More information about the generation of test data can be found in section 8.2. The SNR for all training and test mixtures is set at -5 dB, which is the most challenging SNR of the three employed throughout this research.

Figures 8.26, 8.27 and 8.28 show the results for each system tested using two sources, three sources and five sources in an acoustic environment with 300 ms reverberation time. Each subplot consists of three results, which show the improvement in performance of the three estimators compared with the unprocessed mixture for each of the three different numbers of sources. Figures 8.26 (a) left plot, 8.27 (a) middle plot and 8.28 (a) right plot are the reference results, for the cases in which the total number of sources (N) used in the training phase is equal to the total number of sources (M) used in the testing phase (i.e. N and M are both equal to two, three and five, respectively). Similar analyses have been performed previously in sections 8.4.1 and 8.4.2 and these results are in close agreement with them. The results indicate that the OCM ratio mask performs best in the three systems, followed by the OCM binary mask and the CLAS binary mask. The STOI speech quality is improved by 13.9 %, 8.7 % and 10.0 % for the estimators specifically trained and tested using two sources, three sources and five sources, respectively. These results serve as a reference against which the performance differences for the test cases where $N \neq M$ can be compared. For ease of comparison, the reference results (i.e. when N = M) are also shown as horizontal red lines on each of the corresponding generalised results (i.e. when $N \neq M$). In general, the OCM ratio mask yields the best STOI improvement in all three figures. It is approximately 2 % to 3 % better than the corresponding OCM binary mask system. The CLAS system sits in third place and shows a further reduction in performance of approximately 1 % compared with the OCM binary mask estimator.

All the estimators evaluated here demonstrate the ability to generalise and to perform satisfactorily when the number of interfering sources differs from the training stage. Figure 8.26 (a) shows the generalisation ability of the estimators, O2-S300 and C2-S300, trained using two sources and tested with three sources and five sources. In detail, the ratio mask estimator ERMO2-S300 improves the STOI score most when testing the three-source and five-source configurations, with 8.6 % (figure 8.26 (a) middle plot) and 9.8 % (figure 8.26 (a) right plot) STOI improvements, respectively. The corresponding mask sets which are trained using the three-source configuration (figure 8.27 (a) middle plot) and the five-source configuration (figure 8.28 (a) right plot) show improvements of 0.1 % and 0.2 % more than ERMO2-S300 trained using two sources and tested using three and five sources, respectively.

In figure 8.27, the systems are trained using the three-source configuration and tested using two, three and five sources. Similarly, all the systems have the ability to generalise to the two-source and five-source configurations. The improvements in STOI scores are 13.4 % and 10.0 % for ERMO3-S300 when testing using two sources and five sources, respectively. The performance of the OCM binary mask estimator, EBMO3-S300, is approximately 2 % lower and the CLAS binary mask estimator, EBMC3-S300, is approximately 3 % lower. Estimator ERMO3-S300, trained using three sources and tested using two sources, scores 0.5 % lower than estimator ERMO2-S300, trained and tested using two sources (figure 8.26 (a) left plot). ERMO3-S300, when tested using five sources, compared with ERMO5-S300, trained and tested using five sources (figure 8.28 (a) right plot) scores only 0.03 % lower.

In figure 8.28 (a), the results reveal the ability of the OCM and CLAS mask estimators trained using five sources to generalise to two-source and three-source configurations in terms of their STOI performance. The same performance ranking of the three systems is observed here, with the OCM ratio mask estimator as usual showing the greatest improvement in STOI score. The ERM05-S300 estimator yields a 12.8 % and 8.4 % STOI improvement when tested using the two-source and five-source configurations. On the other hand, compared with the estimators both trained and tested using two (figure 8.26 (a) left plot) and five sources (figure 8.27 (a) middle plot), ERMO5-S300 displays a 1.1 % and 0.3 % reduction in STOI score, respectively.

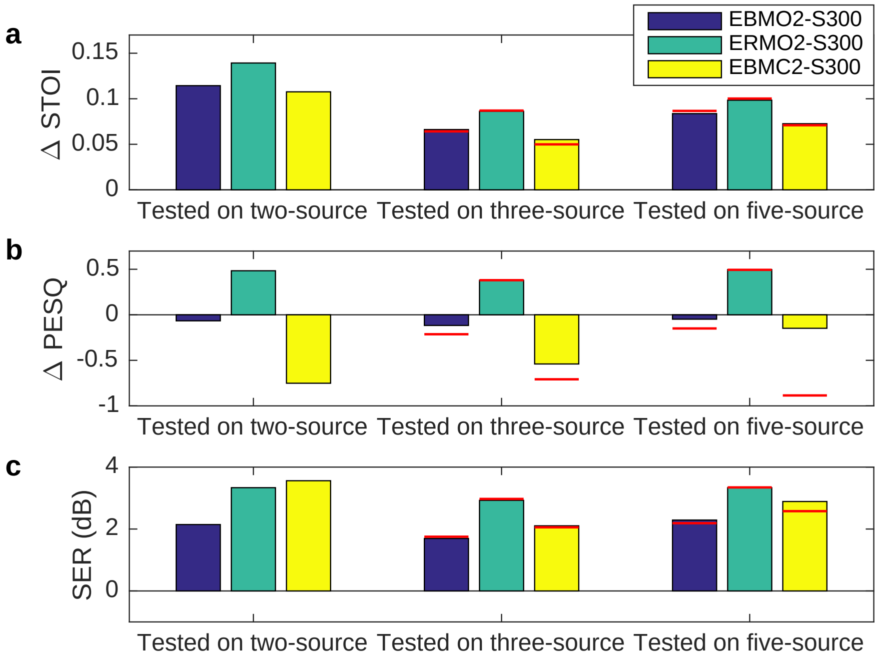


Figure 8.26: The STOI and PESQ improvement compared with the corresponding unprocessed mixtures and SER performance for systems which are trained on signals containing two sources and tested on signals containing two, three and five sources, respectively. (a), (c) and (e) are the STOI improvement, PESQ improvement and SER results, respectively, using -5 dB input mixture SNR (T60 = 300 ms). The red horizontal lines show the corresponding reference results from estimators tested using the same number of sources that they were trained with.

With respect to the generalisation ability of these estimators in terms of speech quality improvement, figures 8.26 (b), 8.27 (b) and 8.28 (b) show that the PESQ score is improved by more than 0.3 for the OCM ratio mask estimator, even when the numbers of interfering sources used in the training and testing phases are not identical. Both the OCM binary mask estimator and CLAS binary mask estimator degrade speech quality, no matter whether the number of interfering sources is identical or not during training and testing. A measure of the numerical similarity between the segregated signal and the clean target signal in terms of the SER metric is also provided in row

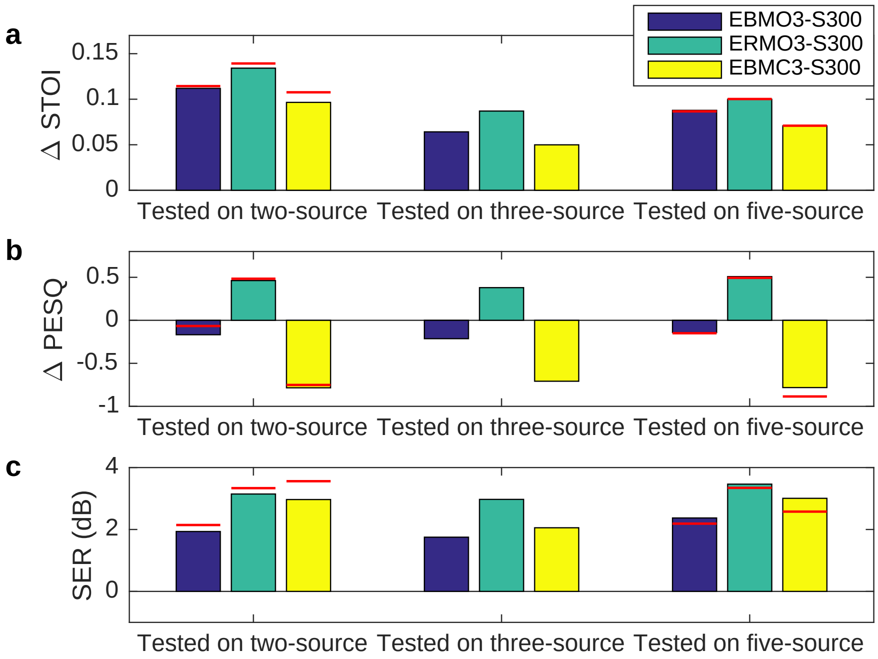


Figure 8.27: The STOI and PESQ improvement compared with the corresponding unprocessed mixtures and SER performance for systems which are trained on signals containing three sources and tested on signals containing two, three and five sources, respectively. (a), (c) and (e) are the STOI improvement, PESQ improvement and SER results, respectively, using -5 dB input mixture SNR (T60 = 300 ms). The red horizontal lines show the corresponding reference results from estimators tested using the same number of sources that they were trained with.

(c) of figures 8.26, 8.27 and 8.28. As for previous analyses, the CLAS binary mask fares somewhat better in comparison to the OCM estimators, though the ratio mask still performs best.

Similar results using the same sets of source setups are presented in Appendix A for two further acoustic environments. Figures A.9, A.10 and A.11 show the results for estimators trained in an anechoic environment using two, three and five sources, respectively. Figures A.12, A.13 and A.14 show corresponding results for estimators trained using two, three and five sources, respectively, in an environment with a 700 ms reverberation time. Broadly

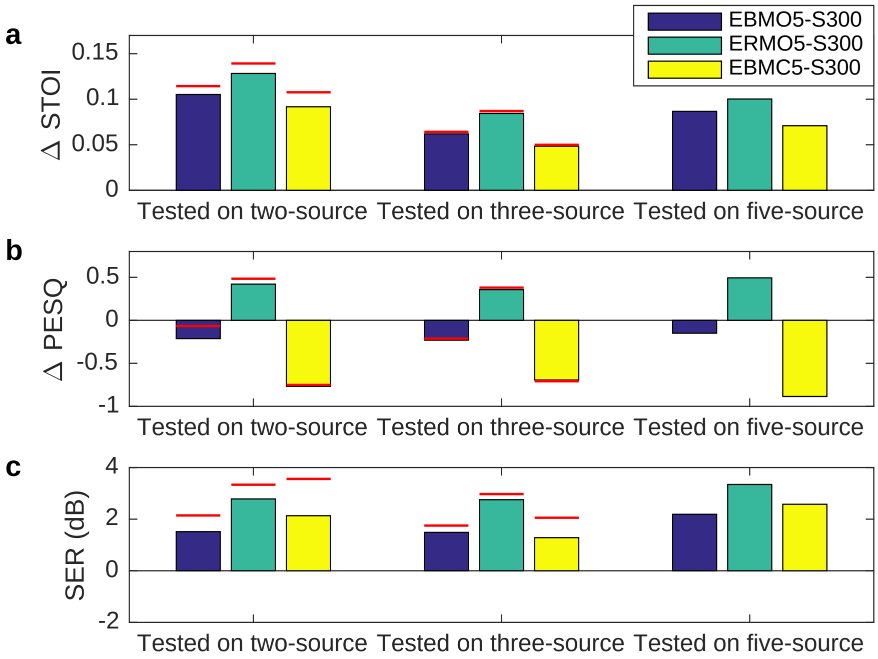


Figure 8.28: The STOI and PESQ improvement compared with the corresponding unprocessed mixtures and SER performance for systems which are trained on signals containing five sources and tested on signals containing two, three and five sources, respectively. (a), (c) and (e) are the STOI improvement, PESQ improvement and SER results, respectively, using -5 dB input mixture SNR (T60 = 300 ms). The red horizontal lines show the corresponding reference results from estimators tested using the same number of sources that they were trained with.

similar patterns of results are observed in both acoustic spaces.

In summary, all the evaluated systems in this section demonstrate an ability to generalise in terms of improving STOI and PESQ scores when the number of interfering sources used for testing differs from the number used for training. There are several unexpected results arising from the analysis conducted in the reverberant environment T60 = 300 ms. Binary masks EBMO2-S300 for OCM and EBMC2-S300 for CLAS (both trained using two sources) very slightly exceed the STOI performance of the reference mask estimator when tested using three sources, as shown in the middle plot of figure 8.26 (a). These two binary mask estimators also outperform the same reference estimators in terms of their PESQ scores (figure 8.26 (b) middle and right plot). These phenomena are not observed in the 700 ms reverberation time condition (see figure A.12). The level of generalisation for the ratio mask estimator trained using two sources in the T60 = 300 ms reverberant environment (see figure 8.26) is surprisingly high for every metric (STOI and PESQ scores, and SER). When tested using three and even five sources, this estimator equals the performance of the corresponding reference estimators. In figure A.10, the three-input anechoic sources ratio mask estimator provides a 24.6 % and 20.9 % improvement in STOI and PESQ scores, respectively, when fed with two-source anechoic test data. This represents a shortfall of 9.5 % and 7.5 % in STOI and PESQ scores, respectively, compared with the reference estimator in this condition. Therefore, some generalisation is occurring, but not as much as might be expected in view of the fact that the source directions in the two-source test data are included in the three directions that this three-source estimator was trained with. Broadly similar results are observed using the OCM and CLAS binary masks. These observations are worthy of further investigation. For now, however, it is sufficient to note that generalisation across the dimension of varying the number of sources has been successful, indeed remarkably so in some conditions.

8.5.2 Generalisation ability for different room acoustics

.

In the previous section, we examined the ability of the three mask estimators to segregate sources successfully when tested using a different number of interfering sources compared with the number they were trained with. All those evaluations were limited to using the same acoustic space. In this section, we investigate how all three estimators perform when the acoustic space is changed by training the estimators using BRIRs for one room and testing them using BRIRs for different rooms.

In section 8.4.3, we trained OCM and CLAS estimators using the real BRIR sets measured by Hummersone (2010). The BRIR sets consist of measurements of four different rooms, labelled A, B, C and D, with different sizes and reflection characteristics (see section 8.4.3). Each estimator is trained using speech examples spatialised in a different individual room and it is then tested using test data generated from all four rooms. Since the focus of this section is on the generalisation ability of the estimators as a function of varying room acoustics, we apply the simplest two-source configuration mixture as the test data. We employ the test strategy described in section 8.4.3. The target source is located at 0° azimuth and the direction of the interfering source ranges from $-90°$ to $-5°$ azimuth in steps of 5°. Each test case contains 50 randomly selected sentences from the test set in the TIMIT corpus (John et al., 1993). More detail about the generation of the test data can be found in section 8.2.

Figure 8.29 shows the target segregation test results at 5 dB SNR in room A for the mask estimators trained on data spatialised in rooms A, B, C and D, respectively. The labels in the legend are defined in table 8.2. Plots (a), (d) and (g) show the STOI, PESQ and SER results, respectively, for four OCM ratio estimators, each trained in a different room. As expected, reference estimator ERMO2-R320, trained and tested in room A, yields the greatest STOI improvement (the highest line in the plots), improving the STOI score by almost 0.2 for most interferer directions. As the direction of the interfering source approaches that of the target, the improvement decreases. The estimators trained using data spatialised in rooms B, C and D increase the STOI score of room A test data slightly less, in most cases by approximately 0.15. When the direction of the interfering source is close to the target, such as at $-5°$ azimuth, the performances of all the estimators fall and become very similar to that of the reference estimator ERMO2-R320.

The speech quality scores behave differently from the speech intelligibility scores. The PESQ scores obtained here for all the systems in room A are very close to each other. This demonstrates that all the three mask estimators trained in different rooms have a similar ability to generalise in terms of their PESQ scores and their performance is almost the same as the mask estimator trained in room A. The SER metric also indicates that the reference estimator, ERMO2-R320, outperforms the other three mask estimators, which all perform similarly to each other.

The STOI and SER metrics for the four OCM binary masks (EBMO2-Rnnn) and the four CLAS masks (EBMC2-Rnnn) in figure 8.29 (b), (e) and (h), and (c), (f) and (i) follow a similar pattern to the OCM ratio masks just discussed. The estimators trained on data spatialised using BRIRs for rooms B, C and D have the ability to segregate the target when tested using room A data. Again, there are improvements in the STOI scores, although not as much as those exhibited by estimators EBMO2-R320 and EBMC2-R320, which are trained and tested in room A. In terms of speech quality, the four EBMO2 binary mask estimators produce a PESQ score which is, on the whole, marginally inferior to that for the unprocessed mixture. The four CLAS binary mask estimators, EBMC2-Rnnn, significantly degrade the speech quality. This observation mirrors the results we obtained in section 8.4.3. The SER metrics also follow a similar pattern to the OCM ratio mask results.

The results for all the other conditions investigated are shown in figures 8.30 (estimators trained on data spatialised in room B), 8.31 (trained in room C) and 8.32 (trained in room D), respectively. They all follow the same pattern as the results shown in figure 8.29.

Compared with the mean STOI scores for the OCM ratio mask reference estimators (i.e. those estimators trained in one environment and tested in the same environment), estimators trained in a different room from which they were tested exhibit a mean STOI score depressed by approximately 0.055, 0.037, 0.055 and 0.017 for rooms A, B, C and D respectively. This corresponds to a range of between 1.7 % and 5.5 %. The results suggest that all the estimators have a substantial ability to generalise along the dimension of varying acoustic environment.

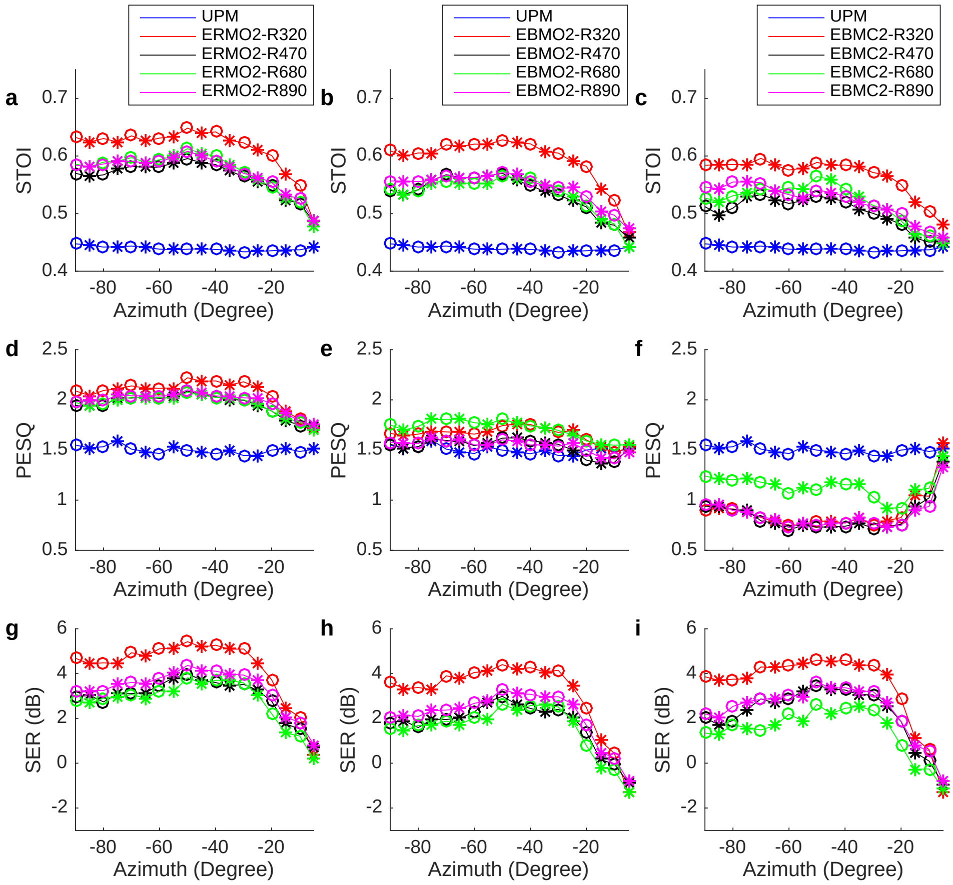


Figure 8.29: The STOI and PESQ and SER performance for systems which are trained on data for room A, B, C and D, respectively, using test mixture with -5 dB SNR in room A (T60 = 320 ms). (a), (d) and (g) are the STOI, PESQ and SER results for OCM ratio masks, respectively. (b), (e) and (h) are the STOI, PESQ and SER results for OCM binary masks, respectively. (c), (f) and (i) are the STOI, PESQ and SER results for CLAS masks, respectively.

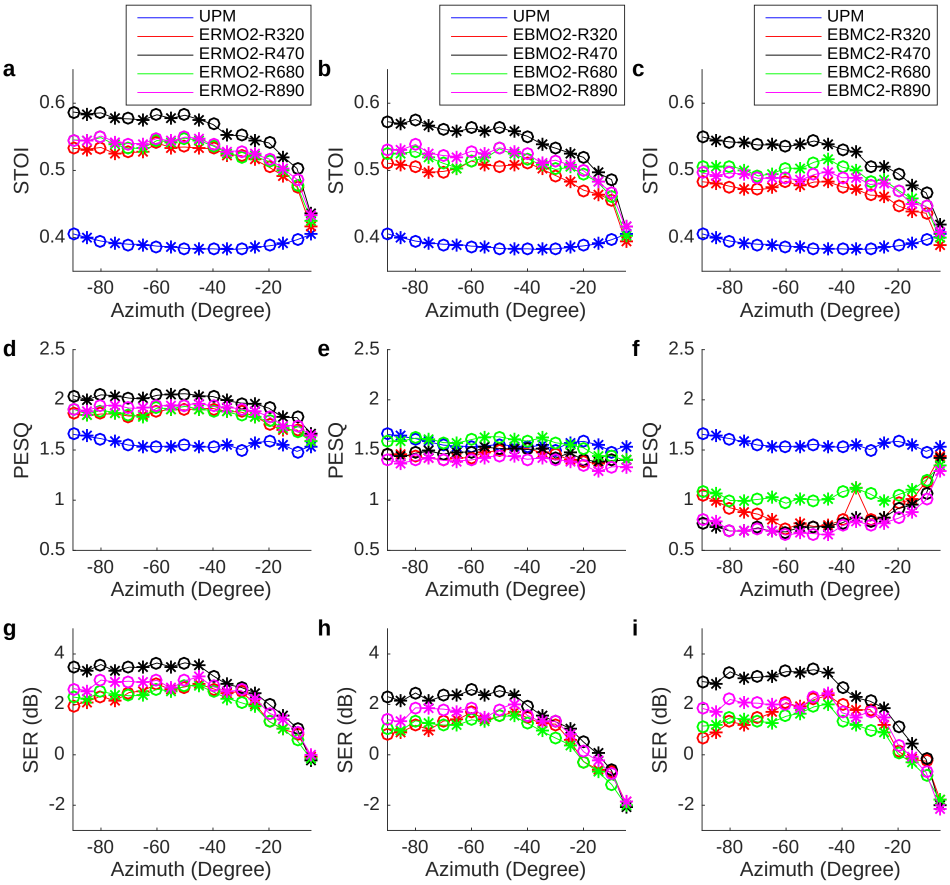


Figure 8.30: The STOI and PESQ and SER performance for systems which are trained on data for room A, B, C and D, respectively, using test mixture with -5 dB SNR in room B (T60 = 470 ms). (a), (d) and (g) are the STOI, PESQ and SER results for OCM ratio masks, respectively. (b), (e) and (h) are the STOI, PESQ and SER results for OCM binary masks, respectively. (c), (f) and (i) are the STOI, PESQ and SER results for CLAS masks, respectively.

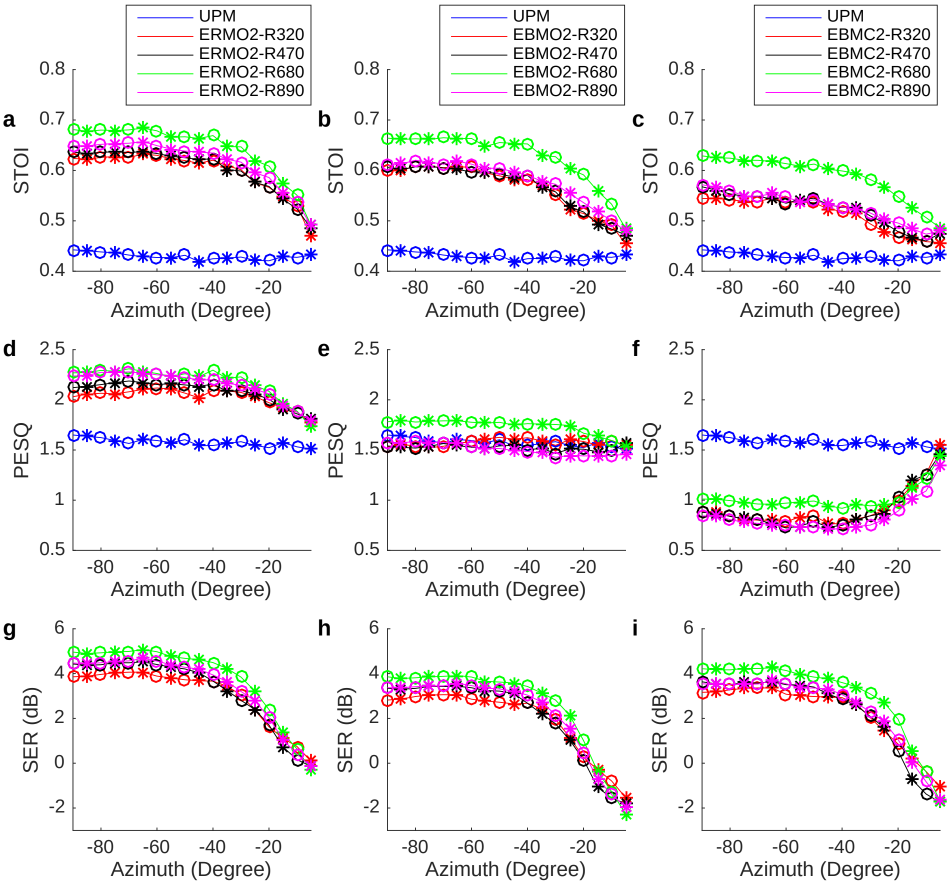


Figure 8.31: The STOI and PESQ and SER performance for systems which are trained on data for room A, B, C and D, respectively, using test mixture with -5 dB SNR in room C (T60 = 680 ms). (a), (d) and (g) are the STOI, PESQ and SER results for OCM ratio masks, respectively. (b), (e) and (h) are the STOI, PESQ and SER results for OCM binary masks, respectively. (c), (f) and (i) are the STOI, PESQ and SER results for CLAS masks, respectively.

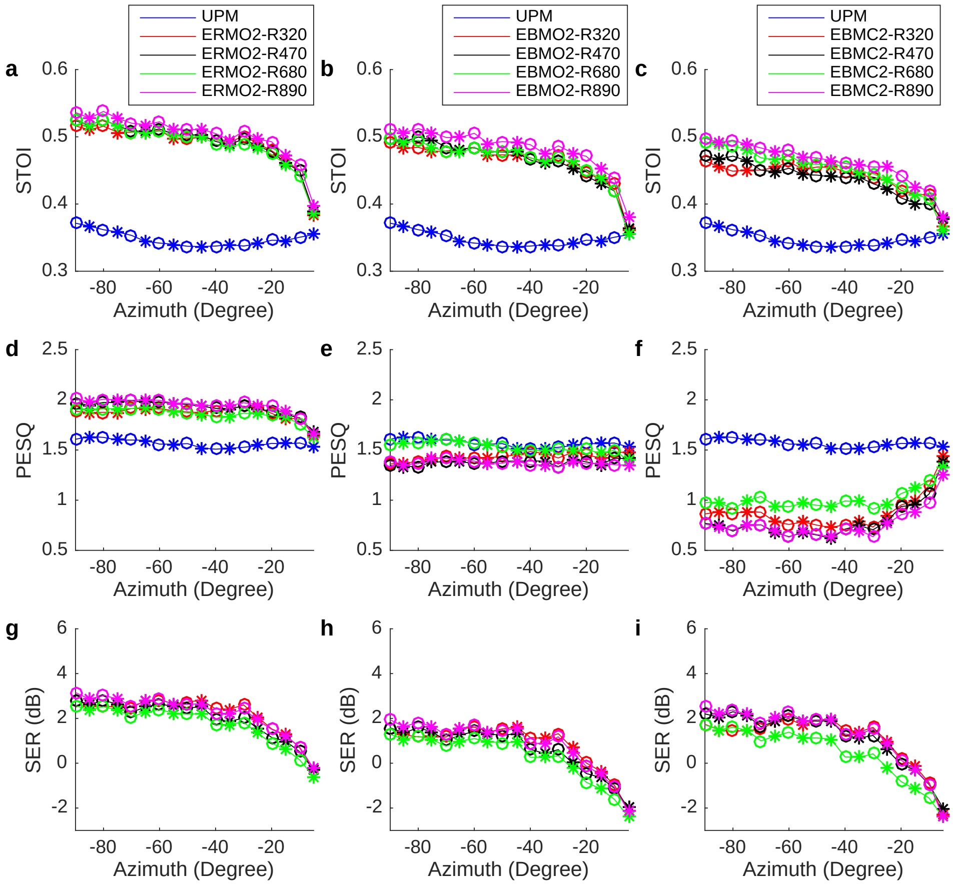


Figure 8.32: The STOI and PESQ and SER performance for systems which are trained on data for room A, B, C and D, respectively, using test mixture with -5 dB SNR in room D (T60 = 890 ms). (a), (d) and (g) are the STOI, PESQ and SER results for OCM ratio masks, respectively. (b), (e) and (h) are the STOI, PESQ and SER results for OCM binary masks, respectively. (c), (f) and (i) are the STOI, PESQ and SER results for CLAS masks, respectively.

8.6 Discussion

This chapter has compared the speech segregation performance of our proposed optimal cue mapping (OCM) approach and a state-of-the-art DNN-based classification (CLAS) algorithm (Jiang et al., 2014). The performance of the systems was examined using three metrics. We first compared the performance of the OCM and CLAS algorithms in anechoic configurations. The comparisons consisted of two-source, three-source and five-source setups using input mixtures with different test SNRs. We also compared them in reverberant conditions with simulated binaural room impulse responses (BRIRs). In addition, the simulated BRIRs were used to generate binaural speech mixtures to assess the OCM and CLAS algorithms in more realistic conditions. These synthesised BRIRs still have short-comings, however, as they simulate the acoustics of a simple rectangular prism room without any obstacles. Due to the generally poor subjective quality of the simulated BRIRs (Hummersone, 2010), we went on to compare the OCM and CLAS algorithms using real recorded BRIRs. The comparisons again included a range of source configurations and used three different input mixture test SNRs. Finally, we assessed the generalisation ability of the estimators in three dimensions: generalisation to different test SNRs, generalisation to different numbers of sources, and generalisation to different room acoustics. These three scenarios are highly relevant performance indicators for speech segregation tasks in applications such as hearing aids and automatic speech recognition systems. This is because the mixture SNR and the number of interfering sources are not constant in real life and the movement of a hearing aid user between one environment and another may cause the characteristics of the acoustic space they are in to vary radically.

The evaluation results illustrate that the OCM ratio and binary masks consistently improve target speech intelligibility in both anechoic and reverberant conditions. Furthermore, the OCM ratio mask improves target speech quality in all the tests which were conducted. The improvement is reduced, however, in reverberant conditions compared with anechoic conditions. To a lesser extent, the OCM binary masks enhance target speech quality in anechoic configurations, but the binary mask has a detrimental effect on speech quality in most of the tested reverberant conditions, both using simulated BRIRs and measured BRIRs of real rooms.

Since IPD and ILD form two of the inputs to our OCM estimators, these estimators partially depend on binaural cues for their operation. In reverberant environments the early and late reflections interact with the direct sound, disturbing the binaural cues, which therefore become less reliable than they are in anechoic conditions. Hence, speech segregation in reverberation is more difficult than it is in anechoic conditions and this degrades the performance of the OCM algorithm in such conditions.

As expected, our comparisons also show that the OCM ratio mask achieves more robust segregation than the corresponding OCM binary mask. Furthermore, the performance of the OCM binary mask exceeds that of the CLAS binary mask in all anechoic configurations. It is revealing to investigate the performance differences between the two binary mask approaches. In two-source anechoic conditions, the OCM binary mask performs better in terms of both target speech intelligibility and quality. However, in the three-source, 5 dB input SNR test condition, its performance is inferior to that of the CLAS binary mask. It also performs worse than the CLAS binary mask for some test cases using the simulated and recorded BRIRs at different input mixture SNRs. It is likely that this is because the OCM algorithm estimates the mask at a finer frequency resolution and it also lacks helpful inputs representing the monaural features of the input mixture. It is thought that these differences account, at least in part, for the increased performance of the CLAS binary mask in some conditions.

The generalisation ability of a system which learns relates to its performance when exposed to test data to which it has not been exposed during training. Generalisation ability not only depends on the nature of a neural network as a universal function approximator, but it also depends on the strategies used in the training process, such as choosing the number of hidden neurons (see section 5.3.2). If the number of hidden neurons is not large enough, the system is unable to learn effectively. Too many hidden neurons, on the other hand, can bestow too much power on the network, so that it is possible for it to overfit the training data, reducing its ability to generalise. The balance and diversity of the training data set are also critical for successful training of the network.

Generalisation is a critical issue for practical speech segregation algorithms. In terms of generalisation ability to different mixture SNRs, estimators were trained using 0 dB SNR mixtures and tested using -5 dB, 0 dB and 5 dB SNR mixtures. All estimators improve speech intelligibility in all the tested SNR conditions. On the other hand, when the test mixture has a low SNR, the PESQ scores are degraded when a binary mask is used, and this applies to both the OCM and CLAS approaches. The OCM ratio mask estimator is the only one that improves speech quality in all the test conditions investigated. When the number of interfering sources between training and testing changes, the results show that all the systems have the ability to generalise. Further generalisation performance evaluation with regards to changes in the acoustic space demonstrate that estimators trained in one room are able to segregate a target speech source in a variety of reverberant conditions and to improve its intelligibility.

Chapter 9

Conclusion

This research has led to a method for improving the intelligibility and quality of a target speech source in a binaural mixture with spatially distinct competing concurrent interfering sounds in anechoic and reverberant conditions. The long term goal is to incorporate the algorithm in a hearing aid capable of assisting people with normal hearing or with a hearing deficit in acoustic conditions which they find challenging.

The proposed optimal cue mapping (OCM) approach, described in Chapter 5, includes a straightforward machine learning algorithm which estimates a binaural ratio mask for speech segregation. Through modelling, it has been demonstrated that OCM has the ability to integrate binaural cues and other acoustic features to successfully segregate the target speech. Using the two relative importance methods outlined in Chapter 5, it is possible to determine the varying relative importance of the input cues for estimating the ratio mask under a range of typical acoustic conditions so that only the most important features are extracted and used for mask estimation according to the prevailing acoustic conditions. The segregated speech shows raised perceptual intelligibility and quality and compares favourably with a current state-of-the-art solution.

In Chapters 7 and 8 we have selected seven cues for the purpose of demonstrating the OCM method. The cues are easily integrated into the OCM algorithm and the range of cues is expandable to include further properties of the acoustic environment and to introduce characteristics of the source. By taking advantage of the generalisation ability of artificial neural networks (ANNs), we have shown how it is possible to reduce the training data size without creating a significant impact on the performance of the binary and ratio masks they generate. The key benefit of doing this is a reduction in training time.

In hearing aid algorithm design, for example, knowledge of the contribution of each type of cue to the mask estimation helps to construct an algorithm that can draw maximum benefit from the limited computational resources available by allowing the least important cues to be identified and eliminated. This knowledge will also assist in the development of segregation algorithms which are able to adapt in a continuous fashion as acoustic conditions change.

9.1 The effect of source segregation

The results of the subjective intelligibility and quality measurements in Chapters 7 and 8 have produced a number of key points. The most important of these is that the speech intelligibility and quality of a target speech sound source can be improved by using a machine learning method that estimates the ratio mask for the target source in a binaural mixture.

In Chapter 6, we carried out a pilot study using a very simplistic scenario of a fixed source direction configuration in anechoic conditions. The target source was fixed at 0 azimuth and two interferers were placed at -30° and 30° azimuth, respectively. This baseline system was constructed using two fundamental localisation cues, IPD and ILD, as inputs to the ANNs which were trained to estimate the ratio mask. Based on the ratio mask estimate, the counterpart binary mask was derived by quantising the ratio mask values using a threshold of 0.5.

Building on this simple start, another system incorporating more cues was considered. The extra cues, ΔIPD, ΔILD and magnitude, were selected based on their expected potential for improving estimation of the ideal masks. Interaural coherence was included to confirm that results of subsequent relative importance analyses were plausible, in that its usefulness in anechoic conditions was expected to be small.

The comparative results revealed, as expected, that the ratio mask estimator performs better than the corresponding binary mask and that integrating extra input features always led to further improvements in segregating the target speech from the binaural mixture in terms of both the intelligibility of the target speech and its quality. Using the STOI metric, the difference in intelligibility between the ratio mask and the binary mask grew as the test SNR was decreased. For example, the ratio mask yielded STOI scores which were approximately 1.5% and 3.5% higher than the equivalent binary mask for the 5 dB and -5 dB SNR conditions, respectively. The estimator using extra input features produced a further 1% improvement compared to the baseline estimator with two inputs. Relative to the original binaural input mixture, the ratio mask estimator with six inputs is able to improve the STOI score by approximately 11%, 18% and 24% for SNRs of 5, 0 and -5 dB, respectively. Similar trends were observed regarding speech quality estimates using the PESQ metric; the ratio mask produced PESQ scores about 2% higher than those produced using the binary mask. The system using the richer set of six cues further improved the quality by approximately 1%. The best-performing system was able to improve target speech quality by over 18% for all the test conditions that we used.

In Chapter 7, to aid mask estimation in the presence of interfering sources with varying directions, and to relax the constraints of needing to know these directions explicitly, a direction cue comprising 27 cross-correlation coefficients was incorporated into the mask estimator. This resulted in a new system with seven distinct types of input. We initially assumed the existence of up to two interferers, one on either side of the listener. Both sources were constrained to lie in any azimuthal direction which was a multiple of 10°, resulting in 91 possible combinations of interferer direction. A system trained using examples drawn from only nine of the 91 combinations, for which the two interferers are paired symmetrically, is able to perform as well as a system which has been trained using the full set of 91 direction combinations. This indicates that the training data size can be reduced significantly by taking advantage of the ability of an ANN to generalise, leading to a reduction in the size of the ANN and a reduction in the training time. Reducing the training data size further led to inferior performance, demonstrating that the training data set had become too small to be able to generalise.

The fully trained seven-input, variable-direction system exhibited an improvement in speech intelligibility, compared with the original binaural mixture, of approximately 11.8%, 18.6% and 24.3% for SNRs of 5 dB, 0 dB and -5 dB, respectively. These results are very similar to the performance of the six-input system with fixed-direction interferers.

The application of OCM to target speech segregation in reverberation shows promise. STOI scores for reverberant binaural mixtures improved by up to 11.5% 17.2% and 20.1% for SNRs of 5 dB, 0 dB and -5 dB , respectively. In terms of speech quality, there were up to approximately 18% and 10% improvements for anechoic and reverberant conditions, respectively.

More challenging configurations, including a five-source setup for the OCM algorithm, were tested in Chapter 8. Our results illustrate that the OCM ratio mask estimates consistently improve target speech intelligibility and quality in both simulated binaural room impulse response (BRIR) conditions and using measured BRIRs. However, the improvement was considerably reduced, especially in the most challenging reverberation conditions compared with anechoic conditions.

A thorough analysis of the OCM algorithm demonstrates its ability to generalise across four dimensions. The OCM estimators successfully estimate masks which function satisfactorily using binaural mixtures with:

- previously unseen source directions;
- previously unseen numbers of sources;
- previously unseen signal-to-noise ratios and
- previously unseen acoustic environments.

9.2 Validation of hypothesis

The hypothesis for this research is stated in Chapter 1. It is repeated here for convenience:

The intelligibility and quality of a target speech source in a binaural mixture with spatially distinct competing concurrent interfering sounds may be increased using a machine learning algorithm which is suitable for implementation in a hearing aid.

The proposed OCM algorithm aims to improve speech intelligibility and quality by segregating the target sound from a binaural mixture based on spatial and other cues. The core of the OCM approach is based on a machine learning algorithm; a neural network with three-layer topology. The objective analysis of OCM in Chapters 6 and 7 indicates that both speech intelligibility and quality are improved for the configuration involving three competing concurrent talkers in both anechoic and reverberant conditions. Furthermore, the performance of OCM compares favourably with the leading, more complex deep neural network approach (the CLAS algorithm) described in Chapter 8. The results indicate that OCM consistently exhibits gains in intelligibility and quality and has equivalent or even better performance than the CLAS algorithm in various test conditions involving up to five competing talkers.

The algorithm operates on the two most recent frames of input signal.

Thus a delay of approximately one frame would be introduced. With a frame duration of 10 ms, the limit stated by Stone and Moore (1999) is met for avoiding issues when the hearing aid user hears their own voice. Due to the use of simple neural networks and powerful DSP technology, OCM can be implemented in real time (this is discussed in greater detail later in section 9.3.7). In summary, the hypothesis is satisfied and has been validated.

Associated with the primary hypothesis are several supplementary research questions. These have been answered by this research and will be discussed in turn.

Does optimal mapping of an increasing number of diverse cues improve the segregation of one sound source in a binaural mixture in terms of intelligibility and quality?

This research question is answered in Chapters 6 and 7. The pilot study in Chapter 6 indicates that integrating extra, appropriate input features leads to better performance in terms of speech intelligibility (using the STOI metric) and mostly improves the speech quality (using the PESQ metric). The relative importance evaluation results in Chapter 7 show that not all the cues selected contribute to the mask estimation process. Taking the interaural coherence as an example, its relative importance depends on the acoustic environment and it makes a greater contribution to mask estimation in reverberation than it does in the anechoic condition. Interaural coherence is shown to be the least important cue, out of those chosen, in the anechoic condition. Nevertheless, its inclusion in the set of inputs to the estimator does no harm to the estimators performance. The general conclusion from this aspect of the research is that an increasing number of relevant diverse cues improves the segregation performance.

Does the use of a ratio mask estimate by neural network compared with the equivalent binary mask improve segregation of one sound source in a binaural mixture in terms of intelligibility and quality?

The answer to this research question can be found in the evaluation and comparison undertaken in, Chapters 6, 7 and 8. The results consistently show that the ratio mask estimator outperforms the corresponding binary mask estimator for both the intelligibility and quality metrics, STOI and PESQ, respectively.

Is it possible to determine the varying relative contribution of diverse cues for estimating a mask in a range of simulated multiple-source and reverberant acoustic conditions?

It is clear from the discussion in section 6.2.2 that both relative importance metrics, namely Garson's method and the connection weights method, have their own strengths and weakness. So in the relative importance analyses in Chapter 6 and 7, we show the results for both approaches. They both suggest that IPD and ILD are the most important cues. IPD is dominant at low frequencies and ILD is dominant at high frequencies. The importance of cross-correlation becomes valuable when the direction of interference is variable, but does not contribute significantly to mask estimation under static conditions of source direction. The importance of interaural coherence also increases in a reverberant environment compared with an anechoic one.

Is it possible to allow maximum benefit to be drawn from limited computational resources by configuring the optimal cue mapping?

By analysing the relative contribution of diverse cues for estimating a mask, it is possible to develop an algorithm which permits maximum benefit to be drawn from the limited computational resources available in a hearing aid. The approach is scalable, with more or less inputs being incorporated according to the computational resources which can be accommodated in a particular application.

9.3 Further work

It has been shown that the optimal cue mapping approach to source segregation provides intelligibility and quality improvements in the various test cases examined in this thesis. However, it is essential to extend this work in a variety of ways, to improve system performance and allow the algorithm to operate in increasingly realistic and challenging situations.

9.3.1 Moving source segregation

The OCM system has been tested in many different configurations in Chapters 7 and 8, including the use of two, three and five sources in anechoic and reverberant conditions. In all these cases, however, the directions of the target and interfering sources have been fixed. We anticipate that the system will be tolerant of interferer movement. Using the three-source configuration as an example, the system has been tested for 91 possible interferer location combinations. For all these (fixed) directions, the system is able to segregate the target speech, demonstrating that as an interferer jumps from one location to the next, as long as the test case falls into one of these 91 combinations, the ANN is capable of producing a near-optimal mask estimate.

But how well does the algorithm work when the interferer lies in between the directions for which it was trained? The ability of the ANN to generalise for semi-matched and unmatched test cases has been demonstrated in Chapter 7 and is illustrated in figures 7.5 (b) and 7.18 (b). These figures show that the ANNs continue to work well when one or both interfering sources lie mid-way between two directions for which they were trained. This strongly suggeasts that these systems will successfully handle cases where an interferer lies in a previously unseen direction or is moving in an arbitrary way through the space for which the ANN was trained.

Throughout this work the target source direction has been fixed at an azimuth of 0°, directly in front of the listener. Although this constraint can be relaxed by training the ANNs with the target fixed in a different direction, it is always the case that the signal of interest must be known. Therefore, the segregation of a moving target signal is not supported in any of the current systems due to a lack of continually updated information about the target's position and direction of movement. In order to overcome this limitation, another layer in the algorithm will eventually be necessary to incorporate head and target tracking, so that the listener can turn their head and the algorithm's segregation focus can remain directed towards the target.

9.3.2 Integration with other features

A variety of features extracted from the binaural input mixture have been used in the OCM algorithm. The performance of the system could be improved further, however, by incorporating other features of the source and acoustic environment. Good candidates of the many possibilities include monaural cues, such as pitch-based features and the amplitude modulation spectrum (AMS) (Wang, 2015), and mel frequency cepstral coefficients (MFCC) (Kallasjoki et al., 2011; Keronen et al., 2013). How such extensions to the current inputs might be prioritised and selected is considered in the following section.

9.3.3 Relative importance measurement

Methods for measuring the relative importance for a set of input features are discussed in Chapter 6 and form an important part of this work. The reliability of the two techniques described there and applied in this research remains a problem. We apply Garson's method and the connection weights method and draw conclusions based on the combined results of both methods. More work is required to analyse more deeply inside the neural networks to derive a more robust and suitable relative importance measurement for each type of input feature.

9.3.4 Dereverberation

Speech intelligibility can be reduced by reverberation. A study carried out by Nábělek and Robinson (1982) shows that longer reverberation times decrease intelligibility more for subjects of all ages. In the current initial implementation of the OCM algorithm we did not consider dereverberation. To segregate the target speech together with its associated reverberation, the OCM estimator was trained using the reverberant target source alone. To segregate an estimate of the dereverberated target source would require the training data to be anechoic. Furthermore, a different set of cues might be needed to assist the estimator with the dereverberation process.

Given an appropriate choice of input features, selected according to their relative importance, and using a clean target source to train the estimator, there are strong reasons to suppose that the ANNs could make a well-informed estimate of a dereverberation ratio mask and so provide a further improvement in speech intelligibility compared with our current results.

9.3.5 Training target

In this research, the means of achieving binaural source segregation has been to estimate the ideal ratio mask using a set of ANNs, and this has been shown to yield improvements in speech intelligibility and quality. However, estimating the ideal ratio mask may not be the most effective method for segregating a target speech source. Other mappings can be considered for ANN training with the goal of improving target speech intelligibility and quality as well. For example, Wang et al. (2014) investigated mappings for supervised monaural speech separation. They demonstrate that the choice of a suitable training target is crucial.

It would be interesting to investigate alternative methods for segregating target speech using ANNs, such as the short-time Fourier transform spectral magnitude and the STFT mask (Wang et al., 2014). Unlike the ideal ratio mask, the STFT mask is not upperbounded by unity. In an alternative approach, Lightburn and Brookes (2015) propose an oracle binary mask. Referred to as the STOI-optimal binary mask, its goal is to maximise the intelligibility of the target speech by optimising its STOI score.

9.3.6 Artificial neural networks

In the ANN training stage of the OCM method, many parameters of the neural networks, such as the training rules and the learning rate, have to be chosen by experiment. Optimal parameter selection and the setting of their values varies depending on the particular application. There is no existing method for determining the best learning rate, for example. It is unlikely that optimal parameters have been used in our experiments. More effort is needed to determine the optimal parameters for the ANNs in this application.

The multi-layer perceptron ANN architecture used in our research falls into the category of shallow networks. Whilst the architecture works well for our current purposes, it might cope less well with a larger number of input features, which is likely to be needed in future. For example, to cope better with some realistic, but complex, scenarios, the number of simultaneous interferers which the ANNs can accommodate is likely to increase. Each additional interferer greatly increases the number of possible combinations of interferer direction. As a result, in order to retain acceptable segregation performance, the number of hidden neurons and number of hidden layers in the ANN will need to increase, along with the size of the associated training data and the training time.

In recent years, there has been much research into deep learning as applied to signal processing. Deep learning neural networks (DNNs) have already been trained to perform spectral mapping for speech dereverberation

(Han et al., 2014), source separation (Jiang et al., 2014) and automatic speech recognition (Narayanan and Wang, 2013; Maas et al., 2014). Ma et al. (2015a) exploited DNNs for binaural localisation of multiple speakers in reverberant conditions. The primary advantage of the DNN is that it is capable of compactly representing a larger set of functions than a conventional ANN (DeepLearning, 2015). For example, a k-layer network (where the number of hidden neurons is a function of the number of inputs) cannot represent as many functions as a $(k+1)$-layer network, unless the former has a very large number of hidden neurons. In addition, DNNs with multiple hidden layers show powerful learning and exhibit the capacity for nonlinear mappings that a conventional ANN cannot learn. Hence, taking advantage of the enhanced properties of DNNs may address some of the limitations of the shallow networks currently used in our research. It should be noted, however, that the conventional multi-layer perceptron ANNs used here are generally much less computationally complex than DNNs. This point is considered further in the next section.

9.3.7 Practical implementations

One of the objectives of this research is to develop a speech segregation technique capable in future of being implemented such that it operates in real time and has modest resource needs. The processing architecture and computational requirement are relatively simple in OCM and satisfy these requirements. Hence it is potentially feasible to implement the algorithm on physical hardware. Although the actual implementation of the OCM algorithm on hardware does not lie within the scope of this research, the possibility of implementing it in a real-time system has always been kept in mind. In particular, the algorithm is suitable for binaural hearing aids with a wireless link such as Bluetooth. In a real-time system, the computational task must be completed within tightly specified time limits and the algorithm must be causal, i.e. it may only use information available in the present and from the past.

Feature extraction is mainly based on the Fourier transform. The STFT temporal frame length has been set to 10 ms throughout this research, corresponding to 320 samples for a 16 kHz sample rate. Therefore, the time limit for processing each frame is 10 ms so that the reconstructed sound can be played back continuously and within the latency limit determined by Stone and Moore (1999). Current digital signal processing (DSP) technology can, for example, perform a 512-point FFT with hardware acceleration in 3740 clock cycles (equivalent to $37.4\,\mu$s with a clock rate of 100 MHz) using a 1.3 volt power supply (Mckeown, 2013). The complex operation of cross-correlation can be executed on such hardware and has been used in a commercial digital hearing aid (Widex Inc., 2015).

Once each mask estimation ANN has been trained, all its weights can be stored in a memory block in the hardware. The ratio mask estimation process inside an ANN principally involves multiplication and addition and a non-linearity in each neuron. Between the input layer and the single hidden layer, the extracted features which form the inputs to the ANN are multiplied by the appropriate weights in the input layer and the products are summed in the hidden layer neurons before being fed into the nonlinearities. A similar process applies between the hidden layer and the output layer. The final output summation is also passed through a nonlinear transfer function and the output forms an estimate for the target speech ratio mask for the same frequency point in each frame. With the pipelined multiplier and accumulator structure in a modern DSP device, the operation can be done in the order of microseconds, which is well inside the 10 ms latency limit mentioned above.

Many ultra-low-power DSPs with integrated Bluetooth are commercially available, such as the CSR series of devices (CSR, 2015), in which the smallest package option is 5.5 mm by 5.5 mm. Hence, the fundamental technology to implement the proposed algorithm in a behind-the-ear binaural wireless hearing aid is already available.

Appendices

Extra figures are presented in this chapter.

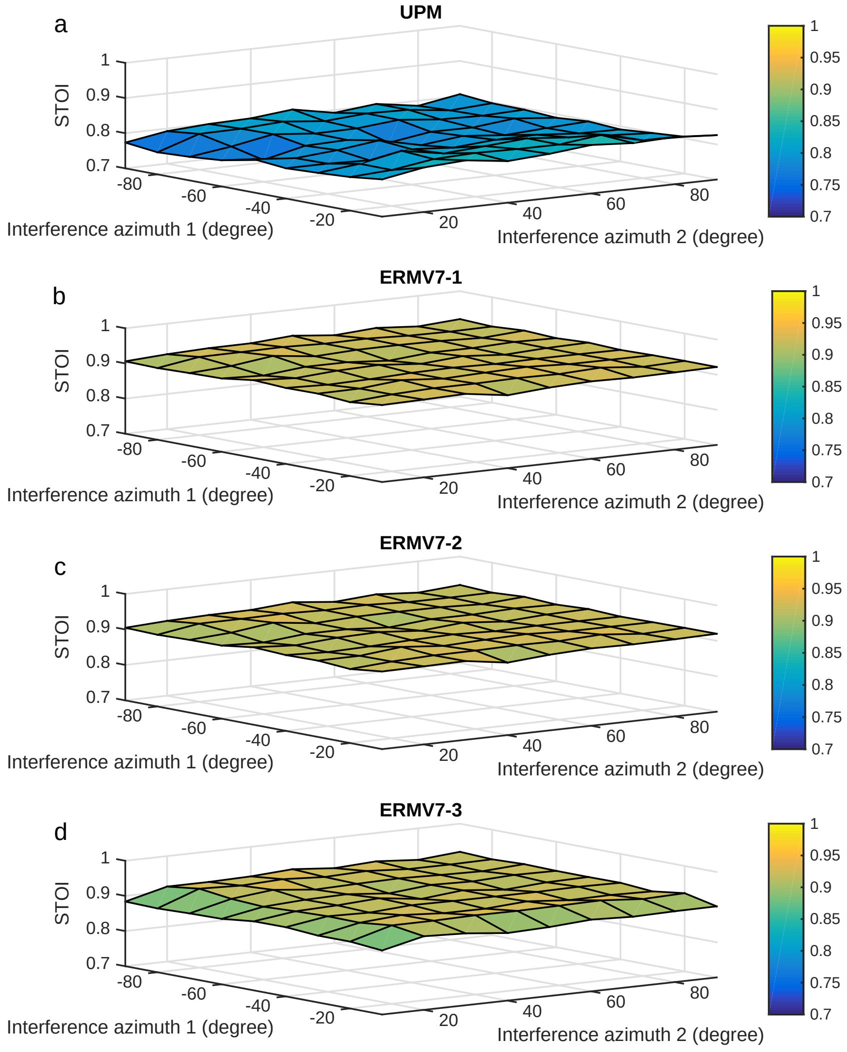


Figure A.1: STOI scores for the unprocessed mixtures and for the three mask estimators ERMV7-1/2/3, using all 91 test cases defined in TSV7-1 in the three-source anechoic configuration. The mixtures are 5 dB local SNR before spatialisation. (a) The STOI scores for the unprocessed mixture with mean value of 0.8007 and standard deviation of 0.0167. (b) Results for mask ERMV7-1 trained using strategy TSV7-1. The mean value is 0.9185 and standard deviation is 0.0055. (c) Results for mask ERMV7-2 trained using TSV7-2. The mean value is 0.9163 and standard deviation is 0.0055. (d) Results for mask ERMV7-3 trained using TSV7-3. The mean value is 0.9123 and standard deviation is 0.0111.

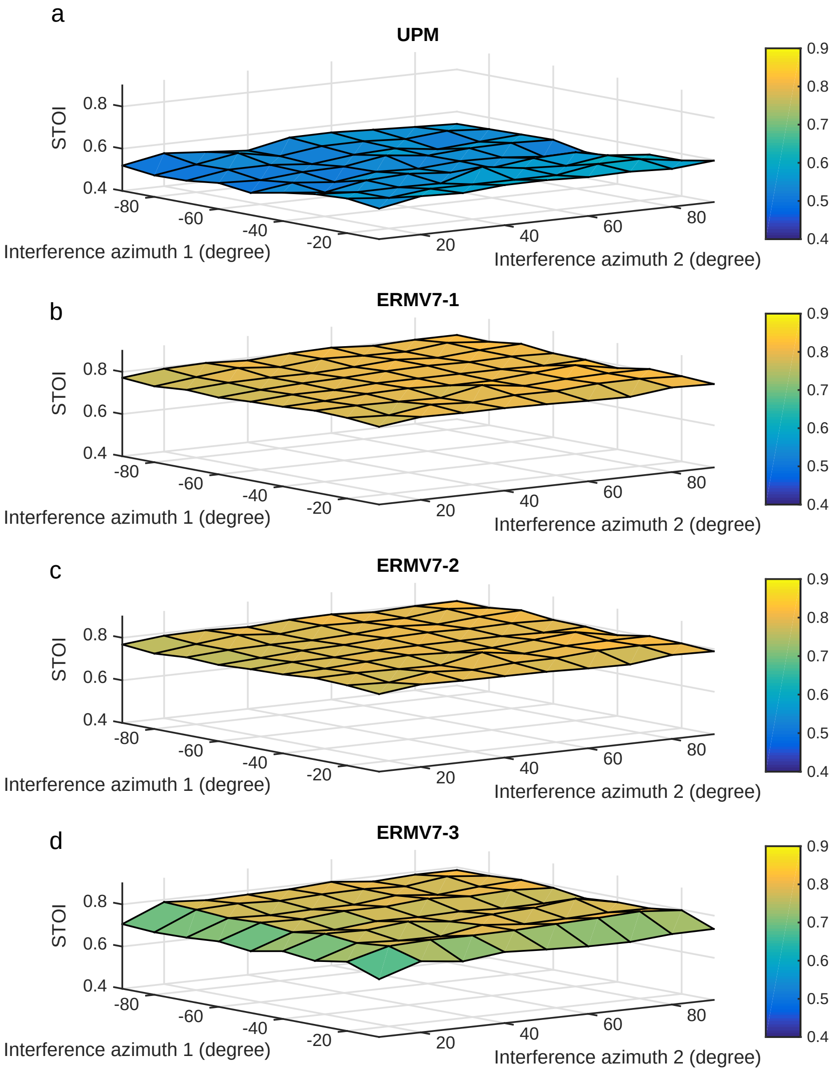


Figure A.2: STOI scores for the unprocessed mixtures and for the three mask estimators ERMV7-1/2/3, using all 91 test cases defined in TSV7-1 in the three-source anechoic configuration. The mixtures are -5 dB local SNR before spatialisation. (a) The STOI scores for the unprocessed mixture with mean value of 0.5501 and standard deviation of 0.0233. (b) Results for mask ERMV7-1 trained using strategy TSV7-1. The mean value is 0.7935 and standard deviation is 0.0123. (c) Results for mask ERMV7-2 trained using TSV7-2. The mean value is 0.7882 and standard deviation is 0.0128. (d) Results for mask ERMV7-3 trained using TSV7-3. The mean value is 0.7679 and standard deviation is 0.0311.

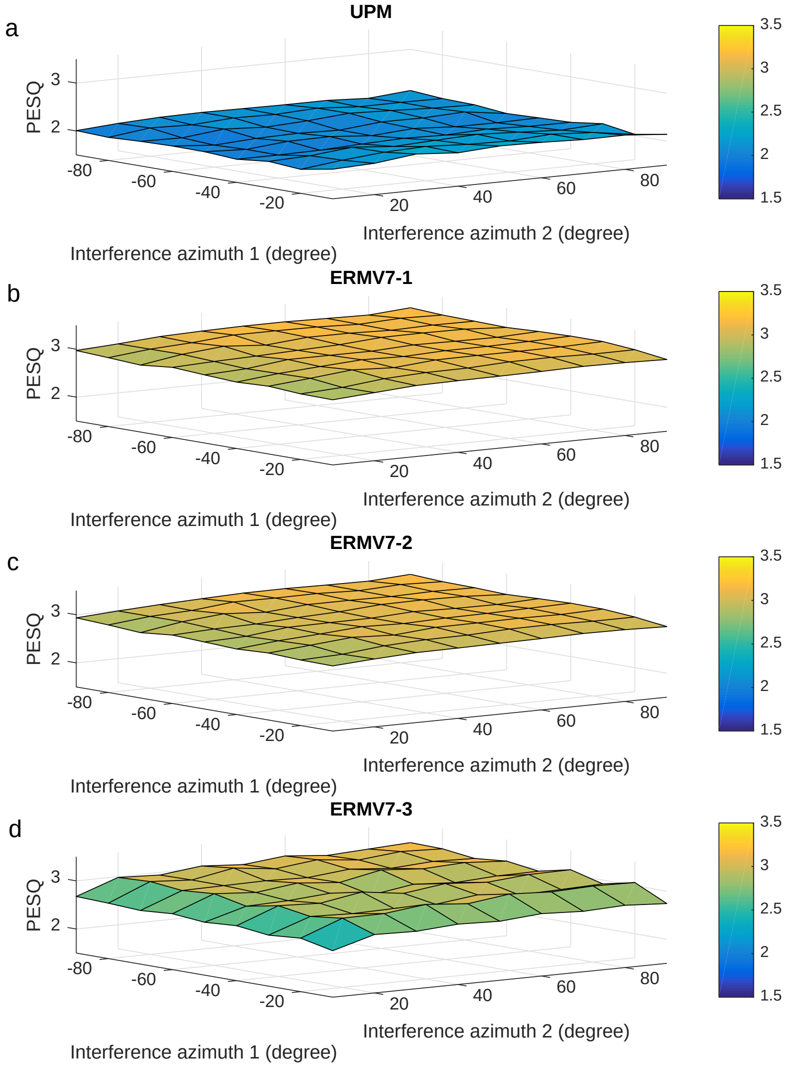


Figure A.3: PESQ scores for the three mask estimators ERMV7-1/2/3, using all 91 test cases defined in TSV7-1 in the three-source anechoic configuration. The mixtures are 5 dB local SNR before spatialisation. (a) The PESQ scores for the unprocessed mixture with mean value of 2.0914 and standard deviation of 0.0698. (b) Results for mask ERMV7-1 trained using strategy TSV7-1. The mean value is 3.0593 and standard deviation is 0.0758. (c) Results for mask ERMV7-2 trained using TSV7-2. The mean value is 3.0274 and standard deviation is 0.0820. (d) Results for mask ERMV7-3 trained using TSV7-3. The mean value is 2.9230 and standard deviation is 0.1533.

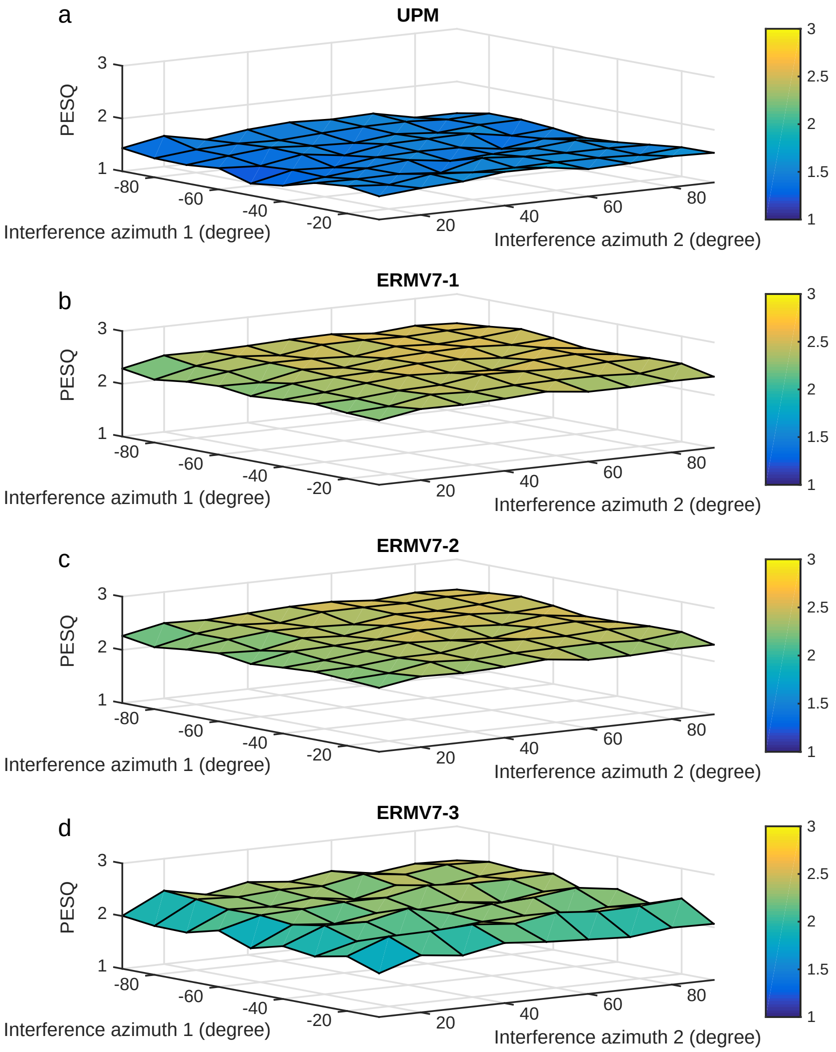


Figure A.4: PESQ scores for the three mask estimators ERMV7-1/2/3, using all 91 test cases defined in TSV7-1 in the three-source anechoic configuration. The mixtures are -5 dB local SNR before spatialisation. (a) The PESQ scores for the unprocessed mixture with mean value of 1.4750 and standard deviation of 0.0746. (b) Results for mask ERMV7-1 trained using strategy TSV7-1. The mean value is 2.4153 and standard deviation is 0.0824. (c) Results for mask ERMV7-2 trained using TSV7-2. The mean value is 2.3834 and standard deviation is 0.0820. (d) Results for mask ERMV7-3 trained using TSV7-3. The mean value is 2.2540 and standard deviation is 0.1605.

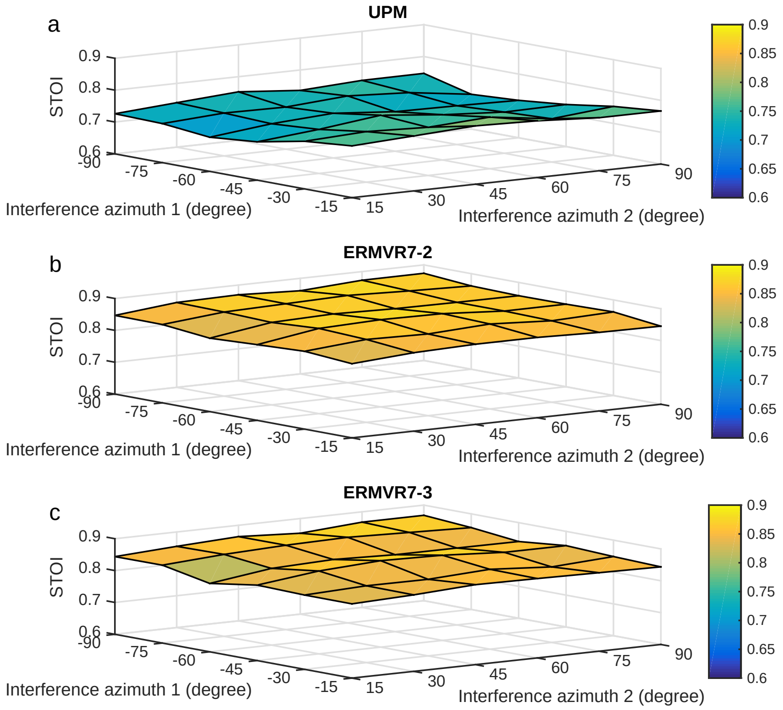


Figure A.5: STOI scores for the unprocessed mixtures and for the outputs from the two mask estimators, ERMVR7-2 and ERMVR7-3, at all direction pair combinations in the three-source reverberant configuration. The mixtures are 5 dB SNR before spatialisation. The target is located at 0°, and two interferers are placed at $-90°$ to 0° and 0° to 90°, with 15° step, respectively. (a) The STOI score of the unprocessed mixture. The mean value is 0.7425 and standard deviation is 0.0189. (b) For estimator ERMVR7-2, trained at all interferer direction pairs on the symmetric diagonal. The mean value is 0.8581 and standard deviation is 0.0115. (c) For estimator ERMVR7-3, trained at alternate combinations of interferer direction pairs on the symmetric diagonal. The mean value is 0.8498 and standard deviation is 0.0126.

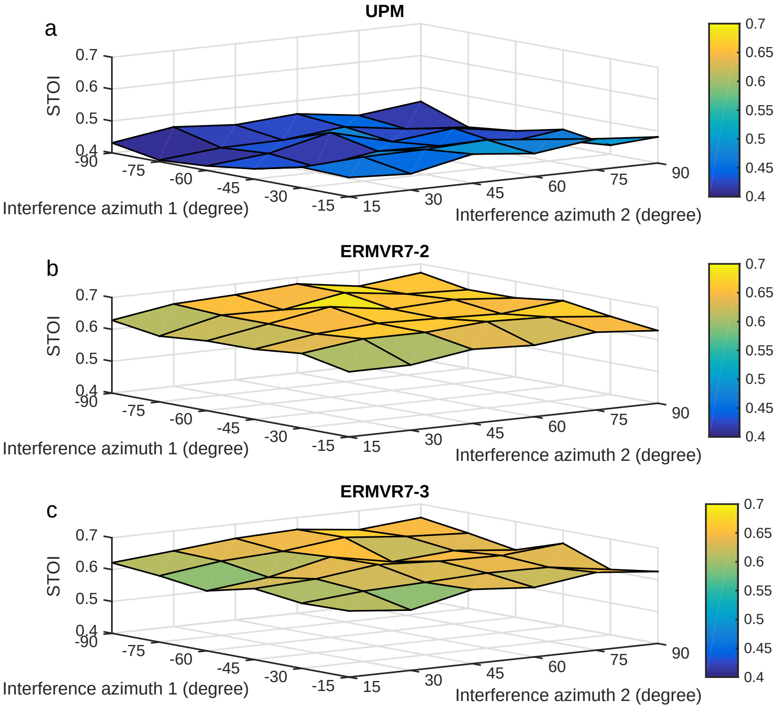


Figure A.6: STOI scores for the unprocessed mixtures and for the outputs from the two mask estimators, ERMVR7-2 and ERMVR7-3, at all direction pair combinations in the three-source reverberant configuration. The mixtures are -5 dB SNR before spatialisation. The target is located at 0°, and two interferers are placed at $-90°$ to 0° and 0° to 90°, with 15° step, respectively. (a) The STOI score of the unprocessed mixture. The mean value is 0.4470 and standard deviation is 0.0254. (b) For estimator ERMVR7-2, trained at all interferer direction pairs on the symmetric diagonal. The mean value is 0.6480 and standard deviation is 0.0209. (c) For estimator ERMVR7-3, trained at alternate combinations of interferer direction pairs on the symmetric diagonal. The mean value is 0.6311 and standard deviation is 0.0201.

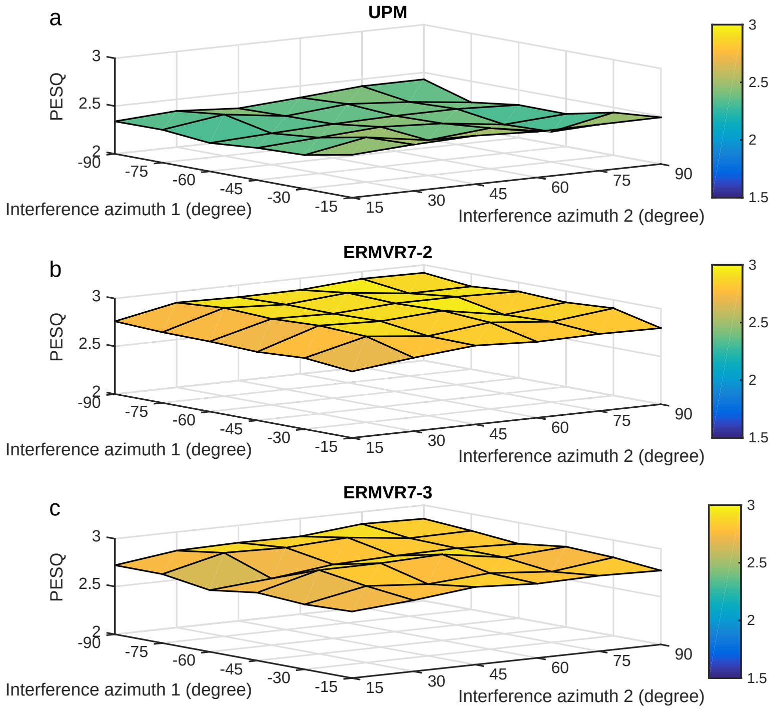


Figure A.7: PESQ scores for the unprocessed mixtures and for the outputs from the two mask estimators, ERMVR7-2 and ERMVR7-3, at all direction pair combinations in the three-source reverberant configuration. The mixtures are 5 dB SNR before spatialisation. The target is located at 0°, and two interferers are placed at $-90°$ to 0° and 0° to 90°, with 15° step, respectively. (a) The PESQ score of the unprocessed mixture. The mean value is 2.3872 and standard deviation is 0.0608. (b) For estimator ERMVR7-2, trained at all interferer direction pairs on the symmetric diagonal. The mean value is 2.8536 and standard deviation is 0.0704. (c) For estimator ERMVR7-3, trained at alternate combinations of interferer direction pairs on the symmetric diagonal. The mean value is 2.7932 and standard deviation is 0.0636.

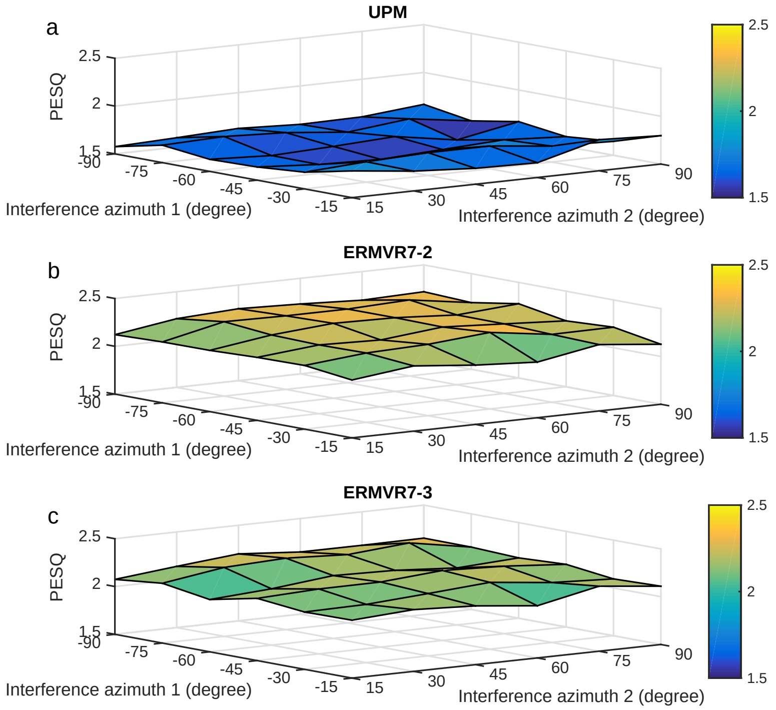


Figure A.8: PESQ scores for the unprocessed mixtures and for the outputs from the two mask estimators, ERMVR7-2 and ERMVR7-3, at all direction pair combinations in the three-source reverberant configuration. The mixtures are -5 dB SNR before spatialisation. The target is located at 0°, and two interferers are placed at $-90°$ to 0° and 0° to 90°, with 15° step, respectively. (a) The PESQ score of the unprocessed mixture. The mean value is 1.6550 and standard deviation is 0.0643. (b) For estimator ERMVR7-2, trained at all interferer direction pairs on the symmetric diagonal. The mean value is 2.2101 and standard deviation is 0.0575. (c) For estimator ERMVR7-3, trained at alternate combinations of interferer direction pairs on the symmetric diagonal. The mean value is 2.1486 and standard deviation is 0.0528.

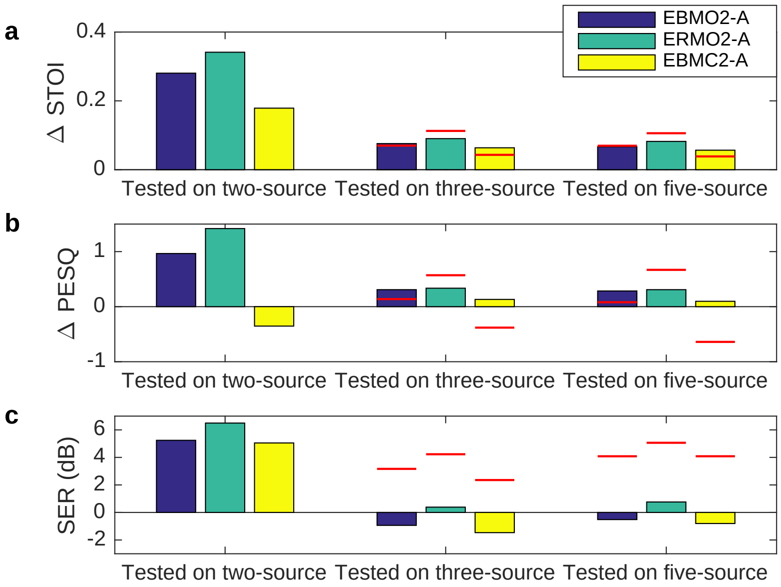


Figure A.9: The STOI and PESQ improvement compared with the corresponding unprocessed mixtures and SER performance for systems which are trained on signals containing two sources and tested on signals containing two, three and five sources, respectively. (a), (c) and (e) are the STOI improvement, PESQ improvement and SER results, respectively, using -5 dB input mixture SNR (T60 = 0 ms). The red horizontal lines show the corresponding reference results from estimators tested using the same number of sources that they were trained with.

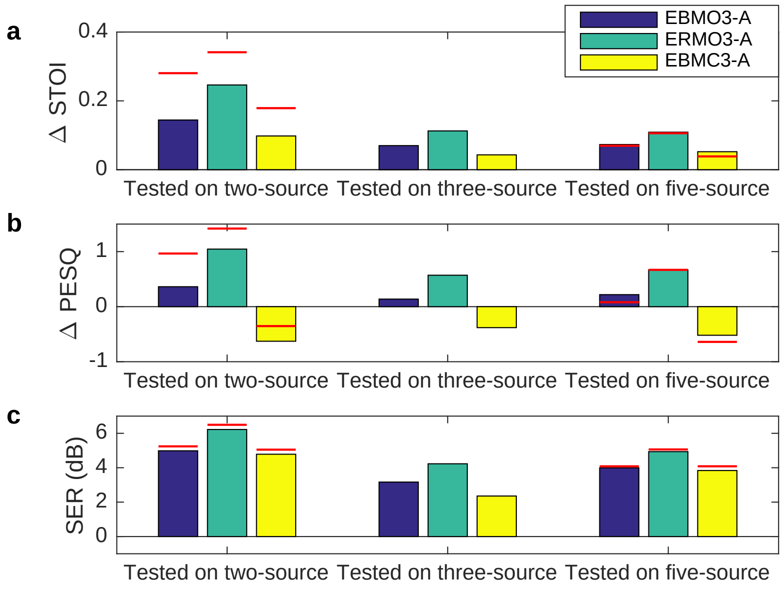


Figure A.10: The STOI and PESQ improvement compared with the corresponding unprocessed mixtures and SER performance for systems which are trained on signals containing three sources and tested on signals containing two, three and five sources, respectively. (a), (c) and (e) are the STOI improvement, PESQ improvement and SER results, respectively, using -5 dB input mixture SNR (T60 = 0 ms). The red horizontal lines show the corresponding reference results from estimators tested using the same number of sources that they were trained with.

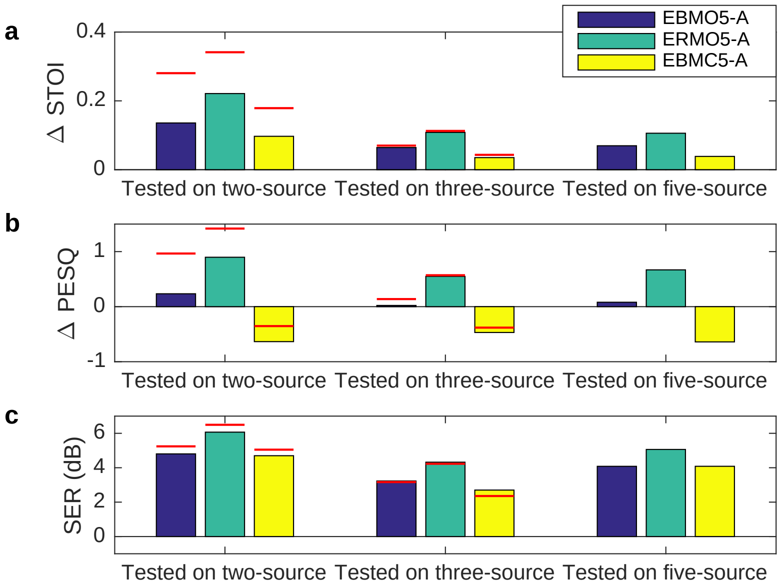


Figure A.11: The STOI and PESQ improvement compared with the corresponding unprocessed mixtures and SER performance for systems which are trained on signals containing five sources and tested on signals containing two, three and five sources, respectively. (a), (c) and (e) are the STOI improvement, PESQ improvement and SER results, respectively, using -5 dB input mixture SNR (T60 = 0 ms). The red horizontal lines show the corresponding reference results from estimators tested using the same number of sources that they were trained with.

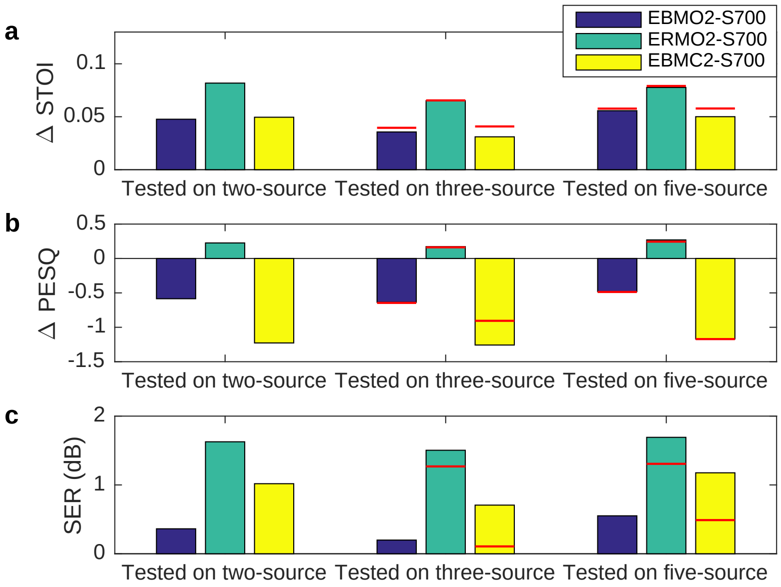


Figure A.12: The STOI and PESQ improvement compared with the corresponding unprocessed mixtures and SER performance for systems which are trained on signals containing two sources and tested on signals containing two, three and five sources, respectively. (a), (c) and (e) are the STOI improvement, PESQ improvement and SER results, respectively, using -5 dB input mixture SNR (T60 = 700 ms). The red horizontal lines show the corresponding reference results from estimators tested using the same number of sources that they were trained with.

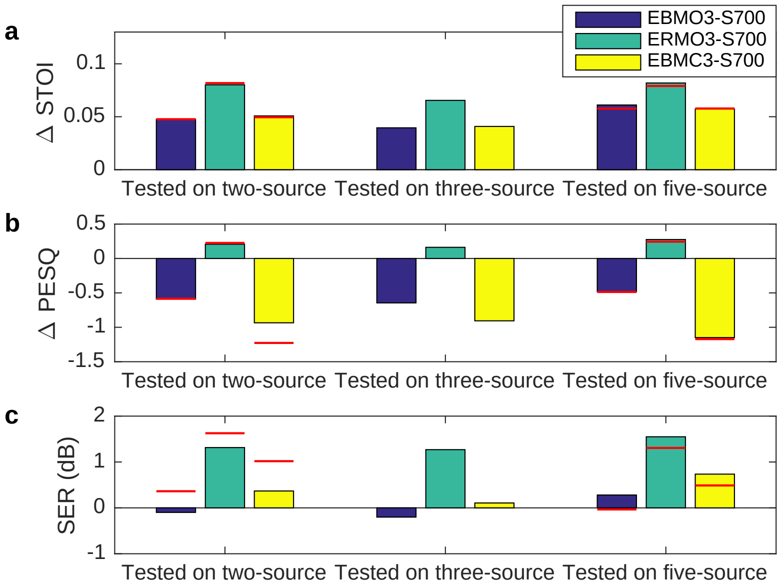


Figure A.13: The STOI and PESQ improvement compared with the corresponding unprocessed mixtures and SER performance for systems which are trained on signals containing three sources and tested on signals containing two, three and five sources, respectively. (a), (c) and (e) are the STOI improvement, PESQ improvement and SER results, respectively, using -5 dB input mixture SNR (T60 = 700 ms). The red horizontal lines show the corresponding reference results from estimators tested using the same number of sources that they were trained with.

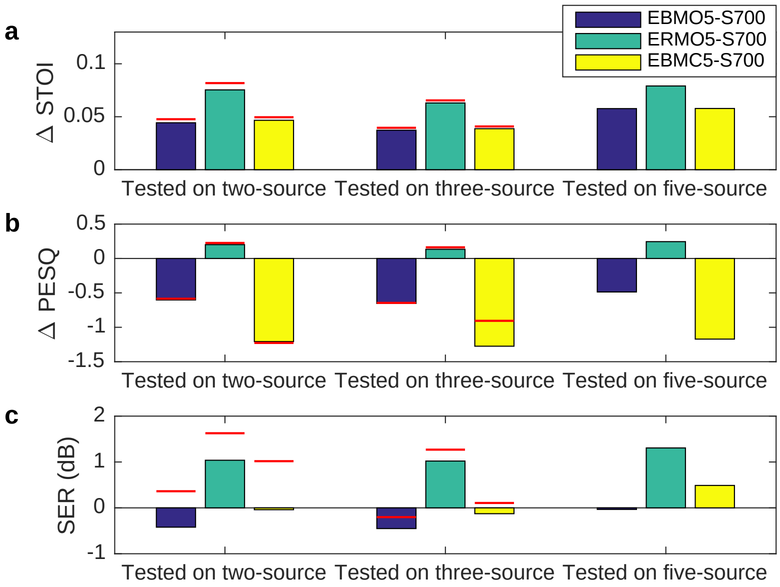


Figure A.14: The STOI and PESQ improvement compared with the corresponding unprocessed mixtures and SER performance for systems which are trained on signals containing five sources and tested on signals containing two, three and five sources, respectively. (a), (c) and (e) are the STOI improvement, PESQ improvement and SER results, respectively, using -5 dB input mixture SNR (T60 = 700 ms). The red horizontal lines show the corresponding reference results from estimators tested using the same number of sources that they were trained with.

List of Abbreviations

AI Articulation Index.

AMS Amplitude Modulation Spectrum.

ANN Articial Neural Networks.

APSD Auto-Power Spectral Density.

ASA Auditory Scene Analysis.

ASR Automatic Speech Recognition.

BILD Binaural Intelligibility Level Difference.

BM Basilar Membrane.

BMLD Binaural Masking Level Difference.

BRIR Binaural Room Impulse Response.

BSS Blind Source Separation.

BTE Behind-the-Ear.

CASA Computational Auditory Scene Analysis.

CCF Cross-Correlation Function.

CIC Completely-in-the-Canal.

CPSD Cross-Power Spectral Density.

DNN Deep Neural Network.

DRR Direct-to-Reverberant Energy Ratio.

DSP Digital Signal Processing.

DUET Degenerate Unmixing Estimation Technique.

EBM Estimated Binary Mask.

ERB Equivalent Rectangular Bandwidth.

ERM Estimated Ratio Mask.

FB Fission Boundary.

FFT Fast Fourier Transform.

FPGA Fileld Programmable Gate Array.

FS-HMM Factorial Scaled Hidden Markov Model.

GFCC Gammatone Frequency Cepstral Coefficient.

HATS Head And Torso Simulator.

HMM Hidden Markov Model.

HRIR Head-Related Impulse Response.

HRTF Head-Related Transfer Functions.

IBM Ideal Binary Mask.

IBM Ideal Ratio Mask.

IC Interaural Coherence.

ICA Independent Component Analysis.

IDFT Inverse Discrete Fourier Transform.

IED Interaural Envelope Difference.

IFFT Inverse Fast Fourier Transform.

IID Interaural Intensity Difference.

ILD Interaural Level Difference.

IPD Interaural Phase Difference.

ITD Interaural Time Difference.

ITDG Initial Time Delay Gap.

LCMV Linearly Constrained Minimum-Variance.

MAA Minimum Audible Angle.

MAF Minimum Audible Field.

MAMA Minimum Audible Movement Angle.

MAP Minimum Audible Pressure.

MESSL Model-based Expectation Maximisation Source Separation and Localisation.

MFCC Mel Frequency Cepstral Coefficients.

MOS Mean Opinion Score.

MSE Mean Square Error.

N-FHMM Non-Negative Factorial Hidden Markov Model.

NMF Non-Negative Matrix Factorisation.

OCM Optimal Cue Mapping.

PESQ Perceptual Evaluation of Speech Quality.

RI Relative Importance.

RIR Room Impulse Response.

SBR Signal-to-Background Ratio.

SER Signal-to-Error Ratio.

SII Speech Intelligibility Index.

SNR Signal-to-Noise Ratio.

SPIN Speech Perception in Noise.

SPL Sound Pressure Level.

SRT Speech Reception Threshold.

STF System Transfer Function.

STFT Short-Time Fourier Transform.

STI Speech-Transmission Index.

STOI Short-Time Objective Intelligibility.

TCB Temporal Coherence Boundary.

TSFT System Transfer Function.

UPM Unprocessed Mixture.

VAD Voice Activity Detector.

XC Cross-Correlation Coefficient.

References

Adel, H., Souad, M., Alaqeeli, A., and Hamid, A. (2012). Beamforming techniques for multichannel audio signal separation. *arXiv preprint arXiv:1212.6080*.

Akansu, A. N. and Haddad, R. A. (2001). *Multiresolution signal decomposition: transforms, subbands, and wavelets*. Academic Press.

Algazi, V. R., Avendano, C., and Duda, R. O. (2001). Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3):1110–1122.

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*, 47(11):937–947.

Alinaghi, A. (2013). Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation. *Acoustics, Speech and Signal Processing*, pages 684–688.

Allred, D. J. (2006). *Evaluation and comparison of beamforming algorithms for microphone array speech processing*. PhD thesis, Georgia Institute of Technology.

Amari, S.-i., Douglas, S. C., Cichocki, A., and Yang, H. H. (1997). Multichannel blind deconvolution and equalization using the natural gradient. In *Signal Processing Advances in Wireless Communications, First IEEE Signal Processing Workshop on*, pages 101–104. IEEE.

Anemüller, J. and Kollmeier, B. (2000). Amplitude modulation decorrelation for convolutive blind source separation. In *Proc. ICA*, pages 215–220.

ANSI, A. (1997). S3. 5-1997, methods for the calculation of the speech intelligibility index. *New York: American National Standards Institute*, 19:90–119.

Barker, J., Josifovski, L., Cooke, M., and Green, P. (2000). Soft decisions in missing data techniques for robust automatic speech recognition. *Proc. ICSLP*.

Barry, D., Coyle, E., Fitzgerald, D., and Lawlor, R. (2005). Single channel source separation using short-time independent component analysis. In *Audio Engineering Society Convention 119*. Audio Engineering Society.

Begault, D. R. et al. (1994). *3-D sound for virtual reality and multimedia*, volume 955. Citeseer.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159.

Bernstein, L. R. and Trahiotis, C. (1985). Lateralization of sinusoidally amplitude-modulated tones: Effects of spectral locus and temporal variation. *The Journal of the Acoustical Society of America*, 78(2):514–523.

Best, V., van Schaik, A., Jin, C., and Carlile, S. (2005). Auditory spatial perception with sources overlapping in frequency and time. *Acta acustica united with Acustica*, 91(3):421–428.

Bird, J. and Darwin, C. (1998). Effects of a difference in fundamental frequency in separating two sentences. *Psychophysical and physiological advances in hearing*, pages 263–269.

Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization.* MIT press.

Bluetooth Technology Website (2015). *https://www.bluetooth.com/*, (Last accessed: 27/Dec/2015).

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

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

Brebbia, C. A. and Dominguez, J. (1996). *Boundary elements: an introductory course.* WIT press.

Bregman, A. S. (1990). *Auditory scene analysis.* Cambridge, MA: MIT Press.

Bregman, A. S. and Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 32(1):19.

Brown, C. P. and Duda, R. O. (1997). An efficient hrtf model for 3-d sound. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New York: IEEE*, pages 298–301. Citeseer.

Brown, C. P. and Duda, R. O. (1998). A structural model for binaural sound synthesis. *Speech and Audio Processing, IEEE Transactions on*, 6(5):476–488.

Brown, G. J. and Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech & Language*, 8(4):297–336.

Brown, G. J. and Wang, D. (2005). Separation of speech by computational auditory scene analysis. In *Speech enhancement*, pages 371–402. Springer.

Brungart, D. S. (1998). Control of perceived distance in virtual audio displays. In *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*, volume 3, pages 1101–1104. IEEE.

Brungart, D. S. (1999). Auditory parallax effects in the hrtf for nearby sources. In *Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on*, pages 171–174. IEEE.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America*, 120(6):4007–4018.

Brungart, D. S., Durlach, N. I., and Rabinowitz, W. M. (1999). Auditory localization of nearby sources. ii. localization of a broadband source. *The Journal of the Acoustical Society of America*, 106(4):1956–1968.

Brungart, D. S. and Rabinowitz, W. M. (1999). Auditory localization of nearby sources. head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3):1465–1479.

Brungart, D. S. and Rabiowitz, W. R. (1996). Auditory localization in the near-field.

Butler, R. A. and Humanski, R. A. (1992). Localization of sound in the vertical plane with and without high-frequency spectral cues. *Perception & psychophysics*, 51(2):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*, 67(1):220–229.

Campbell, D., Palomaki, K., and Brown, G. (2005). A MATLAB Simulation of “Shoebox” Room Acoustics for use in Research and Teaching. *Computing and Information Systems.*

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418.

Chandler, D. W. and Grantham, D. W. (1992). Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity. *The Journal of the Acoustical Society of America*, 91(3):1624–1636.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979.

Coleman, P. D. (1963). An analysis of cues to auditory depth perception in free space. *Psychological Bulletin*, 60(3):302.

Commission, I. E. et al. (2003). Sound system equipment–part 16: Objective rating of speech intelligibility by speech transmission index. *International Standard IEC*, pages 60268–16.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

Cooke, M. (2005). *Modelling auditory processing and organisation*, volume 7. Cambridge University Press.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3):1562–1573.

Cooke, M. and Ellis, D. P. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech communication*, 35(3):141–177.

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

Darwin, C. J. (1997). Auditory grouping. *Trends in cognitive sciences*, 1(9):327–333.

Davis, A. (1995). *Hearing in adults: the prevalence and distribution of hearing impairment and reported hearing disability in the MRC Institute of Hearing Research's National Study of Hearing*. Whurr Publishers London.

DeepLearning (2015). Deep networks: Overview. *http://ufldl.stanford.edu/wiki/index.php/Deep_Networks:_Overview*, (Last accessed: 20/Dec/2015).

Denbigh, P. and Zhao, J. (1992). Pitch extraction and separation of overlapping speech. *Speech Communication*, 11(2):119–125.

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

Dillon, H. (2001). *Hearing aids*, volume 362. Boomerang press Sydney.

Dmour, M., Davies, M., et al. (2011). A new framework for underdetermined speech extraction using mixture of beamformers. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(3):445–457.

Douglas, S. C. and Sun, X. (2003). Convolutive blind separation of speech mixtures using the natural gradient. *Speech Communication*, 39(1):65–78.

Duda, R. O. and Martens, W. L. (1998). Range dependence of the response of a spherical head model. *The Journal of the Acoustical Society of America*, 104(5):3048–3058.

Durlach, N. I., Mason, C. R., Kidd Jr, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (2003). Note on informational masking (l). *The Journal of the Acoustical Society of America*, 113(6):2984–2987.

Egan, J. P. and Hake, H. W. (1950). On the masking pattern of a simple auditory stimulus. *The Journal of the Acoustical Society of America*, 22(5):622–630.

Ephraim, Y. and Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(2):443–445.

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

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

Fastl, H. and Zwicker, E. (2007). *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media.

Fausett, L. (1994). *Fundamentals of neural networks: architectures, algorithms, and applications.* Prentice-Hall, Inc.

Feddersen, W., Sandel, T., Teas, D., and Jeffress, L. (1957). Localization of high-frequency tones. *the Journal of the Acoustical Society of America*, 29(9):988–991.

Feng, A. S. and Jones, D. L. (2006). Localization-based grouping. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, chapter 6*, pages 187–207.

Fletcher, H. (1940). Auditory patterns. *Reviews of modern physics*, 12(1):47.

Fletcher, H. and Steinberg, J. (1929). Articulation testing methods. *Bell System Technical Journal*, 8(4):806–854.

French, N. and Steinberg, J. (1947). Factors governing the intelligibility of speech sounds. *The journal of the Acoustical society of America*, 19(1):90–119.

Frost III, O. L. (1972). An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935.

Gao, J. and Tew, A. I. (2015). The segregation of spatialised speech in interference by optimal mapping of diverse cues. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 2095–2099. IEEE.

Garson, D. G. (1991). Interpreting neural network connection weights. *AI Expert*, pages 47–51.

Gimsing, S. (2008). [use of hearing aids five years after issue]. *Ugeskrift for laeger*, 170(43):3407–3411.

Glasberg, B. R. and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1):103–138.

Gold, B., Morgan, N., and Ellis, D. (2011). *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons.

Gomez, A. M., Schwerin, B., and Paliwal, K. K. (2011). Objective intelligibility prediction of speech by combining correlation and distortion based techniques. In *INTERSPEECH*, pages 1225–1228.

Grant, K. W. and Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. In *AVSP 2001-International Conference on Auditory-Visual Speech Processing*.

Grantham, D. W. (1986). Detection and discrimination of simulated motion of auditory targets in the horizontal plane. *The Journal of the Acoustical Society of America*, 79(6):1939–1949.

Graupe, D., Grosspietsch, J., and Basseas, S. (1986). A single-microphone-based self-adaptive filter of noise from speech and its performance evaluation. *Journal of rehabilitation research and development*, 24(4):119–126.

Greenberg, J. E. and Zurek, P. M. (1992). Evaluation of an adaptive beamforming method for hearing aids. *The Journal of the Acoustical Society of America*, 91(3):1662–1676.

Greenwood, D. D. (1961). Auditory masking and the critical band. *The journal of the acoustical society of America*, 33(4):484–502.

Griffiths, L. J. and Jim, C. W. (1982). An alternative approach to linearly constrained adaptive beamforming. *Antennas and Propagation, IEEE Transactions on*, 30(1):27–34.

Guillon, P. and Zolfaghari, R. (2012). Creating the sydney york morphological and acoustic recordings of ears database. *Multimedia and Expo (ICME), 2012 IEEE International Conference on. IEEE*, 16(1):37–46.

Han, K., Wang, Y., and Wang, D. (2014). Learning spectral mapping for speech dereverberation. *Proc. ICASSP, to appear*, 23(6):982–992.

Harding, S., Barker, J., and Brown, G. (2006). Mask estimation for missing data speech recognition based on statistics of binaural interaction. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):58–67.

Harris, J. D. (1965). Monaural and binaural speech intelligibility and the stereophonic effect based upon temporal cues. *The Laryngoscope*, 75(3):428–446.

Harris, J. D. and Sergeant, R. L. (1971). Monaural/binaural minimum audible angles for a moving sound source. *Journal of Speech, Language, and Hearing Research*, 14(3):618–629.

Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (1999). Speech intelligibility and localization in a multi-source environment. *The Journal of the Acoustical Society of America*, 105(6):3436–3448.

Hearing Aids (2015). Concealed hearing devices of the 20th century. *http://beckerexhibits.wustl.edu/did/20thcent/part6.htm*, (Last accessed: 27/Dec/2015).

Henning, G. B. (1974). Detectability of interaural delay in high-frequency complex waveforms. *The Journal of the Acoustical Society of America*, 55(1):84–90.

Hershey, J. R., Rennie, S. J., Olsen, P. A., and Kristjansson, T. T. (2010). Super-human multi-talker speech recognition: A graphical modeling approach. *Computer Speech & Language*, 24(1):45–66.

Hirsh, I. J. (1950). The relation between localization and intelligibility. *The Journal of the Acoustical Society of America*, 22(2):196–200.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*.

Howard, D. M. and Angus, J. (2009). *Acoustics and psychoacoustics*. Taylor & Francis.

Hu, G. and Wang, D. (2004). Monaural speech segregation based on pitch tracking and amplitude modulation. *Neural Networks, IEEE Transactions on*, 15(5):1135–1150.

Hu, G. and Wang, D. (2008). Segregation of unvoiced speech from nonspeech interference. *The Journal of the Acoustical Society of America*, 124(2):1306–1319.

Humanski, R. A. and Butler, R. A. (1988). The contribution of the near and far ear toward localization of sound in the sagittal plane. *The Journal of the Acoustical Society of America*, 83(6):2300–2310.

Hummersone, C. (2010). Dynamic precedence effect modeling for source separation in reverberant environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1867–1871.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430.

Ihlefeld, A. and Shinn-Cunningham, B. (2008). Spatial release from energetic and informational masking in a selective speech identification task. *The Journal of the Acoustical Society of America*, 123(6):4369–4379.

Ikram, M. Z. and Morgan, D. R. (2002). A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, pages I–881. IEEE.

Irie, B. and Miyake, S. (1988). Capabilities of three-layered perceptrons. *Neural Networks, 1988., IEEE International Conference on. IEEE*, pages 641–648.

Jeffress, L. A. (1948). A place theory of sound localization. *Journal of comparative and physiological psychology*, 41(1):35.

Jeub, M., Schäfer, M., Esch, T., and Vary, P. (2010a). Model-based dereverberation preserving binaural cues. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7):1732–1745.

Jeub, M., Schäfer, M., and Krüger, H. (2010b). Do we need dereverberation for hand-held telephony? *Proc. Int. Congress on Acoustics*, (August):1–7.

Jeub, M., Schäfer, M., and Vary, P. (2009). A binaural room impulse response database for the evaluation of dereverberation algorithms. pages 1–5.

Jiang, Y., Wang, D., Liu, R., and Feng, Z. (2014). Binaural classification for reverberant speech segregation using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):2112–2121.

Jin, C. T., Guillon, P., Epain, N., Zolfaghari, R., van Schaik, A., Tew, A., Hetherington, C., Thorpe, J., et al. (2014). Creating the sydney york morphological and acoustic recordings of ears database. *Multimedia, IEEE Transactions on*, 16(1):37–46.

John, G., Lori, L., William, F. J. F., David, P., Nancy, D., and Victor, Z. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. *Web Download. Philadelphia: Linguistic Data Consortium.*

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. *The Journal of the Acoustical Society of America*, 61(5):1337–1351.

Kallasjoki, H., Keronen, S., Brown, G. J., Gemmeke, J. F., Remes, U., and Palomäki, K. J. (2011). Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments. *Proc. CHiME*, pages 58–63.

Kasturi, K., Loizou, P. C., Dorman, M., and Spahr, T. (2002). The intelligibility of speech with "holes" in the spectrum. *The Journal of the Acoustical Society of America*, 112(3):1102–1111.

Kates, J. M. and Weiss, M. R. (1996). A comparison of hearing-aid array-processing techniques. *The Journal of the Acoustical Society of America*, 99(5):3138–3148.

Kendrick, P. and Shirley, B. (2008). Performance of independent component analysis when used to separate competing acoustic sources in anechoic and reverberant conditions. In *Audio Engineering Society Convention 124*. Audio Engineering Society.

Keronen, S., Kallasjoki, H., Remes, U., Brown, G. J., Gemmeke, J. F., and Palomäki, K. J. (2013). Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment. *Computer Speech & Language*, 27(3):798–819.

Kidd Jr., G., 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. *The Journal of the Acoustical Society of America*, 104(1):422–431.

Killion, M. C. (1978). Revised estimate of minimum audible pressure: Where is the "missing 6 db"? *The Journal of the Acoustical Society of America*, 63(5):1501–1508.

Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (2009). An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 126(3):1486–94.

Kochkin, S. (2000). MarkeTrak V : " Why my hearing aids are in the drawer ": The consumers ' perspective. *The Hearing Journal*, 53(2):34–41.

Kochkin, S. (2010). Marketrak viii: Consumer satisfaction with hearing aids is slowly increasing. *The Hearing Journal*, 63(1):19–20.

Koffka, K. (2013). *Principles of Gestalt psychology*, volume 44. Routledge.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145.

Köhler, W. (1970). *Gestalt psychology: An introduction to new concepts in modern psychology*. WW Norton & Company.

Kollmeier, B. and Koch, R. (1994). Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. volume 95, pages 1593–1602. Acoustical Society of America.

Kondo, K. (2012). *Subjective quality measurement of speech: its evaluation, estimation and applications*. Springer Science & Business Media.

Kubovy, M. (1981). Concurrent-pitch segregation and the theory of indispensable attributes. *Perceptual organization*, pages 55–98.

Kuhn, G. (1977). Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America*, 62(May 2012):157–167.

Kumaresan, R. and Rao, A. (1999). Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications. *The Journal of the Acoustical Society of America*, 105(3):1912–1924.

Li, N. and Loizou, P. C. (2007). Factors influencing glimpsing of speech in noise. *The Journal of the Acoustical Society of America*, 122(2):1165–1172.

Li, N. and Loizou, P. C. (2008). Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *The Journal of the Acoustical Society of America*, 123(3):1673–1682.

Licklider, J. C. R. (1951). A duplex theory of pitch perception. *The Journal of the Acoustical Society of America*, 23(1):147–147.

Lightburn, L. and Brookes, M. (2015). Sobm-a binary mask for noisy speech that optimises an objective intelligibility metric. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5078–5082. IEEE.

Little, A. D., Mershon, D. H., and Cox, P. H. (1992). Spectral content as a cue to perceived auditory distance. *Perception*, 21(3):405–416.

Lyon, R. F. (1983). A computational model of binaural localization and separation. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, volume 8, pages 1148–1151. IEEE.

Ma, N., Brown, G. J., and May, T. (2015a). Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions. In *Sixteenth Annual Conference of the International Speech Communication Association.*

Ma, N., May, T., Wierstorf, H., and Brown, G. (2015b). A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions. In *40th IEEE International Conference on Acoustics, Speech and Signal Processing.*

Maas, A. L., Hannun, A. Y., Lengerich, C. T., Qi, P., Jurafsky, D., and Ng, A. Y. (2014). Increasing Deep Neural Network Acoustic Model Size for Large Vocabulary Continuous Speech Recognition.

Mandel, M., Weiss, R., and Ellis, D. (2010a). Model-based expectation-maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, pages 1–13.

Mandel, M. I., Weiss, R. J., and Ellis, D. P. (2010b). Model-based expectation-maximization source separation and localization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(2):382–394.

May, T., Ma, N., and Brown, G. J. (2015). Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 2679–2683. IEEE.

Middlebrooks, J. C. and Green, D. M. (1990). Directional dependence of interaural envelope delays. *The Journal of the Acoustical Society of America*, 87(5):2149–2162.

Middlebrooks, J. C., Makous, J. C., and Green, D. M. (1989). Directional sensitivity of sound-pressure levels in the human ear canal. *The Journal of the Acoustical Society of America*, 86(1):89–108.

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin*, 44(2):105–129.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352.

Mills, A. and Tobias, J. V. (1972). *Foundations of modern auditory theory*. Academic Press, New York.

Mills, A. W. (1958). On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4):237–246.

Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. (1995). Head-related transfer functions of human subjects. *Journal of the Audio Engineering Society*, 43(5):300–321.

Moncur, J. P. and Dirks, D. (1967). Binaural and monaural speech intelligibility in reverberation. *Journal of Speech, Language, and Hearing Research*, 10(2):186–195.

Moore, B. C. (2012). *An introduction to the psychology of hearing*. Brill.

Musicant, A. D. and Butler, R. A. (1984). The influence of pinnae-based spectral cues on sound localization. *The Journal of the Acoustical Society of America*, 75(4):1195–1200.

Mysore, G. J., Smaragdis, P., and Raj, B. (2010). Non-negative hidden markov modeling of audio with application to source separation. In *Latent Variable Analysis and Signal Separation*, pages 140–148. Springer.

Nábělek, A. K., Letowski, T. R., and Tucker, F. M. (1989). Reverberant overlap-and self-masking in consonant identification. *The Journal of the Acoustical Society of America*, 86(4):1259–1265.

Nábělek, A. K. and Pickett, J. (1974). Reception of consonants in a classroom as affected by monaural and binaural listening, noise, reverberation, and hearing aids. *The Journal of the Acoustical Society of America*, 56(2):628–639.

Nábělek, A. K. and Robinson, P. K. (1982). Monaural and binaural speech perception in reverberation for listeners of various ages. *The Journal of the Acoustical Society of America*, 71(5):1242–1248.

Nakatani, T., Goto, M., and Okuno, H. G. (1996). Localization by harmonic structure and its application to harmonic sound stream segregation. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference*

Proceedings., 1996 IEEE International Conference on, volume 2, pages 653–656. IEEE.

Narayanan, A. and Wang, D. (2013). Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7092–7096. IEEE.

Narayanan, A. and Wang, D. (2014). Investigation of speech separation as a front-end for noise robust speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(4):826–835.

Okuno, H. G., Nakatani, T., and Kawabata, T. (1996). Interfacing sound stream segregation to automatic speech recognition-preliminary results on listening to several sounds simultaneously. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1082–1089.

Okuno, H. G., Nakatani, T., and Kawabata, T. (1999). Listening to two simultaneous speeches. *Speech communication*, 27(3):299–310.

Olden, J. D., Joy, M. K., and Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3):389–397.

Oticon (2008). he audiology in epoq – a whitepaper. *Oticon whitepaper*.

Ozerov, A., Févotte, C., and Charbit, M. (2009). Factorial scaled hidden markov model for polyphonic audio representation and source separation. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pages 121–124. IEEE.

Parsons, T. W. (1976). Separation of speech from interfering speech by means of harmonic selection. *The Journal of the Acoustical Society of America*, 60(4):911–918.

Patterson, R., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1987). An efficient auditory filterbank based on the gammatone function. In *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, volume 2.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *The Journal of the Acoustical Society of America*, 59(3):640–654.

Perrott, D. R. and Musicant, A. (1977). Minimum auditory movement angle: Binaural localization of moving sound sources. *The Journal of the Acoustical Society of America*, 62(6):1463–1466.

Perrott, D. R. and Saberi, K. (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America*, 87(4):1728–1731.

Plomp, R. (1976). Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise). *Acta Acustica united with Acustica*, 34(4):200–211.

Plomp, R. (1977). Acoustical aspects of cocktail parties. *Acta Acustica united with Acustica*, 38(3):186–191.

Plutowski, M., Sakata, S., and White, H. (1994). Cross-validation estimates imse. *training (as training can be faster on smaller datasets)*, 2:4.

Pollack, I. and Pickett, J. (1958). Stereophonic listening and speech intelligibility against voice babble. *The Journal of the Acoustical Society of America*, 30(2):131–133.

Pralong, D. and Carlile, S. (1994). Measuring the human head-related transfer functions: A novel method for the construction and calibration of a miniature "in-ear"recording system. *The Journal of the Acoustical Society of America*, 95(6):3435–3444.

Ramírez, J. and Górriz, J. M. (2011). *Recent Advances in Robust Speech Recognition Technology.* Number p60. Bentham Science.

Rayleigh, L. (1907). Xii. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232.

Rickard, S. (2007). The DUET Blind Source Separation algorithm. *Blind Speech Separation*, pages 217–237.

Ricketts, T. A. and Hornsby, B. W. (2003). Distance and reverberation effects on directional benefit. *Ear and Hearing*, 24(6):472–484.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 2, pages 749–752. IEEE.

RNID (2007). Annual survey report 2007. *The Royal National Institute for Deaf People, Registered charity numbers 207720 (England and Wales) and SC038926 (Scotland).*

RNID (2015). Annual survey report 2015. *The Royal National Institute for Deaf People, Registered charity numbers 207720 (England and Wales) and SC038926 (Scotland).*

Robinson, D. W. and Dadson, R. S. (1956). A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, 7(5):166.

Roman, N., Srinivasan, S., and Wang, D. (2006). Binaural segregation in multisource reverberant environments. *The Journal of the Acoustical Society of America*, 120(6):4040.

Roman, N., Wang, D., and Brown, G. J. (2003). Speech segregation based on sound localization. *The Journal of the Acoustical Society of America*, 114(4):2236–2252.

Roman, N. and Woodruff, J. (2011). Intelligibility of reverberant noisy speech with ideal binary masking. *The Journal of the Acoustical Society of America*, 130(4):2153–2161.

Rosenthal, D. F. and Okuno, H. G. (1998). *Computational auditory scene analysis.* Lawrence Erlbaum Associates Publishers.

Rossing, T. (1990). *Science of Sound*. Addison-Wesley.

Roweis, S. T. (2000). One microphone source separation. In *NIPS*, volume 13, pages 793–799.

Sawada, H., Mukai, R., Araki, S., and Makino, S. (2004). A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *Speech and Audio Processing, IEEE Transactions on*, 12(5):530–538.

Schum, D. J. (2003). Noise-reduction circuitry in hearing aids:(2) goals and current strategies. *The Hearing Journal*, 56(6):32–33.

Schwerin, B. and Paliwal, K. (2014). An improved speech transmission index for intelligibility prediction. *Speech Communication*, 65:9–19.

Shinn-Cunningham, B. (2000). Distance cues for virtual auditory space. In *Proceedings of the IEEE-PCM*, volume 2000, pages 227–230.

Shinn-Cunningham, B. G. and Best, V. (2008). Selective attention in normal and impaired hearing. *Trends in Amplification*.

SID, V. Y. F. T. (2005). *Blind Audio Source Separation*. University of Cambridge.

Smaragdis, P. (1998). Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1):21–34.

Smith, A. (2015). York advanced research computing cluster. *https://wiki.york.ac.uk/display/RHPC/YARCC+-+York+Advanced+Research+Computing+Cluster*, (Last accessed: 12/Aug/2015).

Smith, L. S. and Fraser, D. S. (2004). Robust sound onset detection using leaky integrate-and-fire neurons with depressing synapses. *Neural Networks, IEEE Transactions on*, 15(5):1125–1134.

Srinivasan, S., Roman, N., and Wang, D. (2006). Binary and ratio time-frequency masks for robust speech recognition. *Speech Communication*, 48(11):1486–1501.

Steeneken, H. J. M. (1992). *On measuring and predicting speech intelligibility*. PhD thesis, TU Delft, Delft University of Technology.

Stone, M. A. and Moore, B. C. (1999). Tolerable hearing aid delays. i. estimation of limits imposed by the auditory path alone using simulated hearing losses. *Ear and Hearing*, 20(3):182.

Strange, W., Jenkins, J. J., and Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *The Journal of the Acoustical Society of America*, 74(3):695–705.

Strutt, J. W. (1907). On our perception of sound direction. *Philosophical Magazine*, 13:214–232.

Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136.

Taal, C. H., Hendriks, R. C., Heusdens, R., Jensen, J., and Kjems, U. (2009). An evaluation of objective quality measures for speech intelligibility prediction. In *INTERSPEECH*, pages 1947–1950.

Tao, Y., Tew, A. I., and Porter, S. J. (2003). The differential pressure synthesis method for efficient acoustic pressure estimation. *Journal of the Audio Engineering Society*, 51(7/8):647–656.

Tellakula, A. K. (2007). Acoustic source localization using time delay estimation. *Degree Thesis. Bangalore, India: Supercomputer Education and Research Centre Indian Institute of Science.*

Thakur, C. S., Wang, R. M., Afshar, S., Hamilton, T. J., Tapson, J. C., Shamma, S. A., and van Schaik, A. (2015). Sound stream segregation: a neuromorphic approach to solve the cocktail party problem in real-time. *Frontiers in neuroscience*, 9.

Van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences.* Institute for Perceptual Research.

Varga, A. (1990). Hidden markov model decomposition of speech and noise. In *Proc. ICASSP 90*, pages 845–848.

Varga, A. and Steeneken, H. J. (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251.

Wang, D., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (2009). Speech intelligibility in background noise with ideal binary time-frequency masking. *The Journal of the Acoustical Society of America*, 125(4):2336–47.

Wang, D. L. and Brown, G. J. (2006). Computational auditory scene analysis: Principles, algorithms and applications. *Wiley IEEE press.*

Wang, Y. (2015). *Supervised speech separation using deep neural networks.* Computer science and engineering, The Ohio State University.

Wang, Y., Narayanan, A., and Wang, D. (2014). On training targets for supervised speech separation. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(12):1849–1858.

Warren, R. M. et al. (1970). Perceptual restoration of missing speech sounds. *Science*, 167(3917):392–393.

Weiss, S. and Kulikowski, C. (1991). Computer systems that learn.

Westermann, A., Buchholz, J. M., and Dau, T. (2013). Binaural dereverberation based on interaural coherence histograms. *The Journal of the Acoustical Society of America*, 133(5):2767–77.

Widex Inc. (2015). *http://www.widex.com/*, (Last accessed: 21/Sep/2015).

Wightman, F. L. and Kistler, D. J. (1989). Headphone simulation of free-field listening. i: Stimulus synthesis. *The Journal of the Acoustical Society of America*, 85(2):858–867.

Wilson, R. H., McArdle, R., Watts, K. L., and Smith, S. L. (2012). The revised speech perception in noise test (r-spin) in a multiple signal-to-noise ratio paradigm. *Journal of the American Academy of Audiology*, 23(8):590–605.

Woodruff, J. (2012). *Integrating monaural and binaural cues for sound localization and segregation in reverberant environments.* PhD thesis, The Ohio State University.

Woodruff, J. (2013). Binaural Detection, Localization, and Segregation in Reverberant Environments Based on Joint Pitch and Azimuth Cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(4):806–815.

Woodruff, J. and Wang, D. (2013). Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(4):806–815.

Woods, W. S., Hansen, M., Wittkop, T., and Kollmeier, B. (1996). A simple architecture for using multiple cues in sound separation. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 2, pages 909–912. IEEE.

Woodworth, R. S. and Schlosberg, H. (1962). *Experimental psychology.* Holt, Rinehart and Winston.

Yi, H. and Loizou, P. C. (2008). Evaluation of Objective Quality Measures for Speech Enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):229–238.

Zwicker, E. and Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society of America*, 68(5):1523–1525.