Adaptive 3D sound systems


DOI:
10.6100/IR525351

Document status and date:
Published: 01/01/1999

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

Take down policy
If you believe that this document breaches copyright please contact us: openaccess@tue.nl providing details. We will immediately remove access to the work pending the investigation of your claim.
Adaptive 3D Sound Systems

John Garas
Adaptive 3D Sound Systems

PROEFONTWERP

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de Rector Magnificus, prof.dr. M. Rem, voor een commissie aangewezen door het College voor Promoties in het openbaar te verdedigen op woensdag 15 september 1999 om 16.00 uur

door

John Garas

goingen te Souhag, Egypte
Preface

Stereo sound systems are able to position the perceived sound image at any point on the line between the two reproduction loudspeakers. On the contrary, three-dimensional (3D) sound systems are designed to position the sound image at any point in a 3D listening space around the listener. Processing a monophonic sound signal by filters containing directional cues moves the perceived sound image to the position defined by the filters. The perceived sound source at this position is, therefore, referred to as a virtual or phantom sound source. It is the objective of the present work to investigate the application of adaptive filters in creating robust virtual sound images in real-time through loudspeakers. Adaptive filters are attractive in such applications for two main reasons. The first is their tracking capability. Adaptive filters may be used to track moving listeners and adjust their coefficients accordingly. The second is allowing in-situ design of the filters, which in turn allows including the listeners' own head-related transfer functions in the filters design.

To the time of this writing, most work in 3D sound systems has been performed based on anechoic listening situations, which simplifies the analysis considerably. However, this often leads to systems that function properly only in anechoic environments. To avoid such conditional results in the present work, listening spaces are assumed to be reverberant rooms. The acoustic impulse responses measured in those rooms are understood to have long duration due to reverberation. Although this may frustrate the real-time implementation, it permits studying the effects of the listening environment on the system performance. It also motivates searching for the most efficient implementation of the system filters.

Stereo and traditional 3D sound systems are often designed to properly operate for a single listener who is assumed to be seated at an ideal position in the listening space. In the present work, adaptive 3D sound systems are introduced in the general context of multichannel sound reproduction. No assumptions are made concerning the number of listeners, nor their positions in the reverberant listening space. Although this choice complicates the analysis, it results in a general framework that is valid for many applications in the general listening situation of multiple listeners listening to multiple audio signals through multiple loudspeakers.

The main goal of 3D sound systems is to convince their users (listeners) that the sound is emitted from a position in the listening space where no physical loudspeaker exists. To successfully achieve this goal, not only physical parameters but also human perception must be taken into account in designing such system. Since the output of
a 3D sound system is the perceived sound image location, system performance must be tested based on well-designed psychoacoustical experiments. Although simulations have been used throughout this dissertation to show performance improvement, psychoacoustical experiments are inevitable in proving the significance of the methods developed. Those experiments have not been performed in the present work. However, I hope that psychoacousticians will find those methods worth the experiments in the near future.

I would like to express my gratitude towards all of those who contributed directly or indirectly to the success of this work. To name a few: prof.dr.ir. P.P.J. van den Bosch, prof.dr.ir. W.M.G. van Bokhoven, dr.ir. P.C.W. Sommen, dr. R.M. Aarts, and Stan Ackermans institute for giving me the opportunity to carry out this research in the first place. Prof.dr. A. Kohlrausch for the valuable discussions related to psychoacoustics and his fine remarks that made the text more accurate and the sentences more understandable. Dr.ir. A.C. den Brinker and B.E. Sarroukh for the many discussions, comments, and suggestions related to unitary transformations. Dr.ir. A.M. Tuckey for reading the first concept of this thesis and correcting many horrible language mistakes. Prof. P.A. Nelson for the discussions and hospitality he showed during my visit to ISVR in autumn 1998. Dr. U.K. Laine and many other researchers around the world for the long e-mail discussions and their valuable suggestions especially related to time and frequency warping. Ir. D.W.E. Schobben, ir. A.J. van Leest, dr.ir. M.J. Bastiaans, ing. A.C.P. van Meer, and all other members of the signal processing group for being understanding colleagues.

Summary

3D sound systems achieve their objectives by filtering a set of monophonic audio signals by a matrix of digital filters. The solution to those filters is, in general, composed of two components. The first is the cross-talk cancellation network, which inverts the matrix of acoustic transfer functions between the reproduction loudspeakers and the listeners' ears. The second is the matrix of desired transfer functions, which contains the directional information the listeners should perceive. Traditional 3D sound systems make use of fixed filters to implement both components. The coefficients of the filters are derived from pre-measured transfer functions, often called the Head-Related Transfer Functions (HRTFs). HRTFs vary significantly among people due to the differences in size and geometrical fine structure of human outer ears. Since it is almost impossible to measure and store the HRTFs for each listener, the fixed filters are usually calculated using generalised HRTFs. These generalised HRTFs are usually obtained by measuring the HRTFs of an artificial head or by averaging the HRTFs measured for several human subjects. Unless the generalised HRTFs match those of the listener, calculating the filters using generalised HRTFs may degrade the perceived sound image.

The above mentioned fixed filters are valid only for a single listening position, namely, the position at which the HRTFs used to calculate the filters have been previously measured. If the listener moves slightly from this ideal position, the filters become in error. Using head trackers, traditional 3D sound systems are able to switch between different pre-designed filters to correct the errors introduced by listener's movements. This requires measuring and storing HRTFs from several discrete directions around the listener.

An alternative approach to 3D sound systems is to replace the fixed filters by adaptive ones. This does not require the above mentioned measurements and operates exclusively using the listeners' own HRTFs. By adjusting the filters' coefficients during system operation, errors introduced due to listeners' movements are automatically corrected. For successful adaptation, a measure of the error criteria must be available. This measure may be obtained by inserting probe microphones inside (or near the entrance of) the listeners' ear canals.

In reverberant environments, the impulse response between two points may last for several hundreds of milliseconds. At 44.1 kHz sampling frequency (the standard for audio compact discs), thousands of FIR filter coefficients are needed to properly model such an impulse response. Therefore, adaptive 3D sound systems require a
huge amount of computations, which makes real-time implementation on reasonable hardware resources a difficult task. Reducing the system complexity is, therefore, essential. Implementing the filtering and adaptation operations in the frequency domain reduces the computational complexity considerably. In multichannel systems, further computational saving may be achieved by replacing the traditional filtered-x LMS algorithm by its efficient counterpart, the adjoint LMS algorithm.

For successful tracking of listeners' movements by on-line adaptation of the adaptive filters, the convergence speed of the adaptive algorithm needs to be faster than the listeners' movements. Both filtered-x and adjoint LMS algorithms are known to be slow due to the coloration introduced by filtering through the HRTFs. Convergence speed of those algorithms may be improved by reducing or totally removing this coloration.

System robustness to small listeners' movements may be improved by enlarging the zones of equalisation created around the listeners' ears. Increasing the number of reproduction loudspeakers and correctly positioning those loudspeakers in the listening space improve the system robustness considerably. The filters' behaviour may also be controlled by applying spatial derivative constraints. Alternatively, filters having decreasing frequency resolution with increasing frequency may be employed for the same purpose. Replacing the constant resolution filters with multiresolution ones is also motivated by the fact that the human auditory system performs a similar multiresolution spectral analysis on sound signals.

Multiresolution adaptive filters may be obtained if the filters are implemented in a transformation domain which has the required frequency resolution. To maintain the low computational complexity achieved by implementing the filters in the frequency domain, the multiresolution transformation must be calculated as fast as the Fast Fourier Transformation (FFT). An efficient implementation of the unitary warping transformation using non-uniform sampling is developed in this dissertation for this purpose. By non-uniformly sampling a segment of a signal and calculating the traditional FFT of those samples, a multiresolution transform of that segment is obtained.
Chapter 1

Introduction

Typical applications of 3D sound systems complement, modify, or replace sound attributes that occur in natural listening situations to gain control on one's spatial auditory perception. This control can not be achieved based on physical attributes only. Psychoacoustic considerations of human sound localisation also play important roles in analysing, designing, and testing 3D sound systems. To manipulate one's spatial auditory perception, a thorough understanding of the psychoacoustical phenomena occurring in natural spatial hearing is essential. By modifying the physical parameters associated with these phenomena, the control goal may be achieved. This chapter introduces the problem of designing 3D sound systems, the challenges faced, and the approach followed in the present work. Natural and virtual spatial hearing and the synthesis of localisation cues using conventional digital filters are introduced in Section 1.1. Implementation using adaptive filters, the challenges and advantages in using such filters are mentioned in Section 1.2. In the present work, special attention has been paid to studying the robustness problem, which is inherent to all 3D sound systems. The notion of this problem and the approach followed to contain it are mentioned in Section 1.3. In Section 1.4, example applications of 3D sound systems are considered. Finally, the conventions used throughout the text are listed in Section 1.5.

1.1 Analysis and Synthesis of Spatial Hearing

1.1.1 Natural Spatial Hearing

In natural listening situations, human listeners are capable of estimating the position of the objects radiating the sound waves. The precision of estimating the position of sound sources located in the horizontal plane containing the listener's ears is fairly accurate to allow relying on spatial hearing in everyday life. A typical example often used to explain this sound localisation ability is when one crosses the street looking to the right. Correctly estimating the position and speed of a car coming from the left, from its sound only, helps avoiding an accident.
Chapter 1: Introduction

Using psychoacoustic experiments [30], the important cues used by the human auditory system to localise sound sources in space are identified. It is widely believed that localisation of sources in the horizontal plane containing the listener’s ears is due to the differences between the sound waves received at each ear. Localisation of sound sources not in the horizontal plane is strongly influenced by the acoustic filtering due to reflection of short wavelength sound waves from the listener’s outer ears and shoulders. Those and other localisation cues are discussed briefly in Chapter 2. The effect of room reverberation on natural spatial listening is also addressed in Chapter 2.

1.1.2 Virtual Spatial Hearing

All physical parameters associated with the localisation cues that occur in natural listening situations are embedded in the pair of acoustic transmissions from the source to each of the listener’s eardrums. Assume that those acoustic transmissions, usually referred to as the Head-Related Transfer Function (HRTF) pair, can be accurately measured and stored in the form of digital filters. Filtering a monophonic sound signal that contains no directional information through the previously measured HRTF pair results in a pair of electric signals that contain all the information concerning the source location. Playing this pair of electric signals (e.g. through headphones) at the listener’s eardrums reproduces the localisation cues, which makes the listener perceive the sound to be coming from the position where the source was at the measurement time. This forms the basic principle of modern 3D sound systems [25, 54, 72, 143]. The perceived sound source in this listening situation is referred to as a virtual (or phantom) source. The physical properties of HRTF pairs and the synthesis of virtual sound sources are discussed in more details in Chapter 2.

1.1.3 Conventional 3D Sound Systems

To complete the control task, the pair of electric signals containing the directional information must be transformed to acoustic waves prior playing them at the listener’s eardrums. This may be accomplished using headphones or a pair of loudspeakers. Headphones allow delivering the electric signals at the listener’s ears with minimum cross-talk. Although this simplifies the reproduction process considerably, loudspeakers are preferred in most applications. When loudspeakers are used, the sound from each loudspeaker is heard by both ears. The 3D sound system must, therefore, employ a mechanism to eliminate or reduce this cross-talk [50, 79, 104]. This is usually accomplished by employing a second set of filters often called the cross-talk canceller. Therefore, a 3D sound system that delivers its output sound through loudspeakers employs control filters that are the combination of the cross-talk canceller filters and the HRTF pair containing the directional information. The basic principles of designing the filters employed in conventional (opposed to adaptive) 3D sound systems are discussed in Section 2.2.

1.2 Adaptive 3D Sound Systems

Although several types of fixed digital filters have been used successfully in 3D sound systems [72, 84, 85, 86], adaptive filters may be preferred in many applications. Unlike conventional fixed filters, adaptive filters allow in-situ calculation of the control filters. This enables recalculating the filters in response to changes in the listening environment by simply readapting the filters. In Chapter 3, adaptive 3D sound systems are discussed in detail in the context of multichannel sound reproduction.

The adaptation of the system filters in multichannel 3D sound systems is complicated by the acoustic transmissions between the loudspeakers and the listeners’ ears. Due to these acoustic transmissions, the conventional Least Mean Square (LMS) algorithm becomes unstable and can not be used. A related algorithm, the filtered-x LMS algorithm, has to be employed. The convergence properties of this algorithm are strongly influenced by the physical properties of the acoustic transmissions. Furthermore, adaptive multichannel 3D sound systems require a huge amount of computations to perform real-time filtering and adaptation operations. Efficient implementation of those operations is, therefore, essential. Reducing the computational complexity of multichannel systems may be achieved by using a more efficient version of the filtered-x LMS algorithm, the adjoint LMS. Further computational saving may be gained by performing the filtering and adaptation operations in the frequency domain. The optimum LMS solution to the control filters, the effects of the acoustic transmissions, the filtered-x LMS, the adjoint LMS, and the frequency domain implementations are treated in greater detail in Chapter 3.

1.3 Robustness of 3D Sound Systems

Since the basic principle of 3D sound systems mentioned above relies on the directional information embedded in the HRTF pairs, it is obvious that the control filters depend on the position of the sound source relative to the listener, independent of the method used to calculate those filters. A set of filters designed for optimum results at a specific listener position becomes in error when the listener moves. This problem, referred to as the robustness problem hereafter, is inherent to all 3D sound systems. The robustness problem and its remedies are investigated in Chapter 4.

Small listener movements are contained by designing the control filters to be valid not only at the eardrums of the listener, but also at points in space in the vicinity of the listener’s ears. This is achieved by employing spatial derivative constraints, which limit the sound field to be the same at and around the listener’s ears. The effect of these spatial derivative constraints is to force the filters to assume increasingly deviating solutions from the optimum solutions as the frequency increases. Implementations of spatial derivative constraints require measuring several transfer functions in the vicinity of the ears and calculating the control filters using a weighted average of those measurements.
A more efficient approach to forcing the filters to the above mentioned solutions that increasingly deviate from the optimum solutions with increasing frequency may be achieved by using control filters of non-uniform frequency resolution. Filters having coarser frequency resolution with increasing frequency increasingly ignore details in the acoustic transmissions with increasing frequency. This is also in full agreement with the multiresolution analysis performed by the human ears [148]. Multiresolution spectral analysis of a segment of a signal may be calculated by sampling the signal at non-uniformly spaced time moments and calculating the FFT of those samples. This non-uniform sampling method and other techniques for designing filters having such non-uniform frequency resolution are discussed in Chapter 5.

The performance of a sound reproduction system is strongly influenced by the number of reproduction loudspeakers and the positions of those loudspeakers in the listening space. Using computer simulations, it is shown in Chapter 4 that better performance can be obtained by increasing the number of reproduction loudspeakers. Examination of the optimum positions of the loudspeakers is also considered in anechoic and reverberant listening environments.

When a listener makes a large movement, the control filters become in great errors, and new filters that are derived from the new listener's position must be used. In adaptive 3D sound systems, the required filters in the new position may be calculated by readapting the filters to their new optimal solutions. This requires measuring the acoustic transmissions between the loudspeakers and the listener's eardrums. For successful operation, the adaptive filters must be updated faster than the listener's movements. Therefore, improving the convergence speed of the adaptive algorithm is essential. On-line estimation of acoustic transmissions, on-line adaptation of the control filters, and improving convergence speed of the adaptive algorithm are treated at length in Chapter 4.

### 1.4 Applications

3D sound systems are important in many applications such as consumer electronics, interactive entertainment, desktop computers, multimedia, and human-machine interfaces. In this section, examples of those applications are mentioned.

A simple application of 3D sound systems is the stereo-base widening [1]. The physical dimensions of a television set or a computer monitor may limit the distance between the built-in stereo loudspeakers. The quality of the stereo effect is known to degrade as the distance between the two loudspeakers decreases. Virtual sound sources may be used to let the pilot perceive the sound of different signals to be coming from different directions. Therefore, decreasing the probability of misunderstanding the messages.

Virtual sources are also essential in game machines and virtual reality systems. The experience of a car race game is closer to reality when sounds travel with the cars compared to the case where sounds always come from the same fixed place of the system loudspeaker.

An example of improving human-machine interface using 3D sound systems may be found in the presentation of signals a pilot may receive in his cockpit [25]. Virtual sources may be used to let the pilot perceive the sound of different signals to be coming from different directions. Therefore, decreasing the probability of misunderstanding the messages.

### 1.5 Conventions

Throughout the text, lower-case letters are used to represent time domain signals while upper-case letters represent frequency domain signals. Boldface characters are used to represent matrices and underlined boldface characters represent vectors. In single channel contexts, matrix and vector notations are used to represent samples of a signal's time history or samples of the Fourier transform of the signal. For example, the last $B$ samples from the time history of the signal $x_i(n)$ is represented by

$$X_i(n) = [x_i(n) \ x_i(n-1) \ \ldots \ x_i(n-B+1)]^T,$$  \hspace{1cm} (1.1)

where $^T$ denotes transpose. Calculating the $B$-point Discrete Fourier Transformation (DFT) of $x_i(n)$ results in the vector of frequency samples $X_i(\omega)$

$$X_i(\omega) = [X_i(\omega_0) \ X_i(\omega_1) \ \ldots \ X_i(\omega_B)]^T,$$  \hspace{1cm} (1.2)

where $\omega_b = \frac{2\pi b}{B}$, $b = 0, 1, \ldots, B - 1$. In multichannel contexts, matrix and vector notations will be used to represent multiple signals at a single time moment or a single frequency. For example, the set of signals $\{x_i(n) : i = 1, 2, \ldots, I\}$ at a time sample $n$ is represented by the vector

$$X(n) = [x_1(n) \ x_2(n) \ \ldots \ x_I(n)]^T.$$  \hspace{1cm} (1.3)

The time history of each signal $x_i(n)$ is expressed by (1.1). Calculating the $B$-point DFT of the time history of each of the $x_i(n)$ signals and evaluating the result at one specific frequency $\omega$ results in the frequency vector $X(\omega)$

$$X(\omega) = [X_1(\omega) \ X_2(\omega) \ \ldots \ X_I(\omega)]^T.$$  \hspace{1cm} (1.4)

may not be convenient. In such situations, virtual sound sources can be generated to replace physical loudspeakers. Since two real loudspeakers can be used to generate $N$ virtual sound sources (see Chapter 2), surround sound can be generated using two front loudspeakers only. This leads to an efficient use of the hardware in applications such as televisions, home theatre, and audio equipment.
Chapter 1: Introduction

In some cases as in Chapter 3, the time histories of multiple signals are needed to be processed simultaneously. In these cases, composite vectors and matrices that stack time samples of all signals in specific structures will be defined.

Many concepts are readily understood using discrete-time representations while others are best explained using continuous-time representations of signals. While the discrete-time representations are considered to be the default throughout the text, the continuous-time representations are used whenever necessary. This will be often the case in Chapter 5, which deals with non-uniform sampling and multiresolution spectral analysis.

Frequent use will be made of the Fourier transformation. Therefore, the Fourier operator will be denoted by the special symbol \( \mathbf{F} \). In the discrete-time domain, \( \mathbf{F} \) is a square matrix which has its \((n,m)\) element defined by 
\[ e^{-j \frac{2\pi}{B} nm}, \]
where \( j = \sqrt{-1} \) and \( B \) is the transformation length. The same symbol will be used for the Fourier transformation in the continuous-time domain and will be defined when needed. Other operators that will be frequently used are the convolution operator denoted by \(*\) and the complex conjugate operator denoted by \(*\).

Both the Cartesian and vertical-polar coordinates will be used throughout the text to express the position of objects in the listening space as shown in Fig. 1.1. In single listener situations, the coordinate system is chosen so that the origin is at the centre point between the listener’s ears (the interaural axis). The +z-axis passes through the listener’s right ear, while the +y-axis passes through the listener’s nose. The xy, yz, and xz planes are referred to as the horizontal, median, and frontal planes, respectively. In the vertical-polar coordinate system, the azimuth angle \( \alpha \) is measured from the median plane to a vertical plane containing the z-axis and the object as shown in Fig. 1.1. The elevation angle \( \theta \) is then the angle from the horizontal plane to the object on that plane.

Figure 1.1: The vertical-polar coordinate system.

Chapter 2

3D Sound Reproduction

Psychoacousticians distinguish between a sound source location and an auditory event location. The former is the position of a physical sound source in the listening space, while the latter is the position where a listener experiences the sound [30]. From everyday experience we know that a monophonic audio signal played through a loudspeaker makes the sound source and the auditory event positions coincide. However, it is possible to process the audio signal so that the auditory event occurs at a completely different position in the listening space than the position of the physical loudspeaker emitting the sound. The listener perceives the sound to be coming from the auditory event position, which is therefore, referred to as a phantom (or virtual) sound source. A simple form of this audio processing is the stereophonic audio system [6], where the amplitude or phase of the sound is panned between two loudspeakers. Stereophonic systems are able to position the auditory event (sound image) at any point on the line connecting the two loudspeakers. A direct extension to this technique is the surround sound system, where more than two loudspeakers surrounding the listener are used. By panning the sound between every two adjacent loudspeakers, the auditory event can be positioned on lines connecting the loudspeakers [121]. As the number of reproduction loudspeakers increases, the auditory event can be accurately placed at any point in a three-dimensional (3D) space. This is exploited in the wave field synthesis technique that is based on the Kirchhoff-Helmholtz integral [27, 31]. In the present work, we are more interested in positioning the auditory event at any point in a virtual 3D space using a pair of loudspeakers arranged in a stereophonic set-up. This is achieved by audio systems based on Head-Related Transfer Functions (HRTF). HRTF-based systems are also able of creating multiple virtual sound images simultaneously at different positions in the same listening space using two loudspeakers only. This chapter introduces the basic psychoacoustical, physical, and signal processing principles behind virtual sound source systems of this type. The chapter is divided into two parts: an analysis section (Section 2.1) and a synthesis section (Section 2.2).

In Section 2.1, psychoacoustical and physical phenomena related to natural spatial hearing are introduced with emphasis on the factors that are exploited in improving the robustness of 3D sound systems in Chapter 4. The basic problem of virtual sound source synthesis using HRTFs is examined in Section 2.2.
2.1 Sound Localisation and Localisation Cues

In natural hearing, the human auditory system makes use of several cues to locate an auditory event in the listening space. The difference between sound signals at the left and right ears (interaural cues) are used to determine the horizontal component (azimuth angle) of the sound source position. Spectral cues resulting from high frequency reflections inside the listener's outer ears (pinnae) help determining the vertical component (elevation angle). Human listeners are also capable of estimating the sound source distance in addition to azimuth and elevation angles. Interaural, spectral, and distance cues are discussed in Sections 2.1.1, 2.1.2, and 2.1.3, respectively. When the above mentioned cues are ambiguous, humans usually move their heads (some animals move their ears), so that more (or less) interaural and spectral cues are introduced, making the localisation task easier. These dynamic cues are mentioned in Section 2.1.4. Room reverberation degrades human localisation performance, since the direction of a reflected sound may be confused with the direction of the sound coming directly from the source. However, thanks to the precedence effect, human listeners are still able to localise sounds in reverberant environments. The precedence effect and the physical and perceptual aspects of room reverberation are discussed in Sections 2.1.5 and 2.1.6, respectively. Most of the above mentioned cues are embedded in the physical transfer functions between a sound source and the eardrums of the listener. The properties of such transfer functions, referred to as the Head-Related Transfer Functions (HRTFs) are mentioned briefly in Section 2.1.7. Section 2.1.8 concludes the analysis part of this chapter with a brief review of the frequency selectivity of the human ears, a property that will be exploited in improving the robustness of 3D sound systems in Chapter 4.

2.1.1 Interaural Cues

In natural listening situations, the difference between sound waves received at the listener’s left and right ears is an important cue used by the human auditory system to estimate the sound source position in space. These difference cues, referred to as the interaural or binaural cues are best explained using the far field anechoic listening situation shown in Fig. 2.1. A sound signal emitted from a sound source S located in the horizontal plane at azimuth angle α and distance r from the centre of the listener’s head travels to the listener’s right (ipsilateral) and left (contralateral) ears through paths SR and SL, respectively. Since SR in this example is shorter than SL, a tonal sound wave reaches the right ear before the left ear. The arrival time difference is referred to as the Interaural Time Difference (ITD). The ITD is zero when the source is at azimuth zero, and is a maximum at azimuth ±90°. This represents a difference of arrival time of about 0.7 ms for a typical-size human head, and is readily perceived [30]. ITD represents a powerful and dominating cue at frequencies below about 1.5 kHz. At higher frequencies, the ITD represents ambiguous cue since it corresponds to a shift of many cycles of the incident sound wave. For complex sound waves, the ITD of the envelope at high frequencies, which is referred to as the Interaural Envelope Difference (IED), is perceived.

On the other hand, the head forms an obstacle for incident sound waves. This leads to a level difference between the two ears, known as the Interaural Intensity Difference (IID). Besides being dependent on azimuth angle, the IID is highly dependent on the frequency of the incident sound. At low frequencies, the wavelength is larger than the listener’s head and the sound wave is diffracted around the head to reach the contralateral ear without noticeable attenuation. As the frequency increases, the head forms a bigger obstacle for the sound wave and the level at the contralateral ear decreases. This is known as the head-shadow effect. The IID is an effective cue in the frequency range above 1.5 kHz, therefore, forming a complementary cue to the ITD. Together, the ITD and IID cover the whole audible frequency range.

Assuming a spherical head without any external ears, a sound source at azimuth angle α and its image about the interaural axis at azimuth 180° − α, as shown in Fig. 2.1, produce the same ITD and IID cues at the listener’s ears. For a real head, the ITD and IID are never equal for the two positions, they are, however, maximally similar. With the absence of other spatial cues than the ITD and IID, this similarity causes potential confusion, which explains the often reported phenomenon of front-back reversals [25, 30]. This phenomenon is a special case, in the horizontal plane, of the general phenomenon of the cones of confusion. The cones of confusions are best explained using the vertical-polar coordinates mentioned in Section 1.5. In this
Chapter 2: 3D Sound Reproduction

The spectral cues discussed in the next section enable resolving those confusions.

2.1.2 Spectral Cues

The primary cues used by the human auditory system for elevation estimation are interaural or binaural cues used for azimuth estimation. Spectral cues stem from reflections of short wavelength sound waves off the listener's upper body (torso) and off the outer ears (pinnae). The pinna with its irregular shape and resonant cavities acts as an acoustical filter. Sound waves reflected off the pinna interfere with the direct sound entering the ear canal constructively at some frequencies, and destructively at other frequencies as shown in Fig. 2.2-a. This leads to spectral peaks at frequencies where constructive interference occurs, and spectral dips at those frequencies where destructive interference takes place. The frequencies at which spectral peaks and dips appear, as well as the magnitude of these features, are dependent on the direction of the sound source. The frequencies at which spectral dips occur are of special interest, since spectral dips are often more localised in frequency than spectral peaks. The first spectral dip, known as the pinna notch is believed to be the major cue for elevation localisation. The frequency at which the pinna notch appears changes from about 6 to 12 kHz as the elevation angle changes from -40° to +60° [72] (see also Section 2.1.7). This is shown in Fig. 2.2-b for the transfer functions measured for a dummy head with the median plane at elevation angles -10°, 0°, and 10°.

Figure 2.2: [a] A schematic diagram of the high frequency reflections off the pinna causing constructive and destructive interferences with the direct sound wave. [b] measured pinna responses for a dummy head with the source in the median plane at elevation angles -10°, 0°, and 10°.

In ambiguous natural listening situations where interaural and spectral cues produce insufficient information for the auditory system to localise the sound source, humans turn their heads to resolve this ambiguity. This often occurs when ambiguous interaural cues are introduced at a listener's ears due to the cone of confusion phenomenon. From familiarity with their own pinnae response, human listeners are able to use the spectral cues to estimate the sound source position. Since spectral cues are mainly due to high frequency reflections, slight changes in the pinna shape may lead to significant changes in its frequency response. Therefore, spectral cues vary significantly among people due to differences in pinnae sizes and fine geometrical structure.

2.1.3 Distance Cues

Many phenomena have been noticed to influence the estimation of the distance of a sound source by the human auditory system. Loudness and direct-to-reverberant energy ratio are believed to be the most effective in influencing distance perception. Loudness cues stem from the fact that the sound pressure in the far field radiated from a sound source decreases with increasing the distance from the source. Therefore, nearby sound sources are perceived louder than distant sources emitting the same acoustic energy. The ratio of the sound intensity of two sources at distances \( r_1 \) and \( r_2 \) from a listener's ear is given by \( I_1/I_2 = r_2^2/r_1^2 \), which is known as the inverse square law [25]. Thus, a distance doubling decreases the sound intensity at the listener's ear by 6 dB. The human auditory system makes use of this fact in estimating the distance of a sound source. However, familiarity with the sound and its characteristics influence the estimated distance based on loudness [25]. Furthermore, the loudness cue is valid only in anechoic environments, since in a reverberant environment the sound distribution is dependent on the reverberation characteristics of the enclosed space. For instance, in a reverberant room, the sound field beyond the reverberation distance may be considered diffuse, and theoretically independent of the distance from the source (see Section 2.1.6). This explains the importance of direct-to-reverberant energy ratio as a distance cue [24, 30]. A recent study of distance perception [33] shows that a model based on a modified direct-to-reverberant energy ratio can accurately predict the source distance. The modification to the true direct-to-reverberant energy ratio is the use of an integration time window of 6 milliseconds in the calculation of the direct sound energy. Because of its importance in distance and environment context perception as well as its influence on cross-talk cancellation (Section 2.2.2), reverberation and diffuse field characteristics are discussed in more details in Section 2.1.6.

2.1.4 Dynamic Cues

In ambiguous natural listening situations where interaural and spectral cues produce insufficient information for the auditory system to localise the sound source, humans turn their heads to resolve this ambiguity. This often occurs when ambiguous interaural cues are introduced at a listener's ears due to the cone of confusion phenomenon.

\(^1\)Data from ftp://sound.media.mit.edu/pub/Data/KEMAR. See also [73] for a description of the measurement procedure.
Chapter 2: 3D Sound Reproduction

A sound source at an azimuth angle $\alpha$ to the right of the listener in the horizontal plane introduces maximally similar interaural cues as a source at azimuth $180^\circ - \alpha$ as mentioned in Section 2.1.1. A human listener would resolve the ambiguous interaural cues by turning his/her head to the right, since the ambiguous cues still suggest that the source is at the listener's right. After turning right, if the interaural cues are minimised, the listener would decide that the source is now in front, otherwise if it is maximised, the decision would be that the source is at the back. Head motions of this type are shown to improve human localisation and decrease front-back reversals [136, 137]. Although head movements improve localisation in natural hearing, they constitute great difficulties to synthetic 3D sound systems. The problem of head movements in synthetic 3D sound systems is discussed in details in Chapter 4.

2.1.5 The Precedence Effect

Natural sound localisation is affected by the above mentioned cues as well as by numerous other psychoacoustical phenomena. In this section, one of those phenomena, the precedence effect, that is directly related to localisation in reverberant environments, is briefly mentioned. The precedence effect, also known as the law of the first wavefront [30], explains an important inhibitory mechanism of the human auditory system that allows human listeners to localise sounds in reverberant environments. When the combination of direct and reflected sounds is heard, the listener does perceive the sound to be coming from the direction of the sound that arrives first at his ears. This is true even if the reflected sound is more intense than the direct sound [82]. However, the precedence effect does not totally eliminate the effect of a reflection. Reflections add a sense of spaciousness and loudness to the sound.

Experiments with two clicks of equal intensity have shown that if the second click arrives within 1 ms after the first, the two clicks are perceived as an integrated entity. The perceived location of this entity obeys the summing localisation regime [82]. Within this regime, there is an orderly weighting such that as the delay increases, the weighting decreases. For delays between 1 and 4 ms, the precedence effect is in operation and it is maximum at about 2 ms delay, where the sound location is perceived to be at the location of the first click. Finally, in the range between 5 to 10 ms, the sound is perceived as two separate clicks and the precedence effect starts to fail. However, it was noticed that the second click not only contributes to spaciousness but the perceived location is also biased towards the position of the second click. Furthermore, the second sound was found to decrease the accuracy of azimuth and elevation localisation compared to anechoic listening situations [24, 25].

In normal listening situations, sound signals last longer than clicks, and reflections arrive at the listener's ears while the direct signal is still heard. In such situations, the precedence effect operates on onsets and transients in the two signals. Furthermore, it was found that the precedence effect for speech signals, better known as the Haas effect, has much different time constants than those mentioned above [82]. Maximal suppression occurs for a delay between 10 to 20 ms, while speech comprehension is affected by reflections later than 50 ms.

2.1.6 Reverberation and Room Acoustics

Spatial sound pressure distribution in an enclosure is rather irregular. Near sound absorbing surfaces, the sound pressure may decrease significantly. Near corners and rigid surfaces, constructive interference of reflections may cause local pressure peaks. Near a sound source, the direct sound dominates. Nevertheless, at areas far from the source, where the reflected sounds play a predominant role, the sound field is often considered as being diffuse, i.e. built up by equally strong uncorrelated plane waves from all directions. At a certain point in the field, the intensities of the waves incident from each direction are equal and the net energy transport in a pure diffuse field is zero [48]. A sound field can be approximated by the diffuse model only if reflections from several directions play a significant role, i.e. there are acoustically hard surfaces at all sides of the enclosure. In this section, a brief review of room acoustics and the diffuse field characteristics is given.

2.1.6.1 Room Impulse Response

The sound field distribution in a room or an enclosure can be calculated by solving the three-dimensional wave equation knowing the boundary conditions in terms of acoustic impedance or reflection coefficients. However, for a certain combination of source-sensor positions, the acoustic transmission can be described by the impulse response from a source to a sensor. Figure 2.3 shows such an impulse response calculated for a fictitious enclosure. The impulse response shows the direct sound arriving
first, followed by distinct reflections, known as the early reflections generated at the room boundaries. Early reflections may be deterministically calculated by solving the eigenfunction problem inside the room for low order eigenmodes. These reflections increase in density and decrease in amplitude with time. After about 80 ms from the direct sound, the reflections arrive equally from all directions and the process is best described statistically as exponentially decaying noise. The statistical part of the impulse response is often referred to as the late reverberation. An important property of the impulse response between two points in a room is that if the source or sensor position changes slightly, the fine structure of the impulse response may change significantly. This is explained by the wide range of wavelengths covered by the audio band and the low speed at which the sound propagates in the air. The audio band covers wavelengths from about 17 mm at 20 kHz to 17 m at 20 Hz. Therefore, a small displacement of a few millimetres corresponds to a large part of a cycle at high frequencies while at low frequencies such a displacement corresponds to a negligible fraction of a cycle. This spatial-frequency dependency of a room impulse response makes it difficult to control the sound field in large zones in a wide frequency range as will be shown in Chapter 4.

### 2.1.6.2 Reverberation Time

A very important parameter in room acoustics is the reverberation time. The reverberation time $T_{60}$ is defined as the time needed for the sound pressure level to decay by 60 dB when a steady state sound source in the room is suddenly switched off. An approximate formula for the reverberation time is $T_{60} \approx \frac{V}{63S}$, where $V$ is the room volume in $m^3$, $\beta$ is the average absorption coefficient of the room boundaries, and $S$ is the surface area of the room in $m^2$. Since the average absorption coefficient $\beta$ is frequency dependent, the reverberation time is also frequency dependent, and is usually given as the average value in an octave band.

At low frequencies, an enclosure may be considered as a three dimensional resonator. The sound waves reflected from the enclosure boundaries constitute standing wave patterns or eigenmodes and resonance may occur for a large number of eigenfrequencies. The number of these eigenmodes increase rapidly with frequency and beyond a certain frequency, the Schroeder frequency, very large number of modes are excited and the sound field may be considered diffuse. The Schroeder frequency depends only on the enclosure volume $V$ and the reverberation time $T_{60}$ and can be approximated by the empirical formula $f_{sh} \approx 2000\sqrt{\frac{T_{60}}{V}}$, where $V$ is in $m^3$, and $T_{60}$ is in seconds. The diffuse approximation is, therefore, valid at lower frequencies in large or more reverberant rooms. For example, in a room of dimensions $V = 4 \times 6 \times 3 m^3$, and a reverberation time $T_{60} = 0.5 s$, the sound field may be approximated by the diffuse field model for frequencies above 167 Hz. Below this frequency, the eigenmode model may be used to calculate the individual modes and their eigenfrequencies.

---

### Figure 2.4: Direct ($L_d$), reverberant ($L_r$), and total ($L_t$) sound pressure levels in an enclosure as functions of the distance from the sound source.

#### 2.1.6.3 Reverberation Distance

The reverberation distance is an indication for the distance from the sound source beyond which the sound field may be considered diffuse. The direct sound pressure level $L_d$ is dependent only on the source characteristics and the source-receiver distance, therefore, it decreases by 6 dB per distance doubling. When the direct sound meets the enclosure boundaries, a fraction of the acoustic energy is reflected to build the reverberation field. When the reverberation field is a pure diffuse field, the reverberant sound pressure level $L_r$ is independent of the distance from the sound source.

The reverberation distance $r_r$ is then defined as the distance from the sound source at which the direct sound pressure and the reverberant sound pressure are equal. The reverberation distance may be approximated by $r_r = 0.25 \sqrt{\frac{3V}{\pi}} \approx 0.06 \sqrt{V/T_{60}}$.

Fig. 2.4 shows the direct ($L_d$), reverberant ($L_r$), and total ($L_t$) sound pressure levels as functions of the distance $r$ from the sound source. At distances close to the sound source ($r/r_r < 1$), the direct sound $L_d$ dominates. At the reverberation distance, the total level $L_t$ is 3 dB above the reverberation level. At distances beyond $3r_r$, $L_r$ exceeds $L_d$ by more than 10 dB and the direct sound may be neglected compared to the reverberant sound. In real rooms, however, the sound level tends to decrease slightly with increasing distance beyond the reverberation distance, and the sound field may be considered diffuse only by approximation.

#### 2.1.6.4 Sound Localisation and Reverberation

In a reverberant environment and beyond the reverberation distance, the total sound pressure level is dominated by the reverberant field rather than by the direct field as shown in Fig. 2.4. Since the loudness cue for distance estimation is based on the decrease in sound pressure with increasing distance from the source as mentioned in Section 2.1.3, it is clear that the loudness cue is valid only in anechoic environments. Beyond the reverberation distance, which may be less than one meter in average
rooms, the total sound pressure level is almost constant, and the loudness cue disappears. At distances smaller than the reverberation distance, the loudness cue may be considered an effective localisation cue.

As the loudness cue becomes less effective with increasing reverberation, the direct-to-reverberant energy ratio \( D/R \) becomes a more effective cue in distance perception as mentioned in Section 2.1.3. This ratio can be shown to be

\[
\frac{D}{R} = \frac{(P_D)_{rms}}{(P_R)_{rms}} = \frac{r_r}{r}, \quad (2.1)
\]

which is dependent only on the reverberation distance \( r_r \), a characteristic of the diffuse field in the enclosure, and the distance \( r \) from the sound source. Therefore, \( D/R \) is considered to be a much more effective distance cue than the loudness cue in reverberant environments. Furthermore, reverberation is considered to be important for the perception of environmental context. The reverberation time and level together with the experience with sounds in reverberant rooms enable a listener to estimate the size and absorptiveness of the surfaces in the environment.

Although reverberation is important for distance and environmental context perception, it was found that localisation accuracy of azimuth and elevation degrades due to reverberation [24, 25]. This is explained by the ability of humans to detect the direction of the early reflections in severe reverberation conditions. The precedence effect mentioned in Section 2.1.5 only partially suppresses the effects of reflected sounds. Moreover, the reverberation makes it difficult for the auditory system to correctly estimate the ITD at low frequencies. This is because in typical rooms, the first reflections arrive before one period of a low frequency sound is completed. Thus, in a reverberant room, low frequency information is essentially useless for localisation and azimuth localisation is severely degraded. In such cases, the important timing information comes from the IED, e.g. from the transients at the onset of a new sound.

### 2.1.7 Head-Related Transfer Functions

Consider the natural listening situation where a sound signal is emitted from a source located at \((r, \alpha, \theta)\) from the centre of the listener's head. The sound pressure at the listener's ears can be modelled by the convolution between the sound signal and the acoustic transmissions between the sound source and the listener's eardrums.

The acoustic transmissions are defined by the impulse responses measured from the source to the listener's left and right ears. The sampled versions of such acoustic transmissions may be expressed as \( h_L(n, r, \alpha, \theta) \) and \( h_R(n, r, \alpha, \theta) \), respectively.

This pair of impulse responses is often referred to as the Head-Related Impulse Response (HRIR) pair. The Discrete Fourier Transform (DFT) of \( h_L, H_L(\omega, r, \alpha, \theta) \), and of \( h_R, H_R(\omega, r, \alpha, \theta) \), are known as the Head-Related Transfer Function (HRTF) pair. An HRTF measured from the source to the listener's eardrum captures all the physical cues to source localisation. Although HRTFs are functions in four variables, for distances greater than about one meter in an anechoic environment, the source is in the far field, and the response falls inversely with the range as mentioned in Section 2.1.6. Most HRTF measurements are anechoic far field measurements, which reduces an HRTF to be a function of three variables, namely, azimuth, elevation, and frequency.

Usually HRTFs are measured in an anechoic environment, and thus do not include the effects of reverberation, which is important for range estimation and environmental context perception. In that case, unless binaural room simulation is used to introduce these important reflections, an improper ratio of direct-to-reverberant sound energy results, and when heard through headphones, the sound often seems to be either very close or inside of the head. It is possible to measure the HRTFs in an actual reverberant setting. This has the disadvantages of limiting the simulated virtual environment to a particular room and leads to very long impulse responses. In the following chapters, both anechoic and reverberant HRTFs will be used.

Anechoic HRTFs of manikins and human subjects have been intensively studied in search for physical characteristics that are related to sound localisation. In the subsequent chapters, frequent use will be made of an anechoic HRTF set measured for the KEMAR manikin [72, 73]. The impulse response of the right ear of KEMAR in the horizontal plane as a function of azimuth angle is shown in Fig. 2.5-[a]. The interaural cues can be readily recognised in Fig. 2.5-[a]. The sound has the highest amplitude and arrives first when it is coming from the right side (azimuth = 90°). Conversely, it has the lowest amplitude and arrives latest when it is coming from the left side (azimuth = 270°). The arrival time varies with azimuth in a more or less sinusoidal fashion as estimated by a spherical head model [30, 54]. Pinna reflections can also be seen in the initial sequence of rapid changes when the source is located at the right side of the head. The cone of confusion phenomenon can also be expected as the response is almost symmetrical about horizontal lines at azimuth = 90° and azimuth = 270°, which constitute the interaural axis.

Fig. 2.5-[b] shows the Fourier transform of the impulse responses in Fig. 2.5-[a]. It is
Figure 2.6: The amplitude in dB of the HRTF of KEMAR’s right ear against linear frequency for a source in the median plane spanning elevation angles from -40° to 90°.

It is also clear from Fig. 2.5-[b] that the response is highest when the source is at the right and weakest when the source is at the left. The pinna notch is also visible around 10 kHz when the source at the right side of the head. For the opposite side, the sound pressure is low due to head shadowing, and the notch is not clear. The broad peak in the range 2 to 3 kHz that is independent of the azimuth can be attributed to the ear canal resonance [72].

The frequency response of the right ear of KEMAR when the source spans the elevation angles between -40° and 90° in the median plane as a function of elevation angle is shown in Fig. 2.6. Fig. 2.5-[b] shows three distinct cross-sections in Fig. 2.6 at elevations -10, 0, and 10 degrees, from which the exact frequency and amplitude scales can be seen. Unlike the response in the horizontal plane, interaural cues in the median plane are negligible. The effect of the pinna reflections is clearly visible in the spectral peaks and notches. The frequency of the first notch and its magnitude are elevation dependent, therefore, this notch is considered a major factor in elevation estimation. The frequency of the first notch ranges from 6 to 12 kHz as the elevation angle changes from -40° to 60°. For elevation angles above about 60°, the notch disappears and there is no spectral dependency on elevation. The ear canal resonance is also visible as the first broad spectral peak which is independent of the elevation angle.

2.1.8 Frequency Selectivity of Human Ears

The fundamental characteristics of the human auditory system can be described in a high-level functional model as a bank of overlapping linear band-pass filters. Psychoacoustical experiments have been used to study the characteristics of this auditory spectral analyser. These experiments led to two main models: the Critical Bands (CB) model [148] and the Equivalent Rectangular Bandwidth (ERB) model [116]. These two models are briefly described below.

2.1.8.1 Critical Bands Model

When the bandwidth of a band-pass noise centred at a frequency $f_c$ is increased while keeping the total energy constant, the loudness remains constant as long as the bandwidth of the band-pass noise is smaller than a certain value. Increasing the noise bandwidth beyond this bandwidth leads to increasing loudness. The critical bandwidth at the centre frequency $f_c$ is defined as that bandwidth at which the loudness starts to increase [148]. The critical bands as a function of frequency may be determined by using the above mentioned loudness experiment. This loudness experiment is just one of five methods described by Zwicker [148] to measure the critical bandwidth. Collecting data from several experiments, it was found that the critical bands show constant bandwidth of 100 Hz for frequencies below 500 Hz, while at frequencies above, the critical bands are about 20% of the centre frequency. This is approximated by

$$\Delta f_{CB} = 25 + 75 [1 + 1.4 f^2]^{0.69},$$

where $f$ is in kHz and $\Delta f_{CB}$ in Hz. This relationship is plotted in Fig. 2.7-[a].

---

Figure 2.7: [a] Critical Bands and Equivalent Rectangular Bandwidth as functions of frequency. [b] Bark and ERB-rate scales as functions of frequency.
Chapter 2: 3D Sound Reproduction

The critical bands concept is so important that it has been defined as a perceptual scale, the Bark scale, to replace the physical frequency scale. This is done by considering each critical band as a unit on the new scale and, therefore, the audio frequency band up to 16 kHz is divided into 24 critical bands or Barks. One Bark on the Bark scale corresponds to a constant distance of about $1.5$ mm along the basilar membrane and, therefore, matches the frequency-place transformation that takes place in the inner ear (cochlea) [148]. The relationship between the Bark scale and the frequency scale is shown in Fig. 2.7-[b] and is expressed mathematically by [148]

$$z = 13 \arctan(0.76 f) + 3.5 \arctan(f/7.5)^2,$$

(2.3)

where $z$ is the Bark value and $f$ is the frequency in kHz.

2.1.8.2 Equivalent Rectangular Bandwidth Model

In the Equivalent Rectangular Bandwidth (ERB) model, the auditory system is considered as a bank of band-pass filters. The shape of the auditory filter at a given centre frequency is measured using a notched-noise masking experiment [116]. This experiment is performed by measuring the threshold of a tone at the centre frequency in question, masked by white noise with a notch (band-stop) centred at the tone frequency, as a function of the notch bandwidth. The auditory filter shape is then the derivative of the threshold curve with respect to the normalised frequency variable $g = f - f_c / f_c$. From these experiments, the auditory filter shape could be approximated by a rounded-exponential filter that takes one of three forms depending on the range over which the filter needs to be modelled. The simplest form of such filter is given by [116]

$$W(g) = (1 + pg) e^{-pg},$$

(2.4)

where $W(g)$ is the auditory filter transfer function, $g$ is the normalised frequency given above, and $p$ is a parameter that controls the bandwidth and the attenuation rate of the filter. The bandwidth of the auditory filter (2.4) centred at $f_c$ is equivalent to the bandwidth of a rectangular filter centred at the same frequency and passing the same noise power. Therefore, the bandwidth of an auditory filter is called the Equivalent Rectangular Bandwidth (ERB) at that specific frequency. The ERB corresponds to a constant distance of about $0.85$ mm on the basilar membrane, and is related to frequency by

$$ERB = 19.5 \times (6.046 f + 1).$$

(2.5)

where ERB is in Hz and $f$ is in kHz. This equation is plotted on the same graph with the critical bandwidth in Fig. 2.7-[a]. The main difference between the two curves is seen at low frequencies. While the CB model shows constant bandwidth at frequencies below $500$ Hz, the ERB model indicates that the bandwidth is frequency dependent in this frequency range. Although many recent experiments suggest that the ERB model is more accurate than the classical CB model, both models are widely used in practical applications. The CB model is especially more used in audio coding applications [80].

2.2 Synthesis of Virtual Sound Sources

Similar to the critical bands, the ERB has also been taken as a perceptual scale replacing the frequency scale, and leading to the ERB-rate scale. The ERB-rate scale relates the number of ERBs to frequency in kHz in the same way as the Bark scale relates the number of CBs to frequency. This relationship is shown in Fig. 2.7-[b] and is given mathematically by

$$\text{ERB-rate} = 11.17 \ln \frac{f + 0.312}{f + 14.675} + 43.$$

(2.6)

Regardless of the discrepancy between the two models at low frequencies, both models suggest that the spectral filtering in the human auditory system can be modelled as a non-uniform filterbank. The bandwidths of the filters increase with increasing centre frequency in a constant-Q (percentage bandwidth) manner. This suggests that human ears perform multiresolution spectral analysis. Lower frequencies are analysed with higher frequency resolution (and, therefore, lower time resolution) than high frequencies. It may also be interpreted as spectral averaging with wider averaging windows at higher frequencies. This property of increasingly coarser resolution with increasing frequency will be exploited in Chapter 4 in improving the robustness of 3D sound systems.

2.2.1 Binaural Synthesis and Headphone Displays

As mentioned in Section 2.1.7, a natural listening situation may be modelled by filtering a monophonic sound signal $x(n)$ by a pair of Head-Related Impulse Responses (HRIR) $h_L(n, r, \alpha, \theta)$ and $h_R(n, r, \alpha, \theta)$ corresponding to a sound source at a position $(r, \alpha, \theta)$. Conversely, filtering a monophonic sound signal through the HRIR pair measured for a sound source at an arbitrary point $(r, \alpha, \theta)$ in a 3D space creates an
auditory event at that point. This is expressed in the frequency domain at a frequency \( \omega \) by (see Section 1.5 for notation conventions)

\[
E(\omega, r, \alpha, \theta) = H(\omega, r, \alpha, \theta) X(\omega),
\]

(2.7)

where \( E(\omega, r, \alpha, \theta) = [E_L(\omega, r, \alpha, \theta) \ E_R(\omega, r, \alpha, \theta)]^T \) is a column vector of left and right ear signals expressed at \( \omega \), and usually referred to as the binaural signals. The column vector \( H(\omega, r, \alpha, \theta) = [H_L(\omega, r, \alpha, \theta) \ H_R(\omega, r, \alpha, \theta)]^T \) contains the HRTF pair, and \( X(\omega) \) is the monophonic sound signal at \( \omega \) that is independent of the spatial coordinates. Provided that the HRIR pair matches those of the listener, playing the binaural signals at the corresponding listener’s eardrums (e.g. through headphones) creates an auditory event (a virtual source) at \((r, \alpha, \theta)\). In general, \( K \) auditory events may be created simultaneously in the same virtual space by extending the scalar sound signal in (2.7) to be a column vector of length \( K \). Omitting the dependency on spatial coordinates, this can be expressed as

\[
\begin{bmatrix}
E_L(\omega) \\
E_R(\omega)
\end{bmatrix}
= 
\begin{bmatrix}
H_{L1}(\omega) & \cdots & H_{Lk}(\omega) & \cdots & H_{Lm}(\omega) \\
H_{R1}(\omega) & \cdots & H_{Rk}(\omega) & \cdots & H_{Rm}(\omega)
\end{bmatrix}
\begin{bmatrix}
X_1(\omega) \\
\vdots \\
X_k(\omega) \\
\vdots \\
X_K(\omega)
\end{bmatrix},
\]

(2.8)

which may be expressed compactly as

\[
E(\omega) = H(\omega) X(\omega).
\]

(2.9)

While (2.8) gives the binaural signals at one frequency only, it should be kept in mind that there are as many equations of the form (2.8) as there are frequencies. Using frequency representations of signals at a single frequency effectively reduces the matrix \( H \) from an array of 3-indexes (ear, input signal, frequency) to a matrix of 2-indexes (ear, input signal). This enables using well known matrix algebra while keeping good understanding of the physical process. In the following, the explicit dependency on \( \omega \) will also be dropped to enhance the readability of the equations.

The element \( X_k \) of the vector \( X \) in (2.9) corresponds to the sound signal played by the \( k \)th virtual source and, therefore, must be filtered by the HRTF pair \( H_k = [H_{Lk} \ H_{Rk}]^T \). This is shown in a block diagram form in Fig. 2.8. In this case, only two transducers are required to deliver the binaural signals to the corresponding listener’s ears. This justifies the statement made in the introduction of this chapter that \( K \) virtual sound sources may be created using two loudspeakers only.

The above principle can be further generalised to create \( K \) virtual sound images at the ears of \( M \) listeners by expanding the column vector \( E \) and the HRTF matrix \( H \) to be of dimensions \([2M \times 1]\) and \([2M \times K]\), respectively, which results in

\[
\begin{bmatrix}
E_{L1} \\
E_{R1} \\
E_{L2} \\
E_{R2} \\
\vdots \\
E_{LM} \\
E_{RM}
\end{bmatrix}
= 
\begin{bmatrix}
H_{L11} & H_{L12} & \cdots & H_{L1K} \\
H_{R11} & H_{R12} & \cdots & H_{R1K} \\
H_{L21} & H_{L22} & \cdots & H_{L2K} \\
H_{R21} & H_{R22} & \cdots & H_{R2K} \\
\vdots & \vdots & \ddots & \vdots \\
H_{LM1} & H_{LM2} & \cdots & H_{LMK} \\
H_{RM1} & H_{RM2} & \cdots & H_{RMK}
\end{bmatrix}
\begin{bmatrix}
X_1 \\
X_2 \\
\vdots \\
X_K
\end{bmatrix}.
\]

(2.10)

In (2.10), \( E_{Lm} \) and \( E_{Rm} \) are the signals at the left and right ears of the \( m \)th listener, respectively, and \( H_{Lm} \) and \( H_{Rm} \) form the HRTF pair measured from the ears of the \( m \)th listener to the position of the \( k \)th source. In this case, \( M \) headphones are required, one for each listener.

In 3D sound systems, the question arises of how to deliver the electrical binaural signals to the listener’s eardrums as acoustic waves. Headphones deliver \( E_L \) at the left ear without any cross-talk from \( E_R \) and \( E_R \) at the right ear without any cross-talk from \( E_L \) and, therefore, form a simple way of performing this transducer function. However, headphones have their own drawbacks: they may not be comfortable to
Chapter 2: 3D Sound Reproduction

use for a long period of time. They also isolate their user from the surrounding environment. Sounds heard over headphones often seem to be too close or inside the listener’s head. Since the physical sources (the headphones) are actually very close to the listener’s ears, compensation is needed to eliminate the acoustic cues to their locations. This compensation is sensitive to the headphones position. Finally, headphones can have notches and peaks in their frequency responses that resemble the pinna responses. If uncompensated headphones are used, elevation perception may be severely compromised [54].

2.2.2 Loudspeaker Displays and Cross-Talk Cancellation

Alternatively, a pair of loudspeakers arranged in a stereophonic setting may be used to transform the binaural signals to acoustic waves. Such 3D sound systems are often referred to as loudspeaker displays. In loudspeaker displays, the sound emitted from each loudspeaker is heard by both ears. Furthermore, on its way to the listener’s ears, the sound is filtered through a mixing matrix of the four acoustic transfer functions between the loudspeakers and the ears as shown in Fig. 2.9. To correctly deliver the binaural signals to the listener’s eardrums, this mixing process has to be inverted. This inversion is usually performed by a network of filters known as the cross-talk canceller, which is represented by the block $G$ in Fig. 2.9. Using frequency domain representations of signals and dropping the frequency and spatial dependency for clarity, the sound waves at the left and right ears can be expressed as

$$
\begin{bmatrix}
E_L \\
E_R
\end{bmatrix} = \begin{bmatrix}
C_{LL} & C_{LR} \\
C_{RL} & C_{RR}
\end{bmatrix} \begin{bmatrix}
G_{11} & G_{12} \\
G_{21} & G_{22}
\end{bmatrix} \begin{bmatrix}
H_L \\
H_R
\end{bmatrix} X,
$$

which may be written compactly as $E = CGH X$. To faithfully deliver the binaural sounds $Z_L$ and $Z_R$ to the corresponding ears, the cross-talk canceller $G$ must equal the inverse of the matrix $C$, so that $CG = I$, and $G$ is given by

$$
G = \frac{1}{C_{LL}C_{RR} - C_{RL}C_{LR}} \begin{bmatrix}
C_{RR} & -C_{LR} \\
-C_{RL} & C_{LL}
\end{bmatrix}.
$$

The elements of the square matrix $C$ are the acoustic transfer functions between the listener’s eardrums and the loudspeakers as shown in Fig. 2.9. Each of these transfer functions $C_{AB}$ is composed of the responses of the corresponding loudspeaker $S_A$, the head-related transfer function $H_{AB}$, the room between the corresponding loudspeaker and ear $R_{AB}$, the microphone used to measure the sound at the eardrum $M_A$, the Analogue-to-Digital Converter (ADC), the Digital-to-Analogue Converter (DAC), and the signal conditioning devices. Therefore, the elements of $C$ are complex functions in frequency and space coordinates. They are often non-minimum phase and contain deep notches due to pinna reflections and room reverberation. Moreover, since loudspeakers are acoustic band-pass filters, their responses have much less energy at the edges of the audio frequency band. For those reasons, the inversion of $C$ required by the cross-talk cancellation matrix in (2.12) is difficult to calculate. This is in addition to the processing complexity required to implement the matrix inversion in real-time.

A computationally attractive approach to loudspeaker displays is to combine the binaural synthesis and the cross-talk cancellation functions as shown in Fig. 2.10. In this configuration, the [2-input x 2-output] cross-talk canceller, which requires four separate filters as given by (2.12) is combined with the [1-input x 2-output] binaural synthesis subsystem. The combined solution is, therefore, [1-input x 2-output] and can be implemented with two filters given by $W = C^{-1} H$, and expressed fully by

$$
\begin{bmatrix}
W_L \\
W_R
\end{bmatrix} = \frac{1}{C_{LL}C_{RR} - C_{RL}C_{LR}} \begin{bmatrix}
C_{RR}H_L - C_{LR}H_R \\
-C_{RL}H_L + C_{LL}H_R
\end{bmatrix}.
$$

This approach does not explicitly require a matrix inversion and embeds the convolution with the HRTF pair in the filters’ coefficients. Therefore, it is computationally more efficient than the separate implementation of binaural synthesis followed by cross-talk cancellation. Furthermore, the convolution with the HRTFs, $H_R$ and $H_L$, may compensate for some of the deep notches in the elements of $C$, which may improve the numerical stability of the total solution, although $C^{-1}$ may be ill-conditioned. Moreover, it can be shown that a causal implementation of $W$ exists provided that $H$ possesses sufficient delay. However, the problem of inverting a non-minimum phase system remains and will be discussed in Chapter 3.

Similar to the binaural synthesis case, a loudspeaker display system may be gener-
alised to accommodate more virtual sound images at the ears of multiple listeners. By analogy with (2.10), to create \( K \) virtual sound images at the ears of \( M \) listeners, \( \mathbf{E} \) and \( \mathbf{H} \) are extended to be of dimensions \([2M \times 1]\) and \([2M \times K]\), respectively. Unlike in binaural synthesis, the number of loudspeakers is not constrained to \( 2M \) since all listeners share listening to the loudspeakers as long as they share the listening space. Let the number of reproduction loudspeakers be \( L \), then the dimension of \( \mathbf{C} \) becomes \([2M \times L]\). For the signals at the listeners’ ears to become equal to the natural listening situation with \( L \) loudspeakers and \( K \) sources, the following equation must be valid at every frequency \( \omega \):

\[
\mathbf{E}(\omega) = \mathbf{H}(\omega) \mathbf{X}(\omega) = \mathbf{C}(\omega) \mathbf{W}(\omega) \mathbf{X}(\omega),
\]

(2.14)

where \( \mathbf{W} \) is an \([L \times K]\) matrix of control filters. The solution to those filters is then given by

\[
\mathbf{W}(\omega) = \mathbf{C}^{-1}(\omega) \mathbf{H}(\omega).
\]

(2.15)

The existence of the solution (2.15), the choice of the number of reproduction loudspeakers \( L \), and the implementation details of multichannel loudspeaker displays are discussed at length in Chapter 3.

Chapter 3

Adaptive Filters Approach to 3D Sound Systems

The control filters in loudspeaker displays must implement both binaural synthesis and cross-talk cancellation functions as discussed in Section 2.2. The cross-talk cancellation subsystem is an inverse of an often non-minimum phase and ill-conditioned matrix of electro-acoustic transfer functions. An exact solution to the cross-talk canceller is difficult to calculate and in some cases, may not be possible at certain frequencies. Alternative to exact direct inversion, a statistical least mean square solution may be obtained for the matrix inverse. The advantage of such an approach is that it can be made adaptive, so that the filters can be designed in-situ. Such an adaptive solution also offers the possibility to track and correct the filters when changes in the electro-acoustic system occur. This chapter discusses theoretical and implementation issues concerning adaptive 3D sound systems.

An adaptive loudspeaker display that creates a single sound image at the ears of one listener requires two loudspeakers and two microphones placed inside the listener’s ears. More microphones are needed as the number of listeners increases, and more filters are needed as the number of virtual sound images increases. Instead of treating any of the above special cases, a generalised model is introduced in Section 3.1. This model describes the general task of controlling the sound field produced by playing \( K \) pre-recorded audio signals through \( L \) loudspeakers at the proximity of \( M \) microphones. In addition to generalising the number of sound sources, loudspeakers, and microphones used in the reproduction system, the model is capable of describing a wide class of applications including synthesis of virtual sound sources, cross-talk cancellation, and active noise control.

The optimum least mean square solution for the control filters in the generalised model is derived in Section 3.2. Although important from the numerical analysis point of view, real-time implementation of the system would require approaching the optimum solution using an iterative algorithm. Such an iterative algorithm, the Multiple Error Filtered-X Least Mean Square (MEFX) algorithm, is introduced in Section 3.3.1. In Section 3.3.2, the convergence properties of the MEFX algorithm are examined, which shows the dependency of the convergence speed on the electro-
Chapter 3: Adaptive Filters Approach to 3D Sound Systems

3.1 A Generalised Model

3.1.1 Block Diagram of the Generalised Model

A generalised block diagram of a multichannel audio reproduction system is shown in Fig. 3.1. A set of \( L \) reproduction loudspeakers \( \{S_1, S_2, \ldots, S_L\} \) are used to play \( K \) pre-recorded audio signals defined at the sample index \( n \) by

\[
\mathbf{x}(n) = \begin{bmatrix} x_1(n) & x_2(n) & \cdots & x_K(n) \end{bmatrix}^T. \quad (3.1)
\]

The sound field generated due to these loudspeakers is required to be controlled at the proximity of a set of \( M \) microphones (receivers) \( \{R_1, R_2, \ldots, R_M\} \) to a set of desired values defined as

\[
\mathbf{d}(n) = \begin{bmatrix} d_1(n) & d_2(n) & \cdots & d_M(n) \end{bmatrix}^T. \quad (3.2)
\]

Sound waves emitted from the loudspeakers are filtered through the \([M \times L]\) matrix of electro-acoustic transfer functions \( \mathbf{C}(\omega) \) before reaching the microphones' positions. The matrix \( \mathbf{C}(\omega) \) contains the Fourier transforms of the impulse responses \( \{\mathbf{c}_{ml} : m = 1, 2, \ldots, M, l = 1, 2, \ldots, L\} \) evaluated at a frequency \( \omega \) and defined as

\[
\mathbf{C}(\omega) = \begin{bmatrix} C_{11}(\omega) & C_{12}(\omega) & \cdots & C_{1L}(\omega) \\
C_{21}(\omega) & C_{22}(\omega) & \cdots & C_{2L}(\omega) \\
\vdots & \vdots & \ddots & \vdots \\
C_{M1}(\omega) & C_{M2}(\omega) & \cdots & C_{ML}(\omega) \end{bmatrix}. \quad (3.3)
\]

The set of desired sound fields given by the vector \( \mathbf{d}(n) \) is usually well correlated with the set of \( K \) input signals \( \mathbf{x}(n) \). Therefore, \( \mathbf{d}(n) \) may be considered as the result of filtering \( \mathbf{x}(n) \) through a \([M \times K]\) matrix of transfer functions \( \mathbf{H}(\omega) \) (not shown in Fig. 3.1). This filtering operation may be expressed in the frequency domain as

\[
\mathbf{d}(\omega) = \mathbf{H}(\omega) \mathbf{X}(\omega), \quad (3.7)
\]

where \( \mathbf{D}(\omega) \) and \( \mathbf{X}(\omega) \) are the DFT of the time history of \( \mathbf{d}(n) \) and \( \mathbf{x}(n) \), respectively, as mentioned in Section 1.5. The matrix \( \mathbf{H}(\omega) \) contains the DFT of the impulse response functions stored in the\[L \times K\] matrix of transfer functions \( \mathbf{C}(\omega) \).
The input signals through $s(n-k)$ responses of the reproduction loudspeakers to obtain better sound quality. This may be achieved by letting cellation, virtual source synthesis, concert hall simulation, correction of the responses to disturbances at the microphones. In this case, the matrix $H$ contains delayed unit impulse responses.

The most popular application of the generalised model is in active noise cancellation, virtual source synthesis, concert hall simulation, correction of the responses. In cross-talk cancellation (see Section 2.2.2), it is desired to exactly reproduce the sound of the filters to achieve the control task. The design of these filters is discussed at length in this chapter. Examples illustrating the use of the generalised model in different audio reproduction applications are discussed in the next section.

### 3.1.2 Example Applications

In the generalised model introduced above, different audio reproduction applications may be described by defining different desired sets of sound fields and filters. Consequently, it is the matrix $H(\omega)$ that defines the nature of the application. Examples of the applications that may be described by the model are cross-talk cancellation, virtual source synthesis, concert hall simulation, correction of the responses of the reproduction loudspeakers, and active noise control.

In cross-talk cancellation (see Section 2.2.2), it is desired to exactly reproduce the input signals at the microphones. In this case, $d(n) = x(n)$, and $H(\omega) = I$, the $[K \times K]$ identity matrix. Applying this for all frequencies, the transfer function between the $m$th microphone and the $k$th reproduction loudspeaker is defined as $H_{mk}(\omega) = H_{mk}(n) = \delta(n)$.

Alternatively, $K$ virtual sound images may be generated at the ears of $M/2$ (even) listeners if $d(n)$ is the result of filtering the input signals through the appropriate matrix of HRTFs as given by (2.10). The generalised model may also be used to simulate a concert hall by setting $H(\omega)$ to be a matrix of transfer functions measured in a concert hall as mentioned in Section 2.1.6. Another application that may be described by the model is correcting the responses of the reproduction loudspeakers to obtain better sound quality. This may be achieved by letting $W(\omega)$ be an inverse model of the transfer functions. The most popular application of the generalised model is in active noise cancellation [56, 93, 109]. In such application, a set of $K$ (undesired rather than desired) sound disturbances $d(n)$ are to be silenced at the microphones. This is achieved by filtering the input signals through $W(\omega)$ to generate sound waves $d(n)$ that are equal in amplitude but opposite in phase to the disturbances at the microphones' positions such that the net microphones' outputs $g(n)$ are minimised, where $g(n)$ is given by

$$g(n) = [e_1(n) \ e_2(n) \ \cdots \ e_M(n)]^T.$$  

The only difference between the above mentioned applications when using the generalised model is the desired response $d(n)$ at the microphones. Therefore, a generalised solution for $W(\omega)$ that is valid for the whole class of applications may be obtained by solving the system equations for an arbitrary desired response. The solution for a specific application is then obtained by substituting its specific desired response into the general solution. The optimum solution $W_{opt}(\omega)$ in the least mean square sense is derived in Section 3.2. An iterative approach to this optimum solution is discussed in Section 3.3.1. Examples of specific applications are discussed for active noise control, synthesis of virtual sound images, and cross-talk cancellation in Sections 3.3.5, 3.3.6, and 3.3.7, respectively.

### 3.2 The Optimum Least Mean Square Solution

The optimum Least Mean Square (LMS) solution for the system shown in Fig. 3.1 is obtained by minimising a performance index function $\xi(n)$, usually taken as the sum of the mean squared error signals

$$\xi(n) = \sum_{m=1}^{M} E\{e^2(n)\} = E\{g^T(n) \ g(n)\},$$  

where $E\{\cdot\}$ denotes the mathematical expectation. In the following discussion, all filters $\{w_{lk} : l = 1, 2, \ldots, L; k = 1, 2, \ldots, K\}$ are assumed to be Finite Impulse Response (FIR) digital filters, each of length $N_w$ and defined as

$$\mathbf{w}_{lk} = [w_{lk,0} \ w_{lk,1} \ \cdots \ w_{lk,N_w-1}]^T.$$  

The signal driving the $l$th reproduction loudspeaker $y_l(n)$ is the sum of the outputs of the filters $\{w_{lk} : k = 1, 2, \ldots, K\}$, which may be expressed as

$$y_l(n) = x_1^T(n) \mathbf{w}_{l1} + x_2^T(n) \mathbf{w}_{l2} + \cdots + x_K^T(n) \mathbf{w}_{lK},$$  

where the time history of the $k$th input signal $x_k(n)$ is defined as

$$x_k(n) = [x_k(n) \ x_k(n-1) \ \cdots \ x_k(n-N_w+1)]^T.$$  

Defining the $[K \times 1]$ composite input signal vector $\mathbf{x}(n)$ and the $[K \times 1]$ composite weight vector $\mathbf{w}_I$ as

$$\mathbf{x}(n) = [x_1(n) \ x_2(n) \ \cdots \ x_K(n)]^T,$$

$$\mathbf{w}_I = [\mathbf{w}_{l1} \ \mathbf{w}_{l2} \ \cdots \ \mathbf{w}_{lK}]^T,$$  

1It will be shown in Section 3.3.7 that $b_{mk}(n) = \delta(n)$ results in a non-causal set of filters $W(\omega)$. Causal filters produce a delayed version of the inputs $d(n) = [x_1(n-\Delta_1) \ x_2(n-\Delta_2) \ \cdots \ x_K(n-\Delta_K)]$ at the microphones. In this case, the matrix $H$ contains delayed unit impulse responses $H_{mk}(n) = \delta(n-\Delta_k) : m = k = 1, 2, \cdots, K$. 

41
equation (3.12) can be written as

\[ y_r(n) = x^T(n) w. \]  

(3.16)

From (3.5) and (3.16), the \([L \times 1]\) vector of control signals \(y(n)\) driving the reproduction loudspeakers is expressed as

\[
\begin{bmatrix}
  y_1(n) \\
  y_2(n) \\
  \vdots \\
  y_L(n)
\end{bmatrix} =
\begin{bmatrix}
  x(n) & 0 & \cdots & 0 \\
  0 & x(n) & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & x(n)
\end{bmatrix}^T
\begin{bmatrix}
w_1 \\
w_2 \\
\vdots \\
w_L
\end{bmatrix}. 
\]  

(3.17)

Further defining the \([KLNw \times 1]\) composite vector \(w = [w_1 w_2 \cdots w_L]^T\), and the \([KLNw \times L]\) composite matrix \(x(n)\) of input signals given by the first factor in the right hand side of (3.17), (3.17) can be written compactly as

\[ y(n) = x^T(n) w. \]  

(3.18)

The system response vector \(\tilde{d}(n)\) results from filtering the input to the loudspeakers \(y(n)\) through the matrix of electro-acoustic transfer functions \(C(\omega)\). The resulting component at the \(m^{th}\) microphone is, therefore, given by

\[ \tilde{d}_m(n) = c_{ml} \ast y_L(n) + c_{ml2} \ast y_L(n) + \cdots + c_{mlN_c} \ast y_L(n), \]  

(3.19)

where all electro-acoustic impulse responses \(c_{ml} : m = 1, 2, \ldots, M, l = 1, 2, \ldots, L\) are assumed to be FIR filters of length \(N_c\) defined as

\[ c_{ml} = \begin{bmatrix} c_{m1.0} & c_{m1.1} & \cdots & c_{m1N_c-1} \end{bmatrix}^T. \]  

(3.20)

Defining the time history of the signal driving the \(l^{th}\) loudspeaker \(y_l(n)\) as

\[ y_L(n) = \begin{bmatrix} y_l(n) & y_l(n-1) & \cdots & y_l(n-N_c+1) \end{bmatrix}^T, \]  

(3.21)

and the \([M \times LN_c]\) composite matrix of acoustic impulse responses \(c\) as

\[ c = \begin{bmatrix} c_{11}^T & c_{12}^T & \cdots & c_{1L}^T \\
  c_{21}^T & c_{22}^T & \cdots & c_{2L}^T \\
  \vdots & \vdots & \ddots & \vdots \\
  c_{M1}^T & c_{M2}^T & \cdots & c_{ML}^T \end{bmatrix}, \]  

(3.22)

the \([M \times 1]\) vector \(\tilde{d}(n)\) of sound waves at the microphones can be expressed as

\[ \tilde{d}(n) = c \ast y(n) = c \ast [x^T(n) w] = x_f(n) w, \]  

(3.23)

\[ x_f(n) = \begin{bmatrix} x_{f11}(n) & \cdots & x_{f1L}(n) & \cdots & x_{f1K}(n) \\
  x_{f21}(n) & \cdots & x_{f2L}(n) & \cdots & x_{f2K}(n) \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  x_{fM1}(n) & \cdots & x_{fM2}(n) & \cdots & x_{fMK}(n) \end{bmatrix}. \]  

(3.24)

where \(x_{fmlk}(n) = c_{ml} \ast x_k(n)\) is the last \(N_w\) samples of the result of filtering the \(k^{th}\) input signal through the electro-acoustic impulse response between the \(m^{th}\) microphone and the \(l^{th}\) loudspeaker

\[ x_{fmlk}(n) = \begin{bmatrix} x_{fmlk}(n) & x_{fmlk}(n-1) & \cdots & x_{fmlk}(n-N_w+1) \end{bmatrix}. \]  

(3.25)

The error vector \(e(n)\) in Fig. 3.1 can then be expressed as

\[ e(n) = d(n) - x_f(n) w. \]  

(3.26)

Substituting in the performance index (3.10) gives

\[ \xi(n) = E\{ d^T(n) d(n) - 2 w^T x_f^T(n) d(n) + w^T x_f^T(n) x_f(n) w \}. \]  

(3.27)

The optimum LMS solution \(w_{opt}(n)\) for the composite weight vector is the vector that minimises the performance index \(\xi(n)\) given by (3.27). The optimum solution is, therefore, obtained by setting the first derivative (gradient) of \(\xi(n)\) with respect to the composite weight vector \(w\) to zero. The gradient is readily obtained from (3.27) to be

\[ \nabla \xi(n) = \frac{\partial \xi(n)}{\partial w} = 2 E\{ x_f^T(n) x_f(n) w - x_f^T(n) d(n) \}. \]  

(3.28)

The optimum LMS solution \(w_{opt}(n)\) is then obtained by setting \(\nabla \xi(n)\) to zero

\[ w_{opt}(n) = E\{ x_f^T(n) x_f(n) \}^{-1} \{ E\{ x_f^T(n) d(n) \} \}. \]  

(3.29)

and the corresponding minimum value of \(\xi(n)\) is given by

\[ \xi_{min} = E\{ d^T(n) d(n) \} - E\{ w_{opt}^T(n) x_f^T(n) d(n) \}. \]  

(3.30)

From (3.29), the optimum weight vector \(w_{opt}(n)\) exists only if the matrix \(\{ x_f^T(n) x_f(n) \}\) is non-singular. Since the matrix \(x_f(n)\) is the convolution between the input signals and the electro-acoustic transfer functions as given by (3.24), not only \(x(n)\), but also \(C(\omega)\) influence the solution. Depending on the dimensions of \(C(\omega)\), three cases may be recognised:

1. The number of loudspeakers equals the number of microphones \((L = M)\): In this case, \(C(\omega)\) is a square matrix. Consider \(K = 1\) for simplicity and using frequency domain representations, the system of equations

\[ E(\omega) = D(\omega) - C(\omega) Y(\omega) \]  

(3.31)
3.3 Iterative LMS Solutions

3.3.1 The Multiple Error Filtered-X LMS Algorithm

An alternative approach to obtaining the optimum LMS solution, other than direct calculation using (3.29) or (3.33), is to use an iterative algorithm such as the Multiple Error Least Mean Square (MELMS) algorithm [56, 58, 59, 108, 109, 110]. Such an adaptive approach is preferred over the direct calculation of \( \mathbf{w}_{\text{opt}}(n) \) since it offers in-situ design of the filters. It also enables a convenient method to readjust the filters whenever a change occurs in the electro-acoustic transfer functions as will be discussed in Section 4. The MELMS algorithm employs the steepest descent approach to search for the minimum of the performance index (3.10). This is achieved by successively updating the filters' coefficients by an amount proportional to the negative of the gradient \( \nabla(n) \),

\[
\mathbf{w}(n+1) = \mathbf{w}(n) + \mu \left( -\nabla(n) \right),
\]

where \( \mu \) is the step size that controls the convergence speed and the final misadjustment [144]. An approximation often used in such iterative LMS algorithms is to update the vector \( \mathbf{w} \) using the instantaneous value of the gradient \( \nabla(n) \) instead of its expected value \( \nabla(n) = E\{\nabla(n)\} \) [144], leading to the well known LMS algorithm. Using (3.26) in (3.28), the gradient can be written as \( \nabla(n) = 2E\{-x_f(n) e(n)\} \). The update equation for the MELMS algorithm is then given by replacing \( \nabla(n) \) in (3.35) by its instantaneous value,

\[
\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu x_f(n) e(n).
\]

This update algorithm is often referred to as the Multiple Error Filtered-X (MEFX) algorithm. Implementation of (3.36) requires calculating the matrix \( \mathbf{x}_f(n) \) given by (3.24), which implies measuring all the electro-acoustical transfer functions \( \mathbf{e}_{\text{act}} \) and filtering each input signal through all \( ML \) transfer functions to construct the \( KLM \) elements of \( \mathbf{x}_f(n) \). This is shown in Fig. 3.2, where the measured matrix of electro-acoustical transfer functions is represented by the block \( \mathbf{C} \) to distinguish it from the physical one represented by the block \( \mathbf{C} \). The MEFX algorithm is known to be robust to estimation errors. It shows stable convergence properties as long as the phase error in any of the measured transfer functions at any frequency is less than \( \pm 90^\circ \) [56, 59, 105, 129, 144], while amplitude errors result in less accurate solutions [124]. On-line measurement of the matrix of electro-acoustical transfer functions \( \mathbf{C} \) is discussed in Section 4.7.
Similarly, the noisy steepest descent method may be used to iteratively approach the optimum solution in the case of weighted performance index. In this case

\[ \nabla(n) = 2E\{-x_f^T(n) \Gamma_e g(n) + \Gamma_w w(n)\}, \]  
(3.37)

which leads to the following update equation

\[ w(n+1) = w(n) - \mu \Gamma_w w(n) + 2 \mu x_f^T(n) \Gamma_e g(n). \]  
(3.38)

For \( \Gamma_w = \text{diag}\{\gamma_1, \gamma_2, \cdots, \gamma_{MKL}\} \), each filter weight \( w_i \) is independently weighted by the weighting factor \( \gamma_i \). In this case, when \( \Gamma_e = I \), and letting \( \Gamma = I - \mu \Gamma_w \), the update equation becomes the multiple error leaky LMS algorithm

\[ w(n+1) = \Gamma w(n) + 2 \mu x_f^T(n) g(n). \]  
(3.39)

### 3.3.2 Convergence Properties of the MEFX Algorithm

The convergence properties of the regular LMS algorithm (when \( C(\omega) = I \)) for a single input are determined by the eigenvalues of the autocorrelation matrix of the input signal [144]. Since the input signals to the MEFX update algorithm are the filtered signals given by the elements of the matrix \( x_f(n) \), the convergence properties of the MEFX are expected to be determined by the matrix \( \mathbf{R}_f(n) = E\{x_f^T(n) x_f(n)\} \) [105, 129]. The matrix \( \mathbf{R}_f(n) \) is, in general, positive definite for \( L \leq M \), and the error surface has a unique global minimum. In this section, the convergence properties of the time domain MEFX are discussed using frequency domain representations of signals [37, 38, 64]. Frequency domain analysis decouples the system to approximately uncorrelated frequency bins, and the convergence properties are approximately determined by independent modes at each frequency. Furthermore, frequency domain analysis allows the study of the effects of the electro-acoustic transfer functions on the system performance, which is discussed in detail in Section 3.3.3.

From the definition of the matrix \( x_f(n) \) given in (3.24), the frequency domain version \( \mathbf{X}_f(n, \omega) \) at a frequency \( \omega \) may be expressed as

\[
\mathbf{X}_f(n, \omega) = \begin{bmatrix}
C_{11} X_1 & \cdots & C_{1L} X_L & \cdots & C_{1L} X_K \\
C_{21} X_1 & \cdots & C_{2L} X_L & \cdots & C_{2L} X_K \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
C_{M1} X_1 & \cdots & C_{M1} X_K & \cdots & C_{ML} X_L \\
C_{M2} X_1 & \cdots & C_{M2} X_K & \cdots & C_{ML} X_K \\
\end{bmatrix},
\]  
(3.40)

where the explicit dependency on \( n \) has been dropped inside the matrix for clarity. Representing the system at one frequency \( \omega \) as in (3.40) reduces the dimensions of \( \mathbf{X}_f(n, \omega) \) from \( [M \times KLN] \) to \([M \times KL]\). Consequently, the dimensions of \( \mathbf{W}(n, \omega) \) reduces to \([KL \times 1]\) corresponding to \( KL \) filters each with a single coefficient at \( \omega \). However, it should be kept in mind that there are as many equations of the form (3.40) as there are frequencies.

Assuming slow adaptation, the adaptive filters may be considered time invariant linear filters during a short period of time. In this period, the MEFX algorithm can be approximated in the frequency domain using the following equations

\[ \mathbf{E}(n, \omega) = \mathbf{D}(n, \omega) + \mathbf{X}_f(n, \omega) \mathbf{W}(n, \omega), \]  
(3.41)

\[ \mathbf{W}(n+1, \omega) = \mathbf{W}(n, \omega) + 2 \mu \mathbf{X}_f^H(n, \omega) \mathbf{E}(n, \omega), \]  
(3.42)

where \( \cdot^H \) denotes the complex conjugate transpose and \( \mu' \) is the frequency domain step size. Upper-case letters are the DFT of the corresponding lower-case variables at iteration \( n \) as mentioned in Section 1.5. Substituting (3.41) into (3.42) and taking the mathematical expectation, (3.42) may be approximated by

\[ \mathbf{E}(\mathbf{W}(n+1, \omega)) - \mathbf{W}_{\text{opt}}(\omega) \approx (1 - 2\mu' \mathbf{R}_f(\omega)) [\mathbf{E}\{\mathbf{W}(n, \omega)\} - \mathbf{W}_{\text{opt}}(\omega)], \]  
(3.43)

where \( \mathbf{R}_f(\omega) = \mathbf{E}\{\mathbf{X}_f^H(n, \omega) \mathbf{X}_f(n, \omega)\} \), and \( \mathbf{W}_{\text{opt}}(\omega) \) is the Fourier transform of the optimal weight vector \( \mathbf{w}_{\text{opt}}(n) \) given by (3.29).

Since \( \mathbf{R}_f(\omega) \) is, in general, not diagonal, the components of \( \mathbf{W}(n, \omega) \) are cross-coupled. Provided that \( \mathbf{R}_f(\omega) \) is symmetric and positive, it may be factored into its eigenvalues and eigenvectors [144],

\[ \mathbf{R}_f(\omega) = \mathbf{Q}(\omega) \mathbf{A}_f(\omega) \mathbf{Q}^H(\omega), \]  
(3.44)

where \( \mathbf{A}_f(\omega) = \text{diag}\{\lambda_{f1}(\omega), \lambda_{f2}(\omega), \cdots, \lambda_{fKL}(\omega)\} \) is a diagonal matrix of the eigenvalues of \( \mathbf{R}_f(\omega) \) and \( \mathbf{Q}(\omega) = \mathbf{Q}_1(\omega) \mathbf{Q}_2(\omega) \cdots \mathbf{Q}_{KL}(\omega) \) is the matrix of eigenvectors of \( \mathbf{R}_f(\omega) \) satisfying \( \mathbf{Q} \mathbf{Q}^H = \mathbf{Q}^H \mathbf{Q} = I \). Substituting (3.44) in (3.43) and using the rotated weight difference vector \( \mathbf{X}(n, \omega) = \mathbf{Q}^H (\mathbf{E}\{\mathbf{W}(n, \omega)\} - \mathbf{W}_{\text{opt}}(\omega)) \) results in

\[ \mathbf{X}(n+1, \omega) = (1 - 2\mu' \mathbf{A}_f(\omega)) \mathbf{X}(n, \omega), \]  
(3.45)

Equation (3.45) may be solved by induction to obtain \( \mathbf{X}(n, \omega) \) as a function of the initial rotated weight difference vector \( \mathbf{X}(0, \omega) \) giving

\[ \mathbf{X}(n, \omega) = (1 - 2\mu' \mathbf{A}_f(\omega))^n \mathbf{X}(0, \omega). \]  
(3.46)

Since \( (1 - 2\mu' \mathbf{A}_f(\omega)) \) is now diagonal, (3.46) can be decomposed into \( LM \) independent equations, each corresponding to a fundamental system mode at \( \omega \). From these decoupled equations, it can be seen that the MEFX is stable at \( \omega \) if

\[ |(1 - 2\mu' \lambda_{fi}(\omega))| < 1, \quad i = 1, 2, \cdots, KL. \]  
(3.47)

This defines the limits for the step size to be \( 0 < \mu' < 1/\lambda_{f_{\text{max}}}(\omega) \), where \( \lambda_{f_{\text{max}}}(\omega) \) is the largest eigenvalue of the matrix \( \mathbf{R}_f(\omega) \) at \( \omega \). The algorithm is stable at all frequencies if \( \mu' \) satisfies

\[ 0 < \mu' < \frac{1}{\max_{\omega}\{\lambda_{f_{\text{max}}}(\omega)\}}. \]  
(3.48)
Chapter 3: Adaptive Filters Approach to 3D Sound Systems

If the condition (3.48) is satisfied, then as \( n \to \infty \), the rotated weight difference vector \( \mathbf{V}(n, \omega) \to \mathbf{0} \), and (on average) the adaptive weights at \( \omega \) approach their optimum solutions \( E\{\mathbf{W}(n, \omega)\} \to \mathbf{W}_{opt}(\omega) \). However, each individual coefficient \( W_i(n, \omega), i = 1, 2, \cdots, K L \) approaches its optimal value following an approximate exponential decaying envelope with time constant \([144]\)

\[
\tau_i(\omega) \approx \frac{1}{2\mu' \lambda_{f_i}(\omega)}.
\]  

(3.49)

The convergence speed of the whole system is determined by the time constant of the slowest mode over the whole frequency range of interest, which is given by [38, 144],

\[
\tau_{\max} > \frac{\max \{\lambda_{f_{\max}}(\omega)\}}{2 \min \{\lambda_{f_{\min}}(\omega)\}}.
\]  

(3.50)

Therefore, the convergence speed of the time domain adaptive weights is subject to the ratio of the maximum to minimum eigenvalues of the matrix \( \mathbf{R}_f(\omega) \) over the whole frequency range of interest. Since both \( \mathbf{X}(\omega) \) and \( \mathbf{C}(\omega) \) contribute to \( \mathbf{R}_f(\omega) \), the convergence speed of the MEFX algorithm is affected by both the input signals’ statistics and the matrix \( \mathbf{C}(\omega) \). The dependency of the convergence speed of the conventional LMS algorithm on the statistics of the input signal is well known in the traditional adaptive filters literature such as [144]. The effects of \( \mathbf{C}(\omega) \) and their physical interpretations are discussed in the next section.

### 3.3.3 Effects of the Electro-Acoustic Transfer Functions

The frequency domain matrix \( \mathbf{X}_f(n, \omega) \) defined in (3.40) may be expressed as

\[
\mathbf{X}_f(n, \omega) = \mathbf{C}(\omega) \mathbf{X}(n, \omega),
\]  

(3.51)

where \( \mathbf{C}(\omega) \) is the matrix of electro-acoustic transfer functions and \( \mathbf{X}(n, \omega) \) is the Fourier transform of \( x^T(n) \) defined in (3.17), both evaluated at \( \omega \). To decouple the effects of \( \mathbf{X} \) and \( \mathbf{C} \) and study the effects of \( \mathbf{C}(\omega) \) only, the input signals \( \{x_k(n) : k = 1, 2, \cdots, K\} \) are assumed white noise with zero mean and variances \( \sigma^2_1, \sigma^2_2, \cdots, \sigma^2_K \) that are uncorrelated with each other and with \( \mathbf{C}(\omega) \). From (3.51), the matrix \( \mathbf{R}_f(\omega) \) may be written as \( \mathbf{R}_f(\omega) = E\{\mathbf{X}_f^H(n, \omega) \mathbf{C}^H(\omega) \mathbf{C}(\omega) \mathbf{X}(n, \omega)\} \). The determinant of \( \mathbf{R}_f(\omega) \) can then be expressed as [38, 39]

\[
\det{\mathbf{R}_f(\omega)} = |\mathbf{R}_f(\omega)| = N_B^2 \sigma^2_1 \sigma^2_2 \cdots \sigma^2_K |\mathbf{C}^H(\omega) \mathbf{C}(\omega)|^K.
\]  

(3.52)

where \( N_B \) is the number of frequency domain filter coefficients (the size of the DFT). Since the determinant of a matrix is equal to the product of its eigenvalues, it is clear from (3.52) that the eigenvalues of \( \mathbf{R}_f(\omega) \) are influenced by the eigenvalues of \( \mathbf{C}^H(\omega) \mathbf{C}(\omega) \). Furthermore, the smallest eigenvalue of \( \mathbf{R}_f(\omega) \) is determined only by the smallest eigenvalue of \( \mathbf{C}^H(\omega) \mathbf{C}(\omega) \) at \( \omega \), or equivalently by the determinant of \( \mathbf{C}^H(\omega) \mathbf{C}(\omega) \) at that frequency. The slowest mode is, therefore, completely determined by the lowest eigenvalue of \( \det{\mathbf{C}^H(\omega) \mathbf{C}(\omega)} \) over the whole frequency range of interest.

To assist understanding the physical meaning of the above mentioned result, we consider the system \( [K \times L \times M = 1 \times 1 \times 1] \). In this case, both \( \mathbf{C}(\omega) \) and \( \mathbf{R}_f(\omega) \) are \( [1 \times 1] \) matrices and there is only one eigenvalue \( \lambda_{f_i}(\omega) \) at each frequency. This eigenvalue can be evaluated from \( \lambda_{f_i}(\omega) = \det{\mathbf{C}^H(\omega) \mathbf{C}(\omega)} = |\mathbf{C}(\omega)|^2 \). From (3.47), and assuming white noise input of zero mean and variance \( \sigma^2 \), the stability condition becomes

\[
0 < \mu' < \frac{1}{\max_{\omega} \{ |\mathbf{C}(\omega)|^2 \} N_B \sigma^2}.
\]  

(3.53)

This shows that the step size \( \mu' \) is limited by the maximum power gain of the single transfer function between the loudspeaker and the microphone. From (3.49), the time constant \( \tau(\omega) \) becomes

\[
\tau(\omega) \approx \frac{1}{2\mu' |\mathbf{C}(\omega)|^2}.
\]  

(3.54)

which shows clearly that the smallest power gain defines the slowest mode. In an extreme case, when \( |\mathbf{C}(\omega)| = 0 \) at any frequency \( \omega \), the time constant at that frequency is infinitely long and the adaptive filter will never reach its optimal solution.

In multichannel systems, the influence of \( \mathbf{C}(\omega) \) is enhanced due to the interaction between the different electro-acoustic transfer functions comprising \( \mathbf{C}(\omega) \). To show this, consider the system \( [K \times L \times M = 2 \times 2 \times 2] \), where

\[
\det{\mathbf{C}^H(\omega) \mathbf{C}(\omega)} = (C_{11} C_{22} - C_{21} C_{12}) (C_{11} C_{22} - C_{21} C_{12}).
\]  

(3.55)

As mentioned above, whenever the determinant in (3.55) equals zero (or a very small value) at any frequency, the whole adaptive process is slowed. From (3.55), the following three cases which result in \( \det{\mathbf{C}^H(\omega) \mathbf{C}(\omega)} \to 0 \) may be recognised:

1. \( C_{11}(\omega) = C_{22}(\omega) = 0 \) or \( C_{22}(\omega) = C_{21}(\omega) = 0 \). This occurs when all transfer functions from all loudspeakers to any of the microphones are zeros at the same frequency, leading to a zero row in the matrix \( \mathbf{C}(\omega) \). This effectively reduces the number of microphones at this frequency and reduces the rank of the matrix \( \mathbf{C}(\omega) \) at that specific frequency. This situation may occur in practice when the frequency responses of all loudspeakers have a notch at the same frequency, or the concerned microphone is insensitive at a given frequency.

2. \( C_{11}(\omega) = C_{21}(\omega) = 0 \) or \( C_{22}(\omega) = C_{12}(\omega) = 0 \). This occurs when all transfer functions from any of the loudspeakers to all microphones are zeros at the same frequency leading to a zero column in the matrix \( \mathbf{C}(\omega) \). In such a situation, the related loudspeaker is not effective at that frequency, which also reduces the rank of \( \mathbf{C} \) by one at that specific frequency. Similar practical causes as those mentioned above may also be valid in this case.
3. Chapter 3: Adaptive Filters Approach to 3D Sound Systems

3.3.4 The Causality Constraint

Reducing the error signals $d(n)$ in Fig. 3.2 to zeros makes the actual system outputs $\hat{d}(n)$ equal to the desired responses $d(n)$. Since the inputs $x(n)$ to the cascade combination of $W(\omega)$ and $C(\omega)$ are the same as the inputs to the system $H(\omega)$, the control problem may be considered as adjusting $W(\omega)$ so that

$$C(\omega) W(\omega) = H(\omega), \quad (3.56)$$

where the dimensions of $C(\omega)$, $W(\omega)$, and $H(\omega)$ are $[M \times L]$, $[L \times K]$ and $[M \times K]$, respectively. This shows that $W(\omega)$ is a combination of forward modelling of $H(\omega)$ and inverse modelling of $C(\omega)$. The success of this modelling problem is, in general, limited by the following factors:

- Being physical systems, the elements of $C(\omega)$ correspond to causal impulse responses. The input signals to the loudspeakers $\mathbf{y}(n)$ are delayed as they go through the physical system $C(\omega)$. Assuming for the moment that $H(\omega)$ does not introduce any delay, the desired responses $d(n)$ arrive at the microphones before $\hat{d}(n)$. This requires the filters $W(\omega)$ to be predictors, a task that can only be performed approximately by a causal filter in a statistical sense [144]. To avoid this prediction solution, $H(\omega)$ must have delay components that are longer than the delays introduced by $C(\omega)$.

- The elements of $C(\omega)$ are often non-minimum phase impulse responses and their transfer functions have zeros outside the unit circle in the z-plane. Inverting such a non-minimum phase response results in poles outside the unit circle. Such an inverse is stable only if the impulse response is left-handed in time (non-causal). Since a delayed non-causal impulse response may be approximated by a causal impulse response truncated in time [144], a delayed version of the non-causal solution may be approximated by $W(\omega)$. This delay may be introduced by delaying the desired response $d(n)$.

3.3.5 Adaptive Active Noise Cancellation

The detailed implementation of the MEFX algorithm is further explained by unpacking the composite vector $\mathbf{w}(n)$ in (3.36) into its individual filters $\{\mathbf{w}_{lk}(n) : l = 1, 2, \cdots, L, k = 1, 2, \cdots, K\}$ giving

$$\mathbf{w}_{lk}(n + 1) = \mathbf{w}_{lk}(n) + 2 \mu \sum_{m=1}^{M} \epsilon_{m}(n) \mathbf{x}_{l m k}(n), \quad (3.57)$$

where $\mathbf{x}_{l m k}(n)$ is the result of filtering the $k^{th}$ input through the measured electro-acoustic impulse response $c_{ml}$ between the $m^{th}$ microphone and the $l^{th}$ loudspeaker. The detailed implementation of (3.57) is visually illustrated in Fig. 3.3 for a $[K \times L \times M = 2 \times 2 \times 2]$ active noise control system. In this system, it is desired to reduce the sound field due to the primary sources $P_1$ and $P_2$ at two receiver points (microphones) $R_1$ and $R_2$ using an anti-sound field generated by the secondary loudspeakers $S_1$ and $S_2$. Using frequency domain representations, the inputs to the secondary sources $Y_1(\omega)$ and $Y_2(\omega)$ are controlled by four adaptive filters $\{\mathbf{w}_{lk}(\omega) : l = 1, 2, k = 1, 2\}$ to achieve the above cancellation task. The desired response is the sound field generated by the primary sources $P_1$ and $P_2$ at the microphones' positions: $\mathbf{d}(\omega) = [D_1(\omega) \ D_2(\omega)]$ = $H(\omega) \ \mathbf{x}(\omega)$, where $H(\omega)$ is the matrix of electro-acoustic transfer functions between the two microphones and the primary sources and $\mathbf{x}(\omega) = [X_1(\omega) \ X_2(\omega)]$. The microphones' outputs are, therefore, the sum of the sound fields due to the primary and the secondary sources $E(\omega) = [E_1(\omega) \ E_2(\omega)] = \mathbf{d}(\omega) + \mathbf{d}(\omega)$, where $\mathbf{d}(\omega) = [D_1(\omega) \ D_2(\omega)] = C(\omega) \ \mathbf{x}(\omega)$ is the (unmeasurable) sound field due to the secondary sources alone at the microphones.

At each time sample, the update algorithm adjusts the coefficients of the adaptive filters to minimise the error signals measured by the microphones using the MEFX algorithm.
Chapter 3: Adaptive Filters Approach to 3D Sound Systems

3.3 Iterative LMS Solutions

3.3.6 Adaptive Virtual Sound Sources

A comparison between (3.58) and (2.15) shows that the system in Fig. 3.3 is exactly (except for a − sign) what is needed to generate two virtual sound images at the positions of the primary sources $P_1$ and $P_2$ at the ears of one listener when the two microphones are positioned inside the listener’s ear canals. By controlling the sound at $M = 2B$ microphones, the same sound images are perceived by B listeners, and by increasing the number of inputs, more images are created. This suggests a procedure for implementing the loudspeaker display system discussed in Section 2.2.2. With probe microphones inserted in the listener’s ear canals, the set of adaptive filters are adjusted to cancel uncorrelated white noise signals from physical loudspeakers placed where virtual sound images are required\(^2\). After successful conversion, the physical primary sources are disconnected and the coefficients of the filters are multiplied by $-1$ to compensate for the sign difference mentioned above. Monophonic signals filtered through the previously obtained filters will move the auditory event to the positions where the primary sources have been. This procedure has been successfully used in [2, 3, 83, 100, 131] and proved to have the following advantages compared to other methods:

- The listener’s own HRTFs are used to design the filters, therefore, correct spectral cues are maintained.
- The solution also includes the room impulse response, therefore, maintains the distance cues and the environmental context. This solves the In-Head-Localisation (IHL) problem and eliminates the need for artificial reverberation.
- Direct inversion of the electro-acoustic transfer function matrix $C(\omega)$ that is required for the cross-talk cancellation is avoided by iteratively searching the optimum solution in the least mean squared sense.
- The binaural synthesis (convolution with the HRTF matrix $H(\omega)$) and the cross-talk cancellation (inversion of $C(\omega)$) are combined. This is both more efficient and numerically more stable.
- The problem is mapped from the difficult domain of virtual sound image synthesis to the well-developed one of multichannel active noise control (ANC). This not only allows employing the techniques used in ANC systems, but also facilitates describing the system performance in terms of the sound attenuation achieved at the listener’s eardrums. The larger this attenuation, the more realistic the virtual source is perceived.
- Active noise control systems are real-time systems, therefore, the above procedure allows real-time design and implementation of the system’s filters. This differs from the commonly used methods of measuring and storing a set of

\[^2\text{Since (3.57) was derived assuming that the error is formed by the difference between rather than the sum of } D(\omega) \text{ and } D(\omega), \text{ the } + \text{ sign in (3.57) must be changed to } - \text{ sign.} \]

\[^3\text{The adaptation process requires first measuring all electro-acoustic transfer functions } \hat{Q}_{m1}. \text{ Online estimation of those transfer functions is discussed in detail in Section 4.7.} \]
3.4 Efficient Implementations

HRTFs, calculating the cross-talk cancellation filters, and real-time filtering the audio signals through those previously calculated fixed filters.

In spite of the above mentioned advantages, adaptive filters approach to virtual source synthesis suffers from the following drawbacks:

- Including the room impulse response in the filters limits the virtual space to the same measurement environment.
- The listener is asked to insert a pair of microphones inside his/her ears, which may be objectionable. However, this is the only approach to obtaining individualised HRTFs.
- White noise or chirp signals that are spectrally rich must be used in the identification and adaptation stages.
- The electro-acoustic transfer functions $C(\omega)$ and $H(\omega)$ are very complex functions of frequency and space coordinates. The impulse response of a room may also last for several hundreds of milliseconds. At a sampling frequency of 44.1 kHz, thousands of FIR coefficients are required to properly model these transfer functions. The adaptive filters are, therefore, of high order, which makes system implementation in real-time a great challenge. Efficient implementation of the filtering and adaptation operations are, therefore, essential.
- At high frequencies, the acoustic wavelength is very small. Therefore, the solution obtained by the adaptive process is valid only in a very small area in space. Therefore, the listener's head must be fixed during the adaptation and filtering operations.

3.3.7 Adaptive Cross-Talk Cancellation

The cross-talk cancellation mentioned in Section 2.2.2 may also be realised using adaptive filters. The system shown in Fig. 3.3 can be used for this purpose after a few modifications. In cross-talk cancellation, it is desired to reproduce $x_1(n)$ at $R_1$ and $x_2(n)$ at $R_2$. The desired response is, therefore, given by $d(n) = x(n)$, and the adaptive filters are required to model the inverse of the matrix $C(\omega)$. For the inverse to be realisable using FIR filters, delayed input signals are used to calculate the desired responses as mentioned in Section 3.3.4.

\[
\begin{bmatrix}
   d_1(n) \\
   d_2(n)
\end{bmatrix} =
\begin{bmatrix}
   x_1(n - \Delta_1) \\
   x_2(n - \Delta_2)
\end{bmatrix},
\]

where $\Delta_1$ and $\Delta_2$ are delays that are assumed longer than that introduced by the electro-acoustic transfer functions comprising $C(\omega)$. The mean square error $J$ is minimised by the adaptive algorithm. After converging, the sound pressure at the microphones $R_1$ and $R_2$ are the best least squares estimations of $x_1(n - \Delta_1)$ and $x_2(n - \Delta_2)$, respectively.

3.4 Efficient Implementations

In reverberant acoustic environments, the impulse response between two points may last for several hundreds of milliseconds. At the standard audio compact disc sampling frequency of 44.1 kHz, thousands of FIR coefficients are needed to properly model and store such an impulse response. Adaptive 3D sound systems requires estimating, filtering through, and updating the coefficients of many of these large filters. This implies a huge amount of computations, which makes real-time implementation on reasonable hardware resources a difficult task. Reducing the system complexity is, therefore, essential for real-time implementation. In this section, two approaches are introduced to decrease the number of required calculations. The adjoint LMS algorithm discussed in Section 3.4.1 uses filtered error signals rather than filtered input signals in the update process. This leads to a huge computational saving as the system dimensions increase. Significant saving may further be achieved by implementing the convolution and correlation operations in the frequency domain using block processing techniques. This leads to the Block Frequency Domain Adaptive Filters (BFDAF) discussed in Section 3.4.2.

3.4.1 The Adjoint LMS Algorithm

Setting the electro-acoustic transfer function matrix $C(\omega)$ in Fig. 3.2 at each frequency to be the identity matrix $I$ results in a regular adaptive filter system [144]. In this case, $d(n) = y(n)$ and the adaptive filters' outputs are directly observable from the error vector $g(n) = d(n) - y(n)$. The steepest descent update given by (3.36) becomes

\[
\mathbf{w}(n + 1) = \mathbf{w}(n) + 2\mu \mathbf{x}^T(n) g(n),
\]

where $\mathbf{x}(n)$ is defined in (3.17). The adaptive process in (3.61) basically aims at minimising the crosscorrelation between the input signals $\mathbf{x}(n)$ and the error signals.
Chapter 3: Adaptive Filters Approach to 3D Sound Systems

Figure 3.4: Block diagram of the adjoint least mean square algorithm.

When the crosscorrelation reaches its minimum value, the filters are as close as possible to their optimal solutions. For correct adaptation, it is essential that the crosscorrelation is performed between correctly aligned history samples of x(n) and y(n).

Introducing the matrix \( \mathbf{C}(\omega) \) into the system results in \( \mathbf{d}(n) \) being a filtered version of \( \mathbf{y}(n) \). This has two severe consequences on the adaptation process. The first is that \( \mathbf{y}(n) \) is delayed while propagating through the matrix of acoustical transfer function \( \mathbf{C}(\omega) \), which leads to an unstable adaptive algorithm. The second is the spectral deformation caused by this case, an unobservable system at that frequency.

The effect of the latter may be explained by the extreme case of \( \mathbf{C}(\omega) \) containing no information about the filters' outputs at \( \omega_0 \). In this case, \( e_m(n) \) contains no information about the filters' outputs at \( \omega_0 \), resulting in an unobservable system at that frequency.

A stable adaptive algorithm with \( \mathbf{C}(\omega) \neq \mathbf{I} \) may only be obtained if the above mentioned filtering effects are compensated. The MEFX algorithm discussed in Section 3.3.1 solves this problem by filtering \( \mathbf{x}(n) \) through estimates of \( \mathbf{C}(\omega) \) prior using them as inputs to the adaptation process as shown in Fig. 3.3. This effectively delays the inputs to be correctly aligned in time with the error signals, and at the same time introduces the required spectral correction. Since the dimensions of matrices may not match and matrices do not commute, the MEFX effectively filters every input signal \( \mathbf{x}_f(n) \) with every electro-acoustic response \( e_m \) to construct the KLM elements of the matrix \( \mathbf{x}_f(n) \). This requires in total KLM convolution operations only for constructing \( \mathbf{x}_f(n) \). Assuming time domain implementation, the total number of multiplications required per iteration for calculating \( \mathbf{x}_f(n) \) is \( KLMN_c \). Updating the weights requires \( KLMN_c (M + 1) \) multiplications, and calculating the filters' outputs requires \( KLMN_w \) multiplications. The total number of multiplications required to implement the MEFX algorithm each iteration is, therefore,

\[
\Psi_{MEFX} = KL \left[ (N_w + N_c) M + 2N_w \right],
\]

which increases rapidly with increasing system dimensions. A much more efficient algorithm may be obtained by advancing the error signals \( e(n) \) in time to be aligned with \( x(n) \). This leads to the Adjoint Least Mean Square (ALMS) algorithm [141], which has its update equation given by

\[
\mathbf{w}(n + 1) = \mathbf{w}(n) + 2\mu \mathbf{x}^T(n - N_c + 1) \mathbf{e}(n - N_c + 1),
\]

where \( \mathbf{e}(n) = [ e_{f_1}, e_{f_2}, \ldots, e_{f_L} ]^T \) is the result of filtering the error signals through the adjoint (time mirror) of the transfer functions comprising \( \mathbf{C}(\omega) \). This non-causal operation can only be performed in real-time after a delay of \( N_c - 1 \) samples, the reason for using delayed input and filtered error signals in (3.63). The time mirror operation is equivalent to calculating the complex conjugate of the frequency response as shown in Fig. 3.4. Expressing the adjoint of \( e_m \) as

\[
\hat{e}_m = [ e_{m,N_c-1}, e_{m,N_c-2}, \ldots, e_{m,1}, e_{m,0} ]^T,
\]

the filtered error signal \( e_f(n) \) may be written as by

\[
e_f(n - N_c + 1) = \hat{e}_{f_1}(n) + \hat{e}_{f_2}(n) + \cdots + \hat{e}_{f_M}(n).
\]

The ALMS algorithm reduces the number of multiplications required for updating the adaptive weights to \( 2KLMN_w \). Adding \( LMN_c \) multiplications for calculating \( e_f(n) \), and \( KLMN_w \) multiplications for calculating the filters' outputs sums to

\[
\Psi_{ALMS} = KL \left[ (N_w + N_c) M + 3N_w \right].
\]

Comparing (3.66) with (3.62) shows that for a \( [K \times L \times M = 1 \times 1 \times 1] \) system, the ALMS and MEFX have the same complexity of \( (3N_w + N_c) \) multiplications. For multichannel systems, however, the ALMS algorithm is much more efficient than the MEFX as shown in Fig. 3.5, which gives the computational saving per iteration \( (\Psi_{MEFX} - \Psi_{ALMS}) \) against the adaptive filter order \( N_w \) for different system configurations.

Figure 3.5: Computational saving per iteration \( (\Psi_{MEFX} - \Psi_{ALMS}) \) against the adaptive filter order \( N_w \) for different system configurations.
Chapter 3: Adaptive Filters Approach to 3D Sound Systems

3.4 Efficient Implementations

The first two tasks are convolution operations between $\mathbf{x}(n)$ and FIR filters comprising $\mathbf{W}(\omega)$ and $\mathbf{C}(\omega)$, respectively. The third task implies calculating the instantaneous gradient that is essentially a crosscorrelation between the input and error signals. Since all FIR filters are of high order, implementing the convolution and correlation in the frequency domain results in considerable computational saving [114, 134]. Since the input signals are infinitely long, real-time implementation of the convolution or correlation in the frequency domain must be performed on successive overlapping blocks of data. Two known, and frequently used, methods to construct and process such overlapping blocks of data are the overlap-add and overlap-save methods [114]. Applying any of those block processing techniques to the LMS adaptive process leads to the Block Frequency Domain Adaptive Filter (BFDAF) [41, 55, 98, 106, 130].

Besides the speed gained by carrying out the convolution and correlation operations in the frequency domain, BFDAF offers the possibility of approximate decorrelation of the adaptive process from the input signals’ statistics. This is done by normalising each frequency bin by the input power in that bin [130]. Unlike the systems discussed previously which update the coefficients of the adaptive filters every sample, BFDAF performs such update every $L \geq 1$ samples, which leads to lower complexity, but at the same time slower convergence rate. Due to block processing, reduction in the step size bound and processing delay of $L$ samples are also introduced. In this section, implementations of 3D sound systems using BFDAF are considered. Since the adjoint LMS algorithm is more efficient for multichannel systems as shown in Section 3.4.1, only BFDAF implementation of that algorithm is considered below.

For clarity, the BFDAF implementation of the ALMS algorithm is illustrated for a single channel active noise control system, since the extension to multichannel systems is straightforward as given by (3.63). The block diagram of the ALMS algorithm implemented using overlap-save BFDAF for $[K \times L \times M = 1 \times 1 \times 1]$ ANC system is shown in Fig. 3.7. In this block diagram, vectors are represented by double parallel lines while scalars are represented by single lines. Since we are dealing with a single channel case, vectors are used here to represent frequency samples of the signals rather than multiple signals at the same frequency.

The overlap-save method requires sectioning the stream of input time samples $\mathbf{x}(n)$ into overlapping blocks, each of length $N_B$. Each block contains $L$ new samples and $N_w + N_c - 2$ samples from the previous block. The sectioning and overlapping operations are represented in Fig. 3.7 by the serial-to-parallel and overlap blocks. Each $L$ samples, a complete block of time samples $\mathbf{x}_B(nL)$ is constructed and processing of this block of data may be commenced. For correct real-time operation, the processing of each block must be completed before the next block is constructed, i.e. in maximum $L$ samples. To perform convolution and correlation in the frequency domain, each
of this time domain vector represent the required linear convolution. During processing the next block, these L samples are sent, one sample each clock cycle, to drive loudspeaker S. This is represented in Fig. 3.7 by the parallel-to-serial block. The first $N_w - 1$ samples represent cyclic convolution results and must be discarded. An extra $N_c - 1$ samples represent correct linear convolution results but are not used since only L samples are needed to be sent out every block. This extra $N_c - 1$ samples are due to the choice of a large block length $N_B$ to accommodate successive convolution and correlation as described below.

2. **Calculation of the filtered error $E_f([n - 1]L)$:** The microphone signal $e(n - L + 1)$ is buffered into a length L vector. This vector is augmented with $N_B - L$ leading zeros and transformed to the frequency domain using an FFT of length $N_B$, resulting in the block error vector $E([n - 1]L)$. A previously measured impulse response $\hat{c}$ of length $N_c$ between loudspeaker S and the microphone is padded with $N_B - N_c$ trailing zeros and transformed to the frequency domain using an FFT of length $N_B$, producing the vector $\hat{C}$. The adjoint (time mirror) operation required for the ALMS is performed by calculating the complex conjugate of $\hat{C}$. The filtered error vector $E_f([n - 1]L)$ is then calculated by element-wise multiplication of $\hat{C}^*$ and $E([n - 1]L)$. The length of the result of this convolution is $L + N_c - 1$, and since $N_B$ is chosen larger than this length, all samples of $E_f([n - 1]L)$ represent correct linear convolution results. Defining the matrix $C = \text{diag}(\hat{C})$, this convolution may be expressed mathematically as

$$E_f([n - 1]L) = C^* E([n - 1]L).$$

Since the microphone signal $e(n - L + 1)$ is due to output samples $y(n - L + 1)$ from the previous block, the error signal is already delayed by L samples. Provided that $L \geq N_c$, this delay will suffice to implement the delay required by the ALMS as mentioned in Section 3.4.1. To further ensure that the optimum filter solution is causal, the delay in the path through loudspeaker P must be longer than that through loudspeaker S. Since the algorithm introduces a delay of $L$ samples, this delay must be compensated. Assuming that the acoustic transfer function $H$ possesses longer delay than $C$, this compensation may be done by delaying the signal to loudspeaker P by $L$ samples. This is represented in Fig. 3.7 by the box $z^{-L}$. An elegant alternative is to move loudspeaker P an equivalent distance away from the microphone to save DSP memory.

3. **Updating the filter weights:** Transforming the update equation of the ALMS (3.63) to the frequency domain results in the following block frequency domain update equation\(^7\)

$$W([n]L) = W([n - 1]L) - \frac{2\mu}{L} GP^{-1} X^*([n - 1]L) E_f([n - 1]L),$$

---

\(^7\)Since signals are summed up at the microphones in Fig. 3.7 rather than subtracted, the + sign in (3.63) has been changed to a − sign in (3.70).
where it is assumed in this equation that $L \geq N_c$, the delay required by the ALMS algorithm. The diagonal matrix $\mathbf{P}$ represents an estimate of the input signal power in each frequency bin and performs the decorrelation mentioned above. The inverse of this power matrix is scaled by the step size $2\mu / L$ and the result is used as the new step size. This frequency-dependent step size vector is used to scale the estimated gradient vector. The instantaneous block mean square gradient estimate vector given by the crosscorrelation $\Sigma^{(n-1)\mathbf{L}} = \mathbf{X}^\ast(n-1)\mathbf{L} \mathbf{E}_f(n-1)\mathbf{L}$ is obtained by first delaying $\mathbf{X}(n\mathbf{L})$, calculating its complex conjugate, and element-wise multiplying the result by the block filtered error as shown in Fig. 3.7. Since $\mathbf{X}^{(n-1)\mathbf{L}}$ is not padded with zeros, this element-wise multiplication corresponds in the time domain to $N_w$ linear correlation coefficients, while the last $N_B - N_w$ samples are the cyclic correlation coefficients and must be discarded. However, the correct correlation coefficients must be extracted in the time domain, which is done using the $N_B \times N_B$ constraining window $\mathbf{G}$. This window transforms the estimated gradient to the time domain, replaces the last $N_B - N_w$ samples by zeros, and transforms the result back to the frequency domain. Finally, the weighted and constrained gradient is used to update the weight vector $\mathbf{W}^{(n-1)\mathbf{L}}$ according to (3.70).

The constraining window $\mathbf{G}$ that costs two FFTs may be dropped if extra computational saving is needed. In this case, the algorithm is referred to as the unconstrained BFDAF [98, 130]. This leads to using wrong cyclic correlation coefficients in the gradient vector, which in turn leads to a slower convergence rate and a reduction in the stable range of the step size [98]. Although the unconstrained filter corresponds to $N_B$ coefficients in the time domain, the 'wrap-around error' prevents it from performing as a filter of more coefficients than $N_w$.

Finally, it is worth mentioning that the frequency domain implementation of a single channel system using the adjoint LMS may lead to more computational saving over that using the filtered-$x$ algorithm if $N_w$ is much larger than $N_c$. This second source of computational saving stems from the fact that the filtered-$x$ algorithm calculates $\mathbf{x}_f$ by filtering the infinitely long input signal through the finite impulse response $\mathbf{C}$. Implementing this convolution in the frequency domain requires the use of an overlapping method, and possibly two extra FFTs to separate the correct convolution samples in the time domain. On the other hand, the frequency domain adjoint LMS algorithm calculates $\mathbf{E}_f$ by filtering (the already in block form) error signal through $\mathbf{C}^\ast$ and no overlapping is needed. BFDAF implementations of the single channel adjoint LMS and several implementations of the single channel filtered-$x$ have already been presented in previous works [3, 83, 100, 131, 140]. A complexity comparison between these implementations shows that for a single channel BFDAF, the filtered-$x$ is more efficient than the adjoint LMS when $N_w < 2 N_c$, while the adjoint LMS is much more efficient when $N_w > 2 N_c$ [3, 131].

Chapter 4

Robustness of 3D Sound Systems

The solution for the control filters in the generalised audio reproduction model discussed in Section 3.1 consists of two elements. The first is the cross-talk canceller, which is the inverse of the matrix of electro-acoustic transfer functions between the reproduction loudspeakers and the listeners' ears. The second is the matrix of electro-acoustic transfer functions representing the desired response. Since both elements are dependent on space coordinates, the solution for the control filters is inherently dependent on space coordinates, independent of the method used to calculate the filters. A set of filters designed to give optimum results at a specific listening situation is, in general, in error when a listener moves. In this chapter, the nature of those errors is discussed and methods to improve the system robustness against listeners' movements are introduced. The problem is divided into two subproblems: the first deals with enlarging the zones of equalisation created around the control microphones and the second tracks the listeners when they move beyond those equalisation zones.

In Section 4.1, the physical nature of the errors introduced in the control filters when a listener moves is discussed. These errors are not only position dependent, but also frequency dependent. Small movements make the control filters slightly in error at low frequencies while the errors increase with increasing frequency. This motivates using control filters having non-uniform frequency resolution to cope with this frequency dependency.

The equalisation zones may be enlarged by designing the control filters to be valid not only at the eardrums of the listeners but also at points in space in the vicinity of the listeners' ears. This is achieved by constraining the derivative of the sound field with respect to the space coordinates in the vicinity of the listeners' eardrums to be zero. The implementation of derivative constrained filters requires measuring several transfer functions in the vicinity of the listeners' ears and designing the filters using a weighted average of these. Methods for designing the filters to meet this constraint are discussed in Section 4.2. These methods achieve their goal by forcing the filters to assume solutions that increasingly deviate from their exact solutions with increasing frequency, which is a second motivation for using control filters having non-uniform frequency resolution. By decreasing the frequency resolution of the control filters as the frequency increases, the filters ignore more details with increasing frequency. This
Chapter 4: Robustness of 3D Sound Systems

makes the filters valid in a wider area in space without the need to measure several transfer functions at adjacent points. Multiresolution control filters are discussed in Section 4.3 and in more details in Chapter 5.

The zones of equalisation can further be improved by using more reproduction loudspeakers to increase system controllability. This is achieved only if the reproduction loudspeakers are correctly positioned in the listening space. Improving system controllability by employing more reproduction loudspeakers and the choice of the positions of those loudspeakers are discussed in Section 4.4 and Section 4.5, respectively.

Large errors are introduced in the control filters when the listeners move beyond the equalisation zones. In case of such large movements, the filters must be recalculated using the transfer functions corresponding to the listeners' new positions. In adaptive 3D sound systems, this may be achieved by readapting the control filters to their new optimal solutions. This in turn requires measuring all electro-acoustic transfer functions between the reproduction loudspeakers and the listeners' ears that are needed for the filtered-x and adjoint LMS adaptive algorithms discussed in Chapter 3. Large movements and head tracking are discussed in Section 4.6, while on-line estimation of the electro-acoustic transfer functions and on-line adaptation of the control filters are discussed in Section 4.7.

Essential for adaptive 3D sound systems is that the adaptive filters reach their new optimum solutions after a listener movement fast enough so that the listener does not loose the sound image. Since the above mentioned adaptive algorithms are known for their slow convergence speed, this requires improving the convergence properties of these algorithms, which is addressed in Section 4.8. Finally, Section 4.9 summarises the techniques introduced in this chapter.

4.1 The Robustness Problem

In the general sound reproduction model shown in Fig. 3.1, the \([L \times K]\) matrix of control filters \(W(\omega)\) that minimises the error vector \(E(\omega)\) for a given desired response \(D(\omega)\) is obtained by solving the system of control equations

\[
C(\omega) W(\omega) = H(\omega),
\]

where \(C(\omega)\) is the \([M \times L]\) matrix of acoustic transfer functions between the reproduction loudspeakers \(\{S_1, S_2, \ldots, S_L\}\) and the microphones \(\{R_1, R_2, \ldots, R_M\}\) at \(\omega\). The \([M \times K]\) matrix \(H(\omega)\) is used to generate the desired response vector

\[
D(\omega) = H(\omega) X(\omega).
\]

Regardless of the method used to calculate the filters \(W(\omega)\), it is clear from (4.1) that the solution is dependent only on the transfer functions \(\{C_{mi}(\omega)\} : m = 1, 2, \ldots, M, i = 1, 2, \ldots, L\) and \(\{H_{mk}(\omega)\} : m = 1, 2, \ldots, M, k = 1, 2, \ldots, K\). Each of these transfer functions is a cascade combination of several components. These components include:

- The impulse responses of the reconstruction filter and the Digital-to-Analogue Converter (DAC) used to transform \(y(n)\) and \(x(n)\) to continuous-time form.
- The impulse responses of the power amplifiers used to drive the loudspeakers.
- The impulse responses of the reproduction loudspeakers \(\{S_1, S_2, \ldots, S_L\}\) for \(C(\omega)\) and the primary loudspeakers \(\{P_1, P_2, \ldots, P_K\}\) for \(H(\omega)\). For some applications, such as cross-talk cancellation, the primary loudspeakers do not physically exist and their responses may be considered as delta functions.
- The room impulse responses between each of the above mentioned loudspeakers and each microphone.
- The impulse responses of the microphones and their corresponding amplifiers.
- The Head-Related Transfer Functions (HRTFs) from each loudspeaker to each of listeners' ears.
- And the responses of the anti-aliasing filters and the Analogue-to-Digital Converter (ADC) that transform the microphones' signals to digital form.

From all these components, the head-related transfer functions and the impulse responses between points in the room are functions in space coordinates as mentioned in Chapter 2. Consequently, the optimal solution for \(W(\omega)\) is dependent on the position of the microphones relative to the loudspeakers. This position dependency is further frequency dependent due to the slow speed at which sound waves propagate in the air. To illustrate this latter statement, consider a 2 cm displacement in one of the microphones away from its original position. This displacement corresponds to a complete wavelength of a 17 kHz sound wave while it is just 0.01 wavelength of a 170 Hz wave. A matrix of filters \(W(\omega)\) designed for optimal solution at the original microphones' positions has higher amplitude errors at higher frequencies than at lower frequencies at the displaced microphone position.

The frequency dependent spatial dependency mentioned above is illustrated in Fig. 4.1. This result was obtained in a single channel active noise control experiment in an anechoic chamber [52, 53, 65]. In this experiment, white noise filtered by a band-pass filter having lower and higher cut-off frequencies of 0.1 and 4 kHz, respectively, is emitted from a primary loudspeaker. An adaptive filter is used to drive a secondary source to cancel the sound at a microphone placed 0.5 m from the secondary source. From Fig. 4.1, substantial decrease in the primary acoustic power is achieved at all frequencies at the microphone position. At points away from this control position, the reduction in primary acoustic power decreases. This decrease is frequency dependent; at the same point in space, a rapid decrease in the sound reduction is noticed as the frequency increases. At 2 cm from the control microphone, only about 5 dB reduction of primary acoustic power is achieved at 1 kHz, which is not satisfactory for any audio reproduction application. At the end of the audible frequency range (at 20 kHz), one expects that even a few millimetres displacement in one of the microphones...
Chapter 4: Robustness of 3D Sound Systems

10 Frequency [Hz]

Figure 4.1: The acoustic power spectrum measured at different points from the control microphone in a single channel active noise control experiment. Curve [1] is the primary noise to be reduced, curve [2] is the resulting acoustic power after cancellation at the microphone position, [3] is that at 2 cm from the microphone, and [4] is at 5 cm from the microphone.

or loudspeakers may cause enough errors that render the performance of the system not acceptable.

In general, loudspeakers and microphones also have directional characteristics, which further complicates the matter. In the present work, the loudspeakers and microphones are considered omnidirectional to simplify the discussion. An omnidirectional microphone produces the same output voltage for an incident sound wave regardless of the arrival angle of the sound. Similarly, an omnidirectional loudspeaker radiates equally strong sound waves in all directions.

4.2 Enlarging the Zones of Equalisation

The control task of the generalised model shown in Fig. 3.1 may be described by the system of control equations

\[ C(\omega, r, \alpha, \theta) \mathbf{W}(\omega) = \mathbf{H}(\omega, r, \alpha, \theta), \]

where \( C(\omega, r, \alpha, \theta) \) is the \([M \times L]\) matrix of electro-acoustic transfer functions between the \( M \) microphones and the \( L \) reproduction loudspeakers. \( \mathbf{H}(\omega, r, \alpha, \theta) \) is the \([M \times K]\) matrix of electro-acoustic transfer functions between the \( M \) microphones and the \( K \) input signals. \( \mathbf{W}(\omega) \) is the \([L \times K]\) matrix of control filters. Since (4.3) does not specify the sound field at points other than the control points (the positions of the microphones), direct solution of (4.3) results in control filters that accurately fulfil the control task only at the positions of the microphones. The sound field at points away from the control points is, therefore, uncontrollable by the filters. This leads to small equalisation zones around the microphones as mentioned above, which makes the system sensitive to changes in the microphones' positions. To enlarge the equalisation zones, the control filters must be designed to control the sound field not only at, but also in the vicinity of, the microphones. Extra constraints that specify the sound field in the vicinity of the microphones must, therefore, be added to the system of control equations (4.3). Several of such constraints are discussed in Sections 4.2.1 through 4.2.3. The performance of the discussed methods is examined by computer experiments in Section 4.2.4

4.2.1 Spatial Derivative Constraints

One approach to enlarge the equalisation zones is to design the control filters \( \mathbf{W}(\omega) \) such that the actual sound field \( \mathbf{D}(\omega) \) is constant and equal to the desired sound field \( \mathbf{D}(\omega) \) not only at the microphones but also in their proximity. This is equivalent to constraining the first derivative of \( C(\omega, r, \alpha, \theta) \mathbf{W}(\omega) \) with respect to space coordinates to zero. Further expansion of the control action in space may be obtained by constraining the first \( J > 1 \) spatial derivatives to zero [10]. Derivative constraints with respect to the vertical-polar coordinates allow designing the filters \( \mathbf{W}(\omega) \) while taking into account radial movements in the \( r \) direction, and head rotations in \( \alpha \) and \( \theta \) directions. Alternatively, the Cartesian coordinate system may be used, where derivative constraints minimise errors in the filters due to translations in the \( x \), \( y \), and \( z \) directions.

Spatial derivative constraints are applied by adding new control equations to the set in (4.3), each of which constrains the derivative of the sound field near the corresponding microphone to be zero. Singling out the control equation for the \( m^{th} \) microphone positioned at coordinates \( (r_m, \alpha_m, \theta_m) \) due to the \( k^{th} \) input signal in (4.3), and dropping the dependency on \( \omega \) and space coordinates for improved readability, we obtain

\[ C_{m1}W_{1k} + C_{m2}W_{2k} + \cdots + C_{mk}W_{lk} = H_{mk}. \]

The corresponding first-order spatial derivative constrained control equations with respect to \( r_m, \alpha_m, \) and \( \theta_m \) are

\[ \sum_{l=1}^{L} \frac{\partial C_{ml}(\omega, r_m, \alpha_m, \theta_m)}{\partial r_m} W_{lk}(\omega) = 0 \]

\[ \sum_{l=1}^{L} \frac{\partial C_{ml}(\omega, r_m, \alpha_m, \theta_m)}{\partial \alpha_m} W_{lk}(\omega) = 0 \]
Chapter 4: Robustness of 3D Sound Systems

The first-order spatial derivative constrained control equations for all microphones can, therefore, be written as

\[ C(w, r, a, B) \frac{\partial H(w, r, a, B)}{\partial \theta_m} = 0, \]

(4.7)

where \( C(w, r, a, B) \) and \( H(w, r, a, B) \) are given by

\[
C = \begin{bmatrix}
C_{11} & \cdots & C_{1L} \\
\frac{\partial C_{11}}{\partial r} & \cdots & \frac{\partial C_{1L}}{\partial r} \\
\ddots & \ddots & \ddots \\
\frac{\partial C_{M1}}{\partial r} & \cdots & \frac{\partial C_{ML}}{\partial r} \\
\end{bmatrix},
\]

(4.8)

Higher order spatial derivative constraints are applied by further adding the higher order derivative control equations and equating each to zero \[10\]. Implementation of the spatial derivative constrained filters requires analytic expressions for the electro-acoustic transfer functions \( C(w, r, a, B) \) for calculating the spatial derivatives. Such analytic expressions are readily calculated in free field listening conditions (with the listener absent), where the transfer functions are dependent only on the distance between the microphones and the reproduction loudspeakers. Analytical expressions may also be obtained for the direct sound and the first few reflections in cases of moderate reverberant conditions. As the reverberation level increases, it becomes more difficult to derive the required analytical expressions and approximations to the spatial derivatives must be employed. The performance of such approximate spatial derivative constrained filters is examined by computer simulations in Section 4.2.4.3.

4.2.2 Spatial Difference Constraints

As mentioned above, exact spatial derivative constraints may be applied in free field conditions, where analytical expressions for the electro-acoustic transfer functions \( C(w, r, a, B) \) for calculating the spatial derivatives. Such analytic expressions are readily calculated in free field listening conditions (with the listener absent), where the transfer functions are dependent only on the distance between the microphones and the reproduction loudspeakers. Analytical expressions may also be obtained for the direct sound and the first few reflections in cases of moderate reverberant conditions. As the reverberation level increases, it becomes more difficult to derive the required analytical expressions and approximations to the spatial derivatives must be employed. The performance of such approximate spatial derivative constrained filters is examined by computer simulations in Section 4.2.4.3.

Consider adding three extra microphones at coordinates \((r_m + \Delta r_m, \alpha_m, \theta_m)\), \((r_m, \alpha_m, \theta_m + \Delta \theta_m)\), and \((r_m, \alpha_m + \Delta \alpha_m, \theta_m)\) in the proximity of the original \(m^{th}\) control point located at \((r_m, \alpha_m, \theta_m)\), where \(\Delta r_m\), \(\Delta \alpha_m\), and \(\Delta \theta_m\) are small displacements from the original position of the \(m^{th}\) microphone in the corresponding space coordinate directions. The control equations for those four microphones due to the \(k^{th}\) source can be written as

\[
\sum_{l=1}^{L} C_m(l, r_m, r_m + \Delta r_m, \alpha_m, \theta_m) W_{lk}(\omega) = H_m(k, r_m, \alpha_m, \theta_m), \quad (4.10)
\]

\[
\sum_{l=1}^{L} C_m(l, r_m, r_m, \alpha_m, \theta_m + \Delta \theta_m) W_{lk}(\omega) = H_m(k, r_m, \alpha_m, \theta_m), \quad (4.11)
\]

\[
\sum_{l=1}^{L} C_m(l, r_m, r_m + \Delta \alpha_m, \alpha_m, \theta_m) W_{lk}(\omega) = H_m(k, r_m, \alpha_m, \theta_m), \quad (4.12)
\]

Subtracting each of (4.11), (4.12), and (4.13) from (4.10) and dividing by the corresponding displacement results in

\[
\sum_{l=1}^{L} \frac{C_m(l, r_m, r_m, \alpha_m, \theta_m + \Delta \theta_m) - C_m(l, r_m, r_m, \alpha_m, \theta_m)}{\Delta \theta_m} W_{lk}(\omega) = 0, \quad (4.14)
\]

\[
\sum_{l=1}^{L} \frac{C_m(l, r_m, r_m, \alpha_m, \theta_m + \Delta \alpha_m) - C_m(l, r_m, r_m, \alpha_m, \theta_m)}{\Delta \alpha_m} W_{lk}(\omega) = 0, \quad (4.15)
\]

\[
\sum_{l=1}^{L} \frac{C_m(l, r_m, r_m, \alpha_m, \theta_m + \Delta \alpha_m) - C_m(l, r_m, r_m, \alpha_m, \theta_m)}{\Delta \alpha_m} W_{lk}(\omega) = 0, \quad (4.16)
\]

Comparing (4.5), (4.6), and (4.7) with (4.14), (4.15), and (4.16), respectively, shows that the latter three control equations approximate the former three by replacing the spatial derivatives by their corresponding spatial differences. The first-order spatial difference constrained control equation for the whole system becomes

\[
\tilde{C}(\omega, r, \alpha, \theta) W(\omega) = \tilde{H}(\omega, r, \alpha, \theta), \quad (4.17)
\]
where \( C(w, r, \alpha, \theta) \) and \( H(w, r, \alpha, \theta) \) are given by

\[
\begin{align*}
C_{11}(w, r, \alpha, \theta) & \quad \ldots \quad C_{1L}(w, r, \alpha, \theta) \\
C_{1L}(w, r, \alpha, \theta) & \quad \ldots \quad C_{1L}(w, r, \alpha, \theta) \\
\vdots & \quad \ldots \quad \vdots \\
C_{M1}(\omega, \tau, \alpha, \theta) & \quad \ldots \quad C_{M1}(\omega, \tau, \alpha, \theta) \\
C_{M1}(\omega, \tau, \alpha, \theta) & \quad \ldots \quad C_{M1}(\omega, \tau, \alpha, \theta) \\
\vdots & \quad \ldots \quad \vdots \\
H_{11}(w, r, \alpha, \theta) & \quad \ldots \quad H_{1K}(w, r, \alpha, \theta) \\
H_{1K}(w, r, \alpha, \theta) & \quad \ldots \quad H_{1K}(w, r, \alpha, \theta) \\
\vdots & \quad \ldots \quad \vdots \\
H_{MK}(w, \tau, \alpha, \theta) & \quad \ldots \quad H_{MK}(w, \tau, \alpha, \theta) \\
H_{MK}(w, \tau, \alpha, \theta) & \quad \ldots \quad H_{MK}(w, \tau, \alpha, \theta) \\
\vdots & \quad \ldots \quad \vdots \\
H_{MK}(w, \tau, \alpha, \theta) & \quad \ldots \quad H_{MK}(w, \tau, \alpha, \theta) \\
\end{align*}
\]

(4.18)

(4.19)

The matrixes \( \tilde{C} \) and \( \tilde{H} \) are composed of electro-acoustic transfer functions only, no derivative terms appear. This makes the system of control equations in (4.17) similar in nature to (4.3) except for the increase in the number of control microphones. Therefore, the control filters \( \tilde{W}(\omega) \) may be directly implemented using adaptive techniques as discussed in Chapter 3, while more control is achieved on the sound field. For sound reproduction applications, this means placing multiple microphones in the vicinity of each listener's ear.

In the previous discussion, only one extra microphone (displaced in each coordinate direction) has been used to approximate the derivative constraint in that coordinate direction. Better approximation of the first derivatives using 3-point or 5-point formulas [36] may be obtained by using more displaced microphones in each coordinate direction in the vicinity of each control point. More microphones can also be used to approximate higher order derivatives to further widen the zone of equalisation as mentioned in Section 4.2.1. Using two displaced microphones in the \( \tau \) direction at \( (r_m - \Delta r_m, \alpha_m, \theta_m) \) and \( (r_m + \Delta r_m, \alpha_m, \theta_m) \) in addition to the original microphone at \( (r_m, \alpha_m, \theta_m) \), the second-order derivative of \( C_{mf}(\omega, r_m, \alpha_m, \theta_m) \) with respect to \( r_m \) may be approximated by \([36]\)

\[
\frac{\partial^2 C_{mf}(\omega, r_m)}{\partial r_m^2} \approx C_{mf}(\omega, r_m + \Delta r_m) - 2C_{mf}(\omega, r_m) + C_{mf}(\omega, r_m - \Delta r_m) / (\Delta r_m)^2,
\]

(4.20)

where the dependency on \( \alpha_m \) and \( \theta_m \) has been dropped in this equation for clarity. The second-order difference constrained control equation at the \( m \)th control point due to the \( k \)th sound source with respect to \( r_m \) may then be written as

\[
\sum_{k=1}^{L} C_{mf}(\omega, r_m + \Delta r_m) - 2C_{mf}(\omega, r_m) + C_{mf}(\omega, r_m - \Delta r_m) / (\Delta r_m)^2 W_{ik}(\omega) = 0.
\]

(4.21)

Second-order difference constrained control equations with respect to \( \theta^2, \alpha^2, \alpha \theta, \) \( \tau \), and \( \theta \) can be similarly obtained. As for the first-order case, second-order difference control equations are implemented by simply constraining the target transfer functions at all three microphones to be \( H_{mk}(\omega, \tau, \alpha, \theta) \).

Since difference equations are correct direct discrete approximations to derivative equations only when the sampling theorem is honoured, the properties of the equalisation zones obtained using difference constraints depend on the spacing between adjacent microphones. Derivative constraints are equivalent to using an infinite number of control microphones around each of the control points, and smooth interpolation of the sound field at all frequencies is achieved. In the difference constrained case however, a single zone of equalisation is created around each microphone. The size of each zone decreases with increasing frequency similar to the behaviour shown in Fig. 4.1, and smooth interpolation at all frequencies is not guaranteed if the distances between adjacent microphones are large compared to the sound wavelength. At low frequencies, the size of the created zones is large enough to overlap and constitute one larger equalisation zone. As the frequency increases, the size of the individual zones decreases and the single zone starts to split leaving unequalised areas between the adjacent microphones. This behaviour of the zones of equalisation has also been observed in active noise control studies (where they are known as the zones of quiet) [57, 87, 109]. Experimental results in reverberant environments show that the zones of quiet around a microphone within which the error is reduced by 10 dB is a sphere of diameter 0.1 \( \lambda_{\text{min}} \), where \( \lambda_{\text{min}} \) is the wavelength of the highest frequency of interest. This suggests that adjacent microphones must be at least 0.1 \( \lambda_{\text{min}} \) apart for a non-splitting zone of quiet of 10 dB error reduction.

In audio applications, the highest frequency of interest (20 kHz) corresponds to \( \lambda_{\text{min}} \approx 17 \) mm, and a separation of 1.7 mm between adjacent microphones is required for 10 dB equalisation. This reveals severe deficiencies of the method of difference constraints for audio applications:

- The equalisation zones created around the microphones and, therefore, around the listener's ears, are too small at high frequencies. They allow the listener to move only a few millimetres.
- The separation distance between the microphones for a 10 dB equalisation zone at high frequencies is very small, even smaller than the physical diameter of most microphones.
- The 10 dB error reduction is not sufficient for most audio applications. Listening tests [51] show that at least 20 dB error reduction is required for audio applications.
only (opposed to audiovisual) applications. This means that an even smaller separation between the microphones is necessary\(^1\).

* Every extra microphone adds \(L\) electro-acoustic transfer functions to the system. When the control filters are implemented as adaptive filters, this adds \(L\) identification filters and \(L\) convolutions to calculate the filtered input (or error) signals through the newly introduced transfer functions, which increases the system complexity considerably\(^2\).

However, the difference constraint approach represents a powerful means of widening the equalisation zones at low and middle frequencies. This is most effective in applications such as active noise control of low frequency disturbances and speech applications. It can also be used to improve the system robustness at low frequencies while high frequencies are managed by other means. The performance of difference constrained filters is examined using computer simulations in Section 4.2.4.4.

### 4.2.3 Spatial Filters

As mentioned in Section 4.2.2, adding extra microphones in the proximity of the original control positions to apply the difference derivative constraints increases the system complexity considerably. Furthermore, increasing the number of microphones requires increasing the number of reproduction loudspeakers to obtain sufficient error reduction. One approach to contain this rapidly increasing complexity is to reduce the number of control equations. Rather than requiring the sound pressure at each extra control microphone to equal that desired at the adjacent original control point as in (4.17), a linear combination of such conditions may be employed. Such a linear combination reduces to spatially filtering all transfer functions around each control point. This reduces the number of control equations for each original control point to only two. The first equation is the original control equation and the second is a weighted average of all control equations corresponding to the displaced microphones in the vicinity of the original control point.

A linear combination for the conditions that all spatial differences up to the \(J\)th-order to vanish also leads to the above mentioned spatial filter. This may be readily shown by considering the first and second difference control equations in the \(r\) direction given by (4.14) and (4.21), respectively. Multiplying (4.14) by \(a_1\) and (4.21) by \(a_2\) and adding the results gives the weighted average spatial difference constrained control equation at the \(m\)th control point due to the \(k\)th sound source

\[
\sum_{j=1}^{L} \left[ \left( a_1 + a_2 \right) C_m(\omega, r_m + \Delta r_m) - \left( a_1 + 2a_2 \right) C_m(\omega, r_m) + a_2 C_m(\omega, r_m - \Delta r_m) \right] W_{lk}(\omega) = 0, \tag{4.22}
\]

where \(a_1\) and \(a_2\) are weighting coefficients for the first and second derivatives, respectively. Extending the above linear combination to include derivatives with respect to

\[^{1}\text{Increasing the equalisation level by using more loudspeakers is addressed in Section 4.4.}\]
\[^{2}\text{Reducing the complexity of difference constrained systems is addressed in Section 4.2.3 and 4.3.}\]
Chapter 4: Robustness of 3D Sound Systems

4.2 Enlarging the Zones of Equalisation

Figure 4.2: Experimental set-up.

The room simulation program is also used to calculate the desired transfer functions $H_i(\omega)$ and $H_j(\omega)$ from an arbitrarily chosen virtual source position $V_1$ to $R_1$ and $R_2$, respectively. Each transfer function is calculated as a FIR filter of 2048 taps at 44.1 kHz sampling frequency. The impulse responses are also convolved with a measured impulse response of a loudspeaker to account for the responses of the loudspeakers, microphones, DAC, and ADC that appear in real experiments.

Both anechoic and reverberant environments are considered in the experiments. Anechoic environments are simulated by choosing the reflection coefficients $\{\beta_i : i = 1, 2, \ldots, 6\}$ of all room boundaries equal to zero in the room simulation program. Reverberant environments are simulated by choosing these reflection coefficients $0 < |\beta_i| < 1$. In all experiments, the room dimensions in the $x$, $y$, and $z$ directions are 6, 6, and 3 m, respectively. The origin of the coordinate system shown in Fig. 4.2 is taken at $x = 3, y = 3, z = 1$ m. Examples of the calculated transfer functions in an anechoic environment and a moderate reverberant environment are shown in Fig. 4.3 and Fig. 4.4, respectively.

4.2.4.2 Experiment Procedure

In all experiments discussed in this section, the microphones $R_1$ and $R_2$ were positioned on the $x$-axis at $x = -10$ and $x = 10$ cm, respectively. The desired response was chosen corresponding to an azimuth angle $\alpha = 60^\circ$ and radial distance $r = 1.5$ m. The loudspeakers $S_1$ and $S_2$ were positioned at $\alpha = \pm30^\circ$, and $r = 1$ m. This simulates a virtual sound source in the horizontal plane at a wider azimuth angle as mentioned in Section 1.4.

To examine the performance of each of the above mentioned constraint methods under the same conditions, the system of constrained control equations for each method was solved in the frequency domain at each frequency to calculate the control filters $W_1(\omega)$ and $W_2(\omega)$. The control filters were implemented as FIR filters, each of 2048 taps. A third microphone was used to probe the sound field at different points in space. The room simulation program was used to calculate the transfer functions between the two loudspeakers $S_1$ and $S_2$, and the position of the third microphone. For simplicity of representing the results, derivative, difference, and weighted average constraints were considered only in the $x$-direction. Therefore, in the experiments presented below, the third microphone was moved on the $x$-axis from $x = -20$ to $x = 0$ cm in steps of 1 mm. Referring to the transfer functions from $S_1$ and $S_2$ to the third microphone as $C_{31}(\omega)$ and $C_{32}(\omega)$, respectively, the actual system response was then calculated at each position of the probing microphone using the relationship

$$H_{31}(\omega) = \text{diag}\{C_{31}(\omega)\} W_1(\omega) + \text{diag}\{C_{32}(\omega)\} W_2(\omega).$$ (4.25)

This response was then transformed to the time domain (using an inverse FFT algorithm), resulting in $\hat{h}_{31}(n)$. It is worth mentioning at this point that direct solution of the system of control equations in the frequency domain may lead to non-causal filters $W_1(\omega)$ and $W_2(\omega)$. This causes a problem if the impulse response of these filters were to be directly calculated using the inverse Fourier transformation. However, frequency domain multiplication with $C_{31}(\omega)$ and $C_{32}(\omega)$, as in (4.25), introduces enough delay, resulting in a causal response $\hat{h}_{31}(n)$.

To examine the system performance in different frequency bands, the actual time response $\hat{h}_{31}(n)$ was filtered through a one octave constant-Q (percentage bandwidth)
Chapter 4: Robustness of 3D Sound Systems

filterbank composed of eight filters covering the frequency range from 80 Hz to 20480 Hz, resulting in eight time responses \( \{ h_{31}(n) : i = 1, 2, \ldots, 8 \} \). The desired responses \( h_{11}(n) \) and \( h_{21}(n) \) at microphones \( R_1 \) and \( R_2 \) were also filtered through the same filterbank, resulting in \( \{ h_{31}(n) : i = 1, 2, \ldots, 8 \} \) and \( \{ h_{21}(n) : i = 1, 2, \ldots, 8 \} \), respectively. In each frequency band \( i \), the actual and desired responses were compared and an error measure for each desired response was calculated as

\[
E_{ij} = 10 \log_{10} \left( \frac{\sum_n [h_{31}(n) - h_{ji}(n)]^2}{\sum_n h_{ji}^2(n)} \right), \quad i = 1, 2, \ldots, 8, \ j = 1, 2.
\] (4.26)

The errors \( E_{ij} \) in all eight frequency bands were then plotted at each position of the probing microphone along the x-axis. Each experiment was then repeated exactly, except that the control filters were calculated using the system of unconstrained control equations given by (4.3). The errors resulting from the unconstrained filters were plotted on the same graphs for reference.

4.2.4.3 Derivative Approximations

An electro-acoustic transfer function between a receiver at \((x_r, y_r, z_r)\) and a point monopole sound source at \((x_s, y_s, z_s)\) in free field and at a far distance from the source may be approximated by

\[
C_{rs}(\omega, x_r, y_r, z_r) = A_{rs} e^{-j\omega d_{rs}/c},
\] (4.27)

where \( d_{rs} \) is the distance between the source and the receiver, \( c \) is the sound velocity in the medium, \( \omega \) is the angular frequency of the sound and \( A_{rs} \) is a position dependent amplitude factor [10]. In the far field, and at small displacements from the receiver position, the amplitude factor \( A_{rs} \) may be considered distance independent. Using this approximation, the first derivative of \( C_{rs}(\omega, x_r, y_r, z_r) \) with respect to \( x_r \) may be written as

\[
\frac{\partial C_{rs}(\omega, x_r, y_r, z_r)}{\partial x_r} = \frac{j \omega (x_s - x_r)}{c d_{rs}} A_{rs} e^{-j\omega d_{rs}/c}.
\] (4.28)

Using transfer functions given by (4.27) and their first derivatives given by (4.28), it is shown in [10] by means of computer simulations that the first-order derivative constrained control filters designed using (4.8) result in wider overall zones of equalisation. Not shown in [10] are the performance in separate frequency bands and the validity of the approximation for practical electro-acoustical transfer functions.

As mentioned in Section 4.2.1, it becomes more difficult to calculate the analytic expressions for the transfer functions as the reverberation increases. In this section, a slight modification to the first derivative given by (4.28) that allows using derivative constraints in moderate reverberant conditions is proposed. The proposed approximation exploits the fact that a transfer function measured in a reverberant environment has an initial delay equals to that given by the exponential factor in (4.27). Since the derivative in (4.28) is derived using this delay factor only, a sound approximation in moderate reverberant environments is to apply the same delay constraint. Using such a delay constraint implies that the control filters force the direct sound to arrive at the same moment at points in the vicinity of the control microphone while no control effort is made to control the reflections. A control effort on the reflections may be added by deriving the derivative from a transfer function measured in the reverberant environment. Since the derivative in (4.28) is a high-pass filtered version of the original transfer function, the derivative equation for a reverberant transfer function may be approximated by

\[
\frac{\partial C_{rs}(\omega, x_r, y_r, z_r)}{\partial x_r} = G_{HP}(\omega) C_{rs}(\omega, x_r, y_r, z_r),
\] (4.29)

where \( G_{HP}(\omega) \) is a high-pass filter function. In severe reverberant conditions, and at distances further than the reverberation distance (see Section 2.1.63), the power of the diffuse sound field may be larger than that of the direct sound, and the above approximation may fail.

The performance of the derivative constrained filters \( \mathbf{W}_1 \) and \( \mathbf{W}_2 \) shown in Fig. 4.2, when the derivatives are calculated using (4.29) is shown in Fig. 4.5. Figure 4.5 shows the error at the probing microphone with respect to the desired response at \( R_1 \) (located at \( x = -10 \) cm) in each octave band as mentioned in Section 4.2.4.2. This result was obtained using transfer functions calculated in a room with reflection coefficients \( \{ \beta_i = 0.5 : i = 1, 2, \ldots, 6 \} \). For 10 dB equalisation level, the constrained filters achieve wider equalisation zones than the unconstrained filters in all but the highest two octaves. The 20 dB equalisation zone is approximately 20 cm in diameter.
in the lowest octave, while the level and the size of the zones decrease as the frequency increases.

The performance of the derivative constrained filters can be understood from the physical nature of the constraint. This is best explained using a single channel system. In this case, the system of derivative constrained control equations is given by

\[ C_{11}(ω)W_{11}(ω) = H_{11}(ω), \]
\[ G_{HP}(ω)C_{11}(ω)W_{11}(ω) = 0. \]  \hspace{1cm} (4.30)

Since the system of equations is overdetermined, there exists no exact solution. A general LMS solution of such an overdetermined system is \( \mathbf{W} = \mathbf{C}^H(\mathbf{C}^H\mathbf{C})^{-1}\mathbf{H} \), where \( \mathbf{C}^H \) denotes the complex conjugate transpose (Hermitian). The LMS solution for (4.30) is, therefore,

\[ W_{11}(ω) = \frac{1}{1 + |G_{HP}(ω)|^2} C_{11}^{-1}(ω) H_{11}(ω). \]  \hspace{1cm} (4.31)

At very low frequencies, \( |G_{HP}(ω)|^2 \) is very small, and the solution to the filter is

\[ W_{11}(ω) \approx C_{11}^{-1}(ω) H_{11}(ω), \]  \hspace{1cm} which is very close to the exact solution. As the frequency increases, the term \( |G_{HP}(ω)|^2 \) increases and the filter drifts from the exact solution. The filter is, therefore, designed to increasingly ignore more details in the acoustic transfer function \( C_{11}(ω) \) as the frequency increases. This makes the filter a good approximation to the solution required at points in the vicinity of the original control point. The area in which the filter is considered a valid solution decreases with increasing frequency. In fact, (4.31) suggests that the general solution for the derivative constrained filter is a low-pass filtered version of the exact solution. The idea of designing the control filters to increasingly ignore more details as the frequency increases is further pursued in Section 4.3.

4.2.4.4 Microphone Clusters

Figure 4.6 shows the error criteria for the first microphone \( \{E_{i1} : i = 1, 2, \cdots, 8\} \) when difference constraints are used to calculate the control filters. This result was obtained by using two extra microphones positioned at \( x = -15 \) cm and \( x = -5 \) cm in addition to the original microphone \( R_1 \) at \( x = 10 \) cm. In the room, the three microphones were positioned at \( (x = 2.85m, y = 3m, z = 1m), (x = 2.95m, y = 3m, z = 1m), \) and \( (x = 2.90m, y = 3m, z = 1m), \) respectively. Transfer functions calculated for a room with reflection coefficients \( \{β_i = 0.5 : i = 1, 2, \cdots, 6\} \) have been used in this simulation. It can be seen from this result that microphone clusters are very effective at low frequencies. The 30 dB zone of equalisation is wider than 20 cm in the lowest octave. The 20 dB zone of equalisation of 20 cm diameter is extended to cover the frequency range up to 640 Hz, while the same zone was limited to 160 Hz for derivative constrained filters. The single zone of equalisation becomes narrower and shallower as the frequency increases. Above a certain frequency (which depends on the separation between the microphones), the single large zone of equalisation splits into separate zones around the individual microphones. This effect is very clear in the highest octave band where very narrow and shallow zones of equalisation are created around the three microphones. Decreasing the distance between the microphones extends the zones to higher frequencies while reducing the diameter of the single large zone. Using more microphones (say at \( x = -15, x = -12.5, x = -7.5, \) and \( x = 5 \) cm) extends the frequency range while keeping the width of the single zone unchanged. However, the equalisation level may decrease as the sound field is controlled at more and more microphones using only two loudspeakers.

4.2.4.5 Spatial High-Pass Filters

As mentioned in Section 4.2.3, the complexity of the system of difference constrained control equations given by (4.17) may be significantly reduced by employing a single constraint derived by spatially filtering all transfer functions measured at each microphone cluster. Extensive simulations show that only spatial high-pass filters are capable of widening the zones of equalisation. This is also suggested by (4.22), since the sign of the centre coefficient is opposite to the other coefficients. Such a spatial high-pass filter rejects common features among the transfer functions in the cluster while retains fast changing features. Since fast changing features increasingly appear as the frequency increases, a spatial high-pass filter may be seen to reject the low frequencies and retain the high frequencies of the transfer function measured at the control point in the centre of the cluster. This is similar to the action of the high-pass filter \( G_{HP}(ω) \) in (4.29), which explains the reason why the spatial filter must be of a high-pass nature. The weighted average constrained control filters, therefore, approximate the derivative constrained filters only if the weighted average is of the high-pass type. The more transfer functions are used to calculate the weighted average, the
The factor $C^{-1}(\omega)$ exactly inverts the electro-acoustic channels between the reproduction loudspeakers and the listeners' ears, and the factor $H(\omega)$ adds the desired response as mentioned before. The main problem with this exact solution is that it is valid only at the listeners' ears, the points at which the transfer functions $C(\omega)$ and $H(\omega)$ have been previously measured. When a listener moves slightly, both $C(\omega)$ and $H(\omega)$ change, and the control filters $W(\omega)$ become in error, which is a problem that is inherent to the exact solution. Due to the short wavelengths at high frequency, two impulse responses measured at adjacent points in a room increasingly differ from each other as the frequency increases. The same short wavelength also makes head-related transfer functions increasingly differ from those measured at adjacent azimuth or elevation angles as the frequency increases. These physical properties cause the errors introduced in $W(\omega)$ due to a listener movement to increase with increasing frequency.

One approach to improving the robustness of $W(\omega)$ against small listener movements is to design the filters to be an average solution that is valid in an area in space around the listeners' ears. This is in contrast to filters that implement the exact solution, which is valid only at the listeners' eardrums. Based on the above mentioned physical observations, the deviation from the exact solution should also increase with increasing frequency to cope with the error increase. This is also in full agreement with the physical interpretation of the constrained filters discussed in Section 4.2.

Many authors have reported obtaining robust systems using control filters with average solution at high frequencies. It is suggested in [46] that using transfer functions that are free from details of specific heads to design the control filters results in filters that are tolerant to variations in head shape and position. It is also suggested in [46] that if robust filters are required, only the envelope at high frequencies and not the detailed solution should be employed. In other work [92], one-third octave constant-Q smoothing filters were used to smooth the HRTF data prior to using them in calculating the cross-talk cancellation and the binaural synthesis filters. The same constant-Q smoothing was also used in [72] for implementing cross-talk cancellers. Furthermore, the cross-talk cancellation function employed in [72] was band-limited to 6 kHz, while only the correct sound power is delivered to the ears at higher frequencies. In yet another application [76, 77], segmented adaptive control filters were used, where the control filters were divided into segments with lower frequency resolution as the frequency increases.

In all the above mentioned examples, the main goal was to average the sharp features that occur at high frequencies from the HRTF data to avoid a solution that deals specifically with those features. Since those sharp and fast changing features increase with increasing frequency, it is essential that the width of the averaging window also increases with frequency. This has already been applied in the constant-Q averaging [92] and the segmented filters [76, 77] mentioned above. The resulting control filters are of non-uniform frequency resolution (multiresolution filters) with decreasing frequency resolution with increasing frequency, therefore, averaging wider frequency bands at shorter wavelengths. This approach, in addition to leading to an average solution at high frequencies, relaxes the problem of inverting deep notches in the cross-talk canceller.

---

**Figure 4.7:** The same as Fig. 4.5 but with the solid curve obtained for weighted average constrained filters. Two extra transfer functions from each loudspeaker to $x = -12$ and $x = -8$ cm were used to calculate the weighted average.

better the approximation becomes.

An example of the performance of weighted average constrained filters is shown in Fig. 4.7. Transfer functions calculated for a room with reflection coefficients $\{\beta_i = 0.5 : i = 1, 2, \ldots, 6\}$ have been used in this simulation. The weighted average transfer functions $C_{1f}$ in (4.24) were obtained by using the following high-pass spatial filter

$$C_{1f} = \frac{2}{6} C_{1f} - \frac{4}{6} C_{2f} + \frac{1}{6} C_{3f}, \quad l = 1, 2,$$

(4.32)

where $C_{1f}$, $C_{2f}$ and $C_{3f}$ are the transfer functions between microphones positioned at $x = -12$, $x = -10$, and $x = -8$ cm and the $l^{th}$ loudspeaker, respectively. This result shows that simple spatial high-pass filters are capable of approximating the derivative constraint and, therefore, result in wider equalisation zones. Similar properties of the zones of equalisation in the different frequency bands may also be observed from this result as in the cases of derivative and difference constraints.

### 4.3 Multiresolution Control Filters

Assuming that there exists an exact solution for the matrix of control filters $W(\omega)$ in the generalised multichannel audio reproduction system shown in Fig. 3.1, this solution is given by

$$W(\omega) = C^{-1}(\omega)H(\omega).$$

(4.33)

80
On the other hand, the multiresolution filters may sound objectionable in some applications such as virtual source synthesis, since sound localisation cues for elevation and cone of confusion resolution are high frequency features that are to be averaged. The effects of this averaging on the perceived source location can only be studied using psychoacoustical experiments. However, the use of multiresolution filters is justified by the fact that the human auditory system performs a similar non-uniform spectral analysis and, therefore, has similar inaccuracies as mentioned in Section 2.1.8. Furthermore, human localisation resolution that depends on high frequency cues (elevation resolution for instance) are rather poor compared to those based on binaural cues [30]. Since a multiresolution filter is still delivering the correct interaural time delay envelope at high frequencies and correct interaural intensity difference, the effects on localisation accuracy may be reduced. This, however, must be confirmed by psychoacoustical experiments.

In traditional 3D sound systems, the control filters are usually designed off-line, and the multiresolution control filters may be calculated by processing the HRTF data using constant-Q smoothing filters. In adaptive 3D sound systems, this frequency-dependent averaging must be performed in real-time. When the adaptive filters are implemented in the frequency domain, this reduces to forcing the adaptive filter to the required resolution by employing a non-uniform, constant-Q-like transformation in place of the constant bandwidth FFT, as shown in Fig. 4.8. The problem of designing the multiresolution filters then reduces to designing the multiresolution transformation \( T \) that replaces the FFT in conventional frequency domain adaptive filters. Since computational saving is one of the main reasons for implementing the filters in the frequency domain as mentioned in Section 3.4.2, the transformation \( T \) must be calculated as fast as the usual FFT to maintain this computational advantage. Chapter 5 is devoted to developing such fast multiresolution transformations, where it is shown that a fast multiresolution transformation may be obtained by calculating the FFT of non-uniformly sampled signals [66, 70]. Improving the robustness of 3D sound systems by employing multiresolution control filters is shown by computer simulations in [68], and repeated in Section 4.3.1 for reference. Most audio signal processing applications may also benefit from this non-uniform spectral resolution that resembles the processing of the human ears. An example of improving the performance of phase vocoder systems has also been presented in [69].

### 4.3.1 Performance of Multiresolution Filters

In this section, it is shown, using a computer experiment, that the robustness of 3D sound systems can be improved by using control filters having non-uniform frequency resolution as mentioned in Section 4.3. The experiment shown in Fig. 4.9 implements a one-point noise canceller, which may also be seen as a simplified version of a virtual sound source generator at one ear only. A white noise signal \( x \) is filtered through an acoustic impulse response \( h(n) \) simulating the sound pressure due to a physical loudspeaker \( P \) at one eardrum of the listener. The filter \( w \) is calculated to minimise the sound pressure at MIC through loudspeaker \( S \) and the acoustic transfer function \( g(n) \). The MIC output is then fed to a cochlea model (gammatone filterbank [117]) consisting of 64 channels and covering the frequency range from 20 Hz to 20 kHz. The output of each fourth channel of these 64 channels are half-wave rectified, low-pass filtered and plotted against time for inspection. These plots represent the probability of firing along the auditory nerve due to the sound signal received at the listener’s eardrum represented here by the microphone MIC.

The optimum solution for the filter \( w \) in this case is given in the frequency domain by

\[
W(\omega) = \text{diag}\{C^{-1}(\omega)\} H(\omega),
\]

where \( C(\omega) \) and \( H(\omega) \) represent the spectra of the corresponding acoustic impulse responses at \( \omega \). From this equation, it is clear that the filter length must be large enough to accommodate this solution. The degree of attenuation in sound pressure at MIC depends on the length of the filter \( w \). A longer filter achieves higher attenuation and, therefore, better emulation of the virtual source.

Using the FFT algorithm to calculate \( C(\omega) \) and \( H(\omega) \) from \( g(n) \) and \( h(n) \) produces a constant resolution filter \( \mathbf{W}(\omega) \). On the other hand, using a multiresolution transformation to obtain \( C \) and \( H \) produces a multiresolution solution for \( W \). To show
Chapter 4: Robustness of 3D Sound Systems

the performance improvement achieved by using multiresolution filters, the following experiment was performed:

Four acoustic impulse responses, each of 512 samples, were calculated using the room simulation program Room Impulse Response 2.1 [71]. A room having a reverberation time of 0.3235 s, and dimensions of 5 m, 4 m, and 3 m in the x, y, and z directions, respectively, has been used. The four transfer functions are as follows:

- \( h_0 \): from loudspeaker \( P \) at \((2, 3, 1.5 \text{ m})\) to MIC position at \((4, 3, 1.5 \text{ m})\)
- \( e_0 \): from loudspeaker \( S \) at \((4, 2, 1.5 \text{ m})\) to MIC position at \((4, 3, 1.5 \text{ m})\)
- \( h_1 \): from loudspeaker \( P \) at \((2, 3, 1.5 \text{ m})\) to MIC position at \((4, 3, 1.52 \text{ m})\)
- \( e_1 \): from loudspeaker \( S \) at \((4, 2, 1.5 \text{ m})\) to MIC position at \((4, 3, 1.52 \text{ m})\)

The filter \( W \) of 1024 coefficients was calculated in two different ways: using uniform and non-uniform spectral resolutions. Only the initial MIC position has been used in calculating \( W \) such that \( W = \text{diag}(\{c_0^{-1}\} H_0) \). The simulation shown in Fig. 4.9 was repeated twice: once using the constant resolution filter and the second time using the multiresolution filter. At the middle of the simulation, MIC is moved 2 cm from its initial position (the simulation switched to \( h_1 \) and \( e_1 \) in place of \( h_0 \) and \( e_0 \)). The output of this experiment is shown in Fig. 4.10, where the horizontal axis represents time in samples and the vertical axis is the sound pressure level (SPL) in dB. The simulation represents moving the microphone 2 cm vertically after 500 samples. The reduction achieved by a robust filter should not be affected by this sudden movement. This may occur if the spectrum of \( e \) contains a deep notch at these frequencies, causing inversion problems. Since coarser resolution is equivalent to spectral smoothing, this produces shallower notches and a better solution for \( W \) may be obtained.

4.4 Underdetermined Systems and Exact Solutions

The other important property of the zones of equalisation, besides their width, is the equalisation level. One expects that using more degrees of freedom in the generalised multichannel audio reproduction system shown in Fig. 3.1 results in a higher equalisation level due to improved controllability. More degrees of freedom may be employed by increasing the number of taps of the control filters. Better results may
be obtained by dividing the number of taps into more filters driving more loudspeakers. As mentioned in Section 3.2, when the number of reproduction loudspeakers $L$ is greater than the number of control microphones $M$, the system of control equations is underdetermined. There are less equations than unknowns, and there exists an infinite number of solutions. However, it is shown in [102, 103] that when $L = M + 1$, there exists a unique and exact (opposed to LMS) solution for the matrix of control filters $W(w)$, which reduces the error exactly to zero provided that

- All impulse responses $\{z_{ml} : l = 1, 2, \ldots, L\}$ from all the loudspeakers to the $m^{th}$ microphone do not contain any common zeros.
- The number of FIR coefficients of the control filters $N_w$ is constrained to be less than the number of FIR coefficients $N_c$ representing $z_{ml}$.

This approach, referred to as the Multiple-input\output INverse Theorem (MINT), is further extended in [107] for $L > M$ not constrained to $L = M + 1$. Furthermore, it is shown in [107] that with the correct choice of both modelling delay and the number of coefficients of the control filters, the design of the filters using MINT or MELMS (see Chapter 3) should in principle yield identical results. Therefore, it is possible to implement the system using adaptive filters, the optimum solutions of which are the exact solutions that reduce the error exactly to zero by simply increasing the number of reproduction loudspeakers and fulfilling the delay and number of coefficients constraints. Provided that the loudspeakers are well-positioned in the listening space (see Section 4.5), the underdetermined system is expected to provide better equalisation level than the fullydetermined system.

The correct modelling delay and the number of coefficients for the adaptive filters to exactly model the solution have also been evaluated in [107]. As mentioned in Section 3.3.4, the target transfer functions $H_{mk}(\omega)$ must contain sufficient delay for a causal and stable solution to exist, which is the same constraint on the modelling delay. Assume that the shortest pure delay in the impulse responses $\{z_{ml} : l = 1, 2, \ldots, L\}$ between any of the reproduction loudspeakers and the $m^{th}$ microphone is $\Delta_{m}^{\text{min}}$ samples, and all $\{z_{ml} : l = 1, 2, \ldots, L, m = 1, 2, \ldots, M\}$ can be accurately modelled by FIR filters, the longest of which contains no non-zero coefficients beyond $\Delta_{m}^{\text{max}} - 1$ samples. If the adaptive filters are FIR filters of equal length $N_w$, then the delay $\Delta_{m}$ in any of the target transfer functions $\{H_{mk} : k = 1, 2, \ldots, K\}$ must be in the range [107]

$$\Delta_{m}^{\text{min}} \leq \Delta_{m} < \Delta_{m}^{\text{max}} + N_w - 1. \quad (4.35)$$

The number of each of the control filters for the exact solution to exist must be

$$N_w = \left[ \frac{\sum_{m=1}^{M} (\Delta_{m}^{\text{max}} - \Delta_{m}^{\text{min}})}{L - M} \right]. \quad (4.36)$$

To examine the improvement brought about by using more reproduction loudspeakers, the experiment described in Section 4.2.4.2 has been repeated twice. Once with two loudspeakers positioned at $\alpha = \pm 45^\circ$ and a radial distance $r = 100$ cm from the Centre of the listener's head and once with three loudspeakers positioned at $\alpha_1 = -45^\circ$, $\alpha_2 = +45^\circ$, and $\alpha_3 = 180^\circ$, and $r = 100$ cm. The error criteria (4.26) were calculated at discrete points in the range from $x = y = -20$ cm to $x = y = 20$ cm with 2 mm resolution in the horizontal plane $y = 100$ cm containing the microphones and the loudspeakers. Plots of the equal contours of $\{E_{il} : i = 1, 2, \ldots, 8\}$ at equalisation levels of 0, -10, and -20 dB are shown in Fig. 4.11 and Fig. 4.12 for the two and
three loudspeakers, respectively. Similar computer simulations in [107] also show that choosing \( L > M \), with the correct delay and number of coefficients, not only decreases the error at the control points considerably, but also results in slightly wider zones of equalisation at mid-band frequencies. Cross-talk cancellation experiments at two microphones conducted in [139] also show higher equalisation levels for three reproduction loudspeakers.

A more important effect of the underdetermined system is the shape of the zones of equalisation. The three-loudspeaker system results in zones that are almost round, therefore, allowing the listener to move his head equally in all directions. This is in contrast to the two-loudspeaker system, which allows the listener to move his head to the front and back more than to the right and left as shown in Fig. 4.11 and Fig. 4.12. This, however, is not the case for arbitrary positions of the loudspeakers. The positions of the loudspeakers in the three loudspeakers experiment has been optimally chosen (see Section 4.5).

Underdetermined systems also allow dividing the control effort on more filters and more loudspeakers. It was noticed from the above mentioned simulations that the amplitude of the filters in the three-loudspeaker system is much lower than that of the two-loudspeaker system at all frequencies. Therefore, it is less likely to overload the loudspeakers and encounter nonlinearities when underdetermined systems are used.

### 4.5 Reproduction Loudspeakers Positioning

It is shown in Section 3.3.3 that the positions of the reproduction loudspeakers relative to the control microphones have considerable effect on the convergence speed of the adaptive filters. This same relative positions also have strong influence on the sensitivity of the filters to listeners' movements. This can be explained by examining the control equations for the \( k \text{th} \) input signal

\[
C(w)W_k(w) = H_k(w),
\]

(4.37)

where \( W_k(w) \) and \( H_k(w) \) are the columns in \( W(w) \) and \( H(w) \) given in (4.3), corresponding to the \( k \text{th} \) input signal. From error estimates in numerical analysis, it is known that in a system of linear algebraic equations such as (4.37), a small change in \( C(w) \) causes a large change in the solution \( W_k(w) \) when \( C(w) \) is ill-conditioned [5, 10, 36]. Considering small listener movements as small changes in \( C(w) \), the difference between the required solution for \( W_k(w) \) before and after a move is large if \( C(w) \) is ill-conditioned. Small movement lead to large errors in the filters, therefore, a system with an ill-conditioned matrix \( C(w) \) is more sensitive to listeners' movements.

Since the condition of the matrix \( C(w) \) is dependent on the positions of the reproduction loudspeakers relative to the microphones, it is important to choose those positions so that the matrix \( C(w) \) is well-conditioned at all frequencies. This may be done by examining the condition number of \( C(w) \) at each frequency in the frequency range of interest. The 2-norm condition number of a matrix \( C(w) \) is defined as the ratio between its largest and smallest singular values \(^4\) [10]

\[
\kappa_2(C(w)) = \frac{\sigma_{\text{max}}(C(w))}{\sigma_{\text{min}}(C(w))}. \quad (4.38)
\]

A well-conditioned matrix has a condition number close to one, while an ill-conditioned matrix has a large condition number. Ill-conditioned situations, such as those described in Section 3.3.3, reduce the rank of the matrix \( C(w) \) and consequently increase its condition number making it more sensitive to listener's movements.

#### 4.5.1 Optimum Stereo Set-Up

A case of special interest in audio reproduction is the stereo system, where two reproduction loudspeakers are used. Optimum positioning of those two loudspeakers may be investigated by examining the condition number (4.38) for different positions of the loudspeakers at each frequency \( \omega \). The experimental set-up shown in Fig. 4.2 has been used to perform this investigation. In the following experiments, the azimuth angle \( \alpha \) was changed from \( 1^\circ \) to \( 90^\circ \) in one degree steps, while the radial distance \( r \) is kept constant. In other words, the two loudspeakers were moved on a circle of radius one meter, the centre of which is the centre of the listener's head. In each position, the room simulation program Room Impulse Response 2.1 [71] was used to calculate the four electro-acoustic transfer functions between the loudspeakers and the microphones. Each transfer function was calculated as an FIR filter of 2048 taps at 44.1 kHz sampling frequency. The FFT algorithm was used to calculate the discrete frequency response of the four transfer functions at each position and (4.38) was used to calculate the condition number of \( C(w) \) at each discrete frequency \( \omega \). The above procedure was repeated for anechoic and reverberant conditions of different reverberation characteristics.

The condition number of \( C(w) \) in an anechoic chamber is shown in a grey-scale graph against the frequency in Fig. 4.13. In this experiment, the two microphones were positioned at \( x = -9cm \) and \( x = 9cm \). The radial distance between the loudspeakers and the centre of the listener's head was \( r = 100 \) cm. Figure 4.14 shows the condition number at selected values of \( \alpha \) from those shown in Fig. 4.13 in two-dimensional graphs. The horizontal axis in Fig. 4.14 represents the frequency and the vertical axis represents the condition number. Since the condition number of \( C(w) \) in an anechoic environment is dependent only on the relative positions of the loudspeakers and the microphones, different results are expected for different \( r \). Figures 4.15 and 4.16 show the condition number of \( C(w) \) for the cases \( r = 50 \) cm, and \( r = 150 \) cm, respectively. The result shown in Fig. 4.16 was obtained by moving the two microphones from their symmetrical positions to \( x = -7cm \) and \( x = 11cm \).
Figure 4.13: The condition number of \( C(\omega) \) in an anechoic room. \( d_m = 18 \) cm, and \( r = 100 \) cm.

Figure 4.14: The condition number of \( C(\omega) \) at discrete angles from those in Fig. 4.13

Figure 4.15: The condition number of \( C(\omega) \) in an anechoic room. \( d_m = 18 \) cm, and \( r = 50 \) cm.

Figure 4.16: The condition number of \( C(\omega) \) in an anechoic room. \( d_m = 18 \) cm, and \( r = 150 \) cm.

For closely spaced loudspeakers, and at low frequencies, the matrix \( C(\omega) \) is almost singular. From (4.37), the amplitude responses of the filters \( W_k(\omega) = C^{-1}(\omega) H_k(\omega) \) are, therefore, very large at low frequencies. The filters tend to amplify the low frequencies, which leads to saturating the audio amplifiers driving the loudspeakers or damaging the loudspeakers. A possible solution for this problem is to divide the audible frequency range into two or more bands. The low frequency band is reproduced through widely spaced loudspeakers while the high frequency band through closely spaced ones. The two closely spaced loudspeakers may be small tweeters since they reproduce high frequencies only while the widely spaced ones must be large loudspeakers (subwoofers) for low frequency reproduction. Although this arrangement uses four loudspeakers, it may still be considered as a two-loudspeaker system, since each tweeter and subwoofer may be mounted on one enclosure, as is usually the case in a closed or vented box loudspeaker system. The two enclosures are then placed with the tweeters close to each other while the subwoofers are widely spaced. For most loudspeakers, this will mean placing the boxes horizontally in front of the listener instead of the conventional vertical position.

4.5.2 Three-Loudspeaker Systems

Another approach to improve the system robustness is to increase the number of reproduction loudspeakers, where each loudspeaker covers the whole audible frequency range (see also Section 4.4). Figures 4.17 and 4.18 show the condition number for a three-loudspeaker system. The result in Fig. 4.17 was obtained by placing two loudspeakers at angles \( \alpha = \pm 5^\circ \) and changing the azimuth angle of the third loudspeaker from 0° to 180° in 2° steps on a circle of radius 100 cm. Fig. 4.18 was obtained using the same procedure while the two fixed loudspeakers were positioned at angles \( \alpha = \pm 45^\circ \).

When the two fixed loudspeakers are positioned very close to each other, the matrix

It may be concluded from the above observations that a small \( \alpha \) creates a wide frequency band at mid-band frequencies where the condition number is very low, while a large \( \alpha \) does better at low frequencies. Although several authors [23, 91, 111] have suggested using closely spaced loudspeakers in stereo reproduction systems to widen the equalisation zones in the middle-band frequencies, this compromises the performance at low frequencies considerably. This problem has also been recognised in [91].
4.5.3 Performance in Reverberant Environments

In most studies of sound reproduction systems, only simplified free field acoustic transfer functions are used. Although this leads to better understanding of the underlying physical principles, sound reproduction systems are rarely designed to operate in free field environments. Therefore, examination of the effects of reverberation on results obtained under free field conditions is necessary. This is readily performed in the experimental set-up used in the present work by changing the reflection coefficients of the room boundaries in the room simulation program. The results of these changes are shown in Fig. 4.19 and Fig. 4.20 after setting all $\beta_i$ to 0.5 in the case of three-loudspeaker systems. This can be seen from Fig. 4.21 and Fig. 4.22 which are obtained by repeating the experiments shown in Fig. 4.17 and Fig. 4.18 after setting all $\beta_i$ to 0.5. In the case of two closely separated loudspeakers, little change is noticed by increasing reverberation. When the two fixed loudspeakers are widely separated, the condition number increases while the third loudspeaker has limited the smearing of traces over frequency. In general, a three-loudspeaker system is more robust in reverberant environments than a two-loudspeaker system.

### 4.6 Large Movements and Head Tracking

Up to this point, only small listener's movements have been considered. For small movements, it is sufficient to design the control filters to have average solutions that are valid within small areas around the listener's ears. When the listener moves beyond this area, the control filters become in great error, and new filters that are derived from the new listener's position must be used. Traditional binaural synthesis systems solve the problem of a moving listener (or a moving virtual source) by switching between pre-measured HRTF pairs [25]. To simulate a moving source, the system switches between stored HRTF pairs that are previously measured at discrete positions on the required source contour in the correct sequence. For a moving listener, a sensor is employed to continuously monitor the listener's position and send the position information to a computer. At every time moment, the position information is used to address the HRTF pair nearest to the current listener's position.
In an adaptive audio reproduction system, the control filters are designed in real-time, and the above mentioned switching technique can be used. Provided that the adaptive filters can be updated faster than the listener’s movements, it is sufficient to re-adapt the control filters to their new optimum solutions that correspond to the current listener’s position. This is complicated by the following factors,

- The filtered-\(x\) LMS and the adjoint LMS algorithms used to update the control filters (see Chapter 3) require estimating all the electro-acoustic transfer functions \(C_{ml}(\omega)\).
- Adaptive 3D sound systems introduced in Chapter 3 are based on cancelling the sound emitted from a physical loudspeaker at the listener’s ears. Using the same technique with the audio signals results in cancelling the sound meant for the listener to hear.
- The filtered-\(x\) and adjoint LMS algorithms are known for their slow conversion speed due to the electro-acoustic transfer functions in front of the filters.

These issues are the subjects of the next few sections. On-line estimation of electro-acoustic transfer functions is addressed in Section 4.7. Methods for updating the control filters while leaving the audio signal for the listener to hear are discussed in Section 4.7.3. Finally, improving the convergence speed of the filtered-\(x\) and the adjoint LMS algorithms is treated in Section 4.8.

### 4.7 On-Line Identification

Both the MEFX and the ALMS algorithms discussed in Chapter 3 require measuring the acoustic transfer functions between the reproduction loudspeakers and the microphones. In a static situation, measuring these transfer functions off-line in an initialisation stage is sufficient. In a dynamic situation when one (or more) of these transfer functions changes, new measurements must be performed and used to recalculate the matrix of control filters \(W(\omega)\). When such changes occur frequently, it is necessary to continuously measure the electro-acoustic transfer functions and use the last measurement to update the filters. Methods to measure the transfer functions while updating the control filters are discussed in this section. Two main methods are known in active noise control literature for on-line identification. The first uses an extra training signal to update an adaptive filter connected in parallel with the acoustic transfer function to be measured. This method is discussed in Section 4.7.1. The second uses no extra training signal and is discussed in Section 4.7.2. The application of these methods to multichannel audio reproduction systems is investigated in Section 4.7.3.

#### 4.7.1 Using Extra Training Signals

In both off-line and on-line identification of acoustic transfer functions, an adaptive identification process is usually employed. Although adaptive filters connected in series with the electro-acoustic transfer function to be measured may be used [61], only FIR adaptive filters connected in parallel are considered here. The basic principle of on-line identification using extra training signals is introduced in Section 4.7.1.1 for a single channel system. Improving the convergence speed and extending the basic principle to the multichannel case are discussed in Section 4.7.1.2 and Section 4.7.1.3, respectively.
4.7.1.1 The Basic Principle

The basic principle of on-line adaptive identification using an extra white noise training signal \[60, 61\] is shown in Fig. 4.23 for a single channel active noise control system. The filtered-x (or alternatively the adjoint) LMS algorithm is used to update the control filter \( W \) to cancel the sound emitted by loudspeaker \( P \) at the microphone position. At the same time, another adaptive filter \( C \) is used to estimate the acoustic transfer function \( C \). An extra white noise signal \( s(n) \) (or any spectrally rich alternative such as a chirp signal) that is uncorrelated with \( x(n) \) is added to the control signal \( y(n) \) and the sum is used to drive loudspeaker \( S \). The superposition of the responses of loudspeakers \( P \) and \( S \) is observed by the microphone Mic. The microphone signal \( e(n) \) can be expressed as

\[
e(n) = s(n) * c + x(n) * [h + w * c],
\]

which shows that \( e(n) \) does not contain the desired component \( s(n) * c \) only but is also contaminated with a disturbance term which is dependent on \( x(n) \). The output of the filter \( C \) is subtracted from the microphone output and the result \( r(n) = e(n) - z(n) \) is used as the error, the square of which is to be minimised by the LMS algorithm. The input to the adaptive filter \( C \) and its LMS update algorithm is the extra white noise training signal \( s(n) \) only. Since \( s(n) \) is chosen to be uncorrelated with \( x(n) \), the adaptive filter \( C \) converges on average to the desired response \( C \), regardless of the above mentioned disturbance \[144\]. However, the disturbance degrades the convergence speed of the filter, a matter that will be shortly discussed in Section 4.7.1.2.

The main disadvantage of the above mentioned method is that the extra training signal remains to be heard at the microphone. In active noise cancellation applications, this represents an additional residual noise, which may be tolerated if the training signal \( s(n) \) is about 25 dB lower in level than \( x(n) \).

4.7.1.2 Improving the Convergence Speed

The training signal \( s(n) \) in Fig. 4.23 must be kept at a low level for adequate system performance. The identification error \( r(n) \) is, therefore, contaminated with a large disturbance due to the components \( d(n) = x(n) * h \) and \( d(n) = y(n) * C \). For a stable identification process, the step size \( \mu \) of the LMS update of \( C \) must be kept small \[93, 144\], which results in a slow convergence speed. The convergence speed can be improved by estimating the undesired component \( d(n) \) and subtracting it from \( e(n) \). This is achieved using a third adaptive filter \( Q \) with input \( x(n) \) to estimate the transfer function \( h + w * c \) as shown in Fig. 4.24. In this case, \( s(n) * c \) represents a disturbance for the adaptive filter \( Q \). Since \( s(n) \) and \( x(n) \) are uncorrelated, and the variance of \( s(n) \) is much smaller than that of \( x(n) \), \( Q \) converges fast to its average final solution. This in turn cancels the disturbance component in \( r(n) \), allowing the use of a larger step size to update \( C \).
4.7.1.3 Multichannel Identification

In multichannel systems, all transfer functions \( \{C_{ml} : m = 1, 2, \ldots, M, l = 1, 2, \ldots, L \} \) between the reproduction loudspeakers and the microphones must be estimated. However, it is not possible to estimate all \( M \times L \) channels simultaneously using one extra training signal \( s(n) \) due to the interchannel coupling effects [99]. Due to this coupling, the estimate of \( C_{ml} \) is biased by the estimation error of other transfer functions from all loudspeakers to the same microphone \( m \). For this reason, only transfer functions from one of the loudspeakers to all microphones \( \{C_{ml} : m = 1, 2, \ldots, M \} \) may be identified simultaneously using a single extra training signal. This identification process is then repeated for the other loudspeakers, which increases the identification time. One approach to speed up the identification of multichannel systems is to employ \( L \) independent training signals, one for each reproduction loudspeaker. Alternatively, a single training signal may be used with inter-channel delay to decorrelate the excitation signals to the loudspeakers [94]. The latter method is more efficient, since only one internal training signal must be generated compared to \( L \) signals in the former.

4.7.2 Using the Control Signals

Using an extra training signal for on-line identification as described in Section 4.7.1 may not be desired in some applications. In such cases, the control signal \( y(n) \) may be used as a training signal. In general, the training signal \( y(n) \) is not persistent and its spectrum is not flat. The spectral contents of such a training signal depends on the spectral contents of the input signal \( x(n) \) and the frequency response of the control filter \( W \). This leads to a slow convergence speed, and the estimated filter \( \hat{C} \) depends on the spectrum of \( y(n) \). However, this approach may be attractive in many situations, especially when the adaptation is performed in the frequency domain, allowing power normalisation in each frequency bin as mentioned in Section 3.4.2.

Another problem that arises when using \( y(n) \) as a training signal is that \( y(n) \) is always well correlated with \( x(n) \). The microphone signal in this case is given by

\[
e(n) = \mathbf{x}^T(n) \mathbf{h} + \mathbf{y}^T(n) \mathbf{g}, \tag{4.40}
\]

which shows that the microphone signal is contaminated with a component that is dependent on \( x(n) \). For proper convergence, the disturbance \( \mathbf{x}^T(n) \mathbf{h} \) in the microphone signal must be removed. This may be achieved by using an additional adaptive filter similar to the approach followed to improve the convergence speed in 4.7.1.2.

The block diagram of a single channel active noise control system employing the control signal as a training signal is shown in Fig. 4.25. The acoustic transfer function \( C \) is estimated using the adaptive filter \( \hat{C} \). Another adaptive filter \( \mathbf{H} \) is used to estimate the transfer function between loudspeaker \( P \) and the microphone. The outputs of the two estimation filters are added to obtain an estimate \( \hat{e}(n) \) of the microphone signal \( e(n) \). The difference between \( e(n) \) and \( \hat{e}(n) \) is used to update both \( C \) and \( \mathbf{H} \). In this structure, the two estimation filters \( C \) and \( \mathbf{H} \) are considered as one long filter composed of two parts. The first part \( H \) has \( x(n) \) as its training signal while the second part \( \hat{C} \) has \( y(n) \) as a training signal. This can be seen by rewriting (4.40) as a vector multiplication as follows

\[
e(n) = \begin{bmatrix} \mathbf{x}^T(n) & \mathbf{y}^T(n) \end{bmatrix} \begin{bmatrix} \mathbf{h} \\ \mathbf{e} \end{bmatrix}. \tag{4.41}
\]

The condition for successful identification may be obtained by left multiplying (4.41) once by \( \mathbf{x}(n) \) and once by \( \mathbf{y}(n) \) and taking the expectation [125]

\[
E \left[ \begin{bmatrix} \mathbf{x}(n) & \mathbf{y}(n) \end{bmatrix} e(n) \right] = E \left[ \begin{bmatrix} \mathbf{x}(n) & \mathbf{y}(n) \end{bmatrix} \begin{bmatrix} \mathbf{x}^T(n) & \mathbf{y}^T(n) \end{bmatrix} \right] \begin{bmatrix} \mathbf{h} \\ \mathbf{e} \end{bmatrix}. \tag{4.42}
\]

The unknown filters \( \mathbf{h} \) and \( \mathbf{e} \) may, therefore, be calculated only if the correlation matrix given by the first factor of the right hand side of (4.42) is non-singular. This matrix is non-singular if its rank is \( N_h + N_c \), where \( N_h \) and \( N_c \) are the length of \( \mathbf{h} \) and \( \mathbf{e} \), respectively. This is fulfilled only if the length of the control filter \( W \) satisfies \( N_w \geq N_h + N_c + 1 \) [125]. This approach provides continuous tracking of both \( \mathbf{H} \) and \( \hat{C} \) simultaneously, which is advantageous for a frequently moving microphone since
both transfer functions change frequently. On the other hand, if the frequent changes are encountered only in \( C \) while \( H \) remains constant, the estimation of \( H \) represents extra overhead calculations.

When using the control signals as training signals for on-line identification, simultaneous identification of multichannel transfer functions such as that mentioned in Section 4.7.1.3 is not possible. This is due to the fact that all training signals \( \{y_{ml}(n) : l = 1, 2, \cdots, L\} \) are correlated with \( x(n) \) and, therefore, correlated with each other. Only simultaneous identification of the transfer functions \( \{\xi_{ml} : m = 1, 2, \cdots, M\} \) is possible. This is also true for multiple input signals, since each of the \( L \) control signals \( \{y_{ml}(n) : l = 1, 2, \cdots, L\} \) are correlated with all input signals \( \{x_k(n) : k = 1, 2, \cdots, K\} \).

### 4.7.3 Audio Reproduction and On-Line Identification

In a sound reproduction system and for large listener’s movements, the control filters are in great error and must be redesigned as mentioned before. Adaptive sound reproduction systems may redesign the filters by readapting them to their new optimum solutions at the new listener’s position. In this section, on-line adaptation of the control filters while the audio reproduction system is in operation is addressed. The new solutions require knowledge of the new electro-acoustic transfer functions \( C \) between the reproduction loudspeakers and the current positions of the listener’s ears. When microphones are placed in the listener’s ear canals, the above mentioned on-line identification techniques may be used to continuously measure the electro-acoustic transfer functions \( C \). In some applications, such as cross-talk cancellation, knowledge of the current \( C \) alone is sufficient to readapt the filters as discussed in Section 4.7.3.1. In other applications, such as virtual sound source synthesis, knowledge of the current matrix \( H \) is also required as will be shortly shown in Section 4.7.3.2.

#### 4.7.3.1 Tracking for Cross-Talk Cancellation

The desired response \( d(n) \) in the single channel cross-talk cancellation system shown in Fig. 4.26 is a delayed version of the input signal \( x(n) \) as mentioned in Section 3.3.7. The system response at the microphone \( d(n) \) is the sound signal that the listener actually hears. The error \( e(n) \) is formed by _electrically_ subtracting the microphones’ output from the desired response. This is in contrast with other applications such as active noise control where both \( d(n) \) and \( d(n) \) are acoustic waves that are added by the microphone. This simplifies both the on-line identification of \( C \) and the on-line adaptation of \( W \) for listener tracking. Furthermore, when the listener moves, only \( C \) has to be re-estimated, since \( H \) is always known pure delay.

Consider first the on-line identification of the electro-acoustic transfer function \( C \). Since the microphone’s output is not contaminated with any primary disturbances, the identification process is considerably simplified. No estimation of such disturbances is needed to be performed as in the general case described above. This simplification is shown in Fig. 4.26 for a single channel case when \( y(n) \) is used as a training signal for the identification process. Using an extra training signal for identification results in a slower convergence speed due to the disturbance component from \( y(n) \) at the microphone as mentioned in Section 4.7.1. This component is also readily estimated and subtracted from the microphone output since only the transfer function \( C \) is involved.

Turning the attention to the on-line adaptation of \( W \), since \( d(n) \) is in electric rather than acoustic form, subtracting \( d(n) \) from \( d(n) \) does not null the acoustical waves that are meant for the listener to hear. Therefore, the audio signal \( x(n) \) may be used as a training signal for the adaptation of the filter \( W \) as shown in Fig. 4.26. This may not be an option in other applications. A work around is discussed in Section 4.7.3.2 for virtual source synthesis systems.

#### 4.7.3.2 Tracking for Virtual Source Synthesis

A single channel virtual sound source synthesis system is shown in Fig. 4.27. The adaptive filter \( W \) in this system is designed off-line by playing a white noise signal \( s(n) \) through loudspeaker \( P \) and adjusting the filter coefficients to cancel the resulting sound at the microphone as mentioned in Section 3.3.6. Clearly, if this approach is used for on-line adaptation of \( W \) with \( e(n) \) as the music signal required to be played by the virtual source, the adaptation process will cancel the music at the microphone, and the listener will hear silence.

The straightforward approach to on-line adaptation of \( W \) is to use an extra training signal similar to the on-line identification mentioned in Section 4.7.1. The music signal \( x(n) \) is multiplied by -1 to compensate for the negative sign in the optimum
solution and added at the input of the filter $W$. When the filter adapts to its optimal solution, the acoustic response due to $x(n)$ at the microphone equals $h * x(n)$, which is the desired response for creating a virtual sound source at the position of loudspeaker $P$. The control signal $y(n)$ is used as a training signal for the on-line identification of $C$ and $H$ as discussed in Section 4.7.2. The white noise signal $s(n)$ is the extra training signal used for the on-line adaptation of the filter $W$. As not to degrade the quality of the music heard at the microphone, the noise $s(n)$ must be kept at a low amplitude. The music response at the microphone, therefore, forms a high level disturbance to the adaptation process of $W$. This can be alleviated by subtracting an estimate of this disturbing component from the microphone signal before using it in updating $W$ as shown in Fig. 4.27. Unlike in active noise control applications, the acoustic response due to the extra training noise at the microphone is masked by the much louder music. Therefore, the upper limit of the amplitude of $s(n)$ is defined by the masking properties of the human auditory system [148].

The above mentioned system assumes that a physical loudspeaker $P$ exists at the position where a virtual sound source is required. In a real-time virtual sound source synthesis system, such a physical loudspeaker is absent and the architecture shown in Fig. 4.27 can not be used. However, this system (with $x(n)$ absent) is very efficient for calculating and storing a set of filters $W$ for later use in virtual source simulation at different positions. A head tracking system may then be used to select one of the pre-calculated set of filters, through which the music signal should be filtered in real-time.

In many virtual sound source synthesis systems, the matrix of electro-acoustic transfer functions $H$ is available as a pre-measured and stored set of HRTFs from different directions. In such systems, the desired responses $g(n)$ may be calculated internally by filtering the music signals $x(n)$ through the appropriate set of transfer functions $H$, and the error vector $e(n)$ is calculated electrically as in the case of cross-talk cancellation. This allows the use of music signals as training signals for the adaptive filters $W$ similar to the structure shown in Fig. 4.26. Unlike in cross-talk cancellation, the optimum solution $W_{opt}$ in this case is dependent on both $H$ and $C$. Provided that a switching mechanism is integrated in the system to select the appropriate new matrix $H$ for every listener movement, the on-line identification may be employed to measure $C$ and the filters $W$ are adapted to the optimum solution for the new listener position. The need for a head tracking system may be avoided if the music signals are stored after processing them by the HRTFs. This reduces the system to a cross-talk canceller and the system shown in Fig. 4.26 may be used.

### 4.8 Fast Adaptive Algorithms

The convergence speed of the single channel filtered-$x$ LMS algorithm discussed in Section 3.3.1 is known to be slow due to the electro-acoustic transfer function between the secondary loudspeaker and the microphone (the secondary acoustic path) in front of the adaptive filter. The same also holds for the adjoint LMS algorithm discussed in Section 4.3.1. It is shown in [144] that the convergence properties of the LMS adaptive algorithm can be improved by decorrelating its input signal. This process results in convergence properties that are independent of the input signal statistics. However, the decorrelation process is not obvious in the filtered-$x$ LMS (FX) case, since the update suffers from coloration not only due to the input signal but also due to the estimate of the secondary acoustic path used to calculate the filtered-$x$ signal. In this section, two methods are presented to improve the convergence properties of that algorithm [67, 132]. A straightforward application of the decorrelation principle to the update equation of the FX algorithm leads to the decorrelated filtered-$x$ LMS (DFX) algorithm described in Section 4.8.1. The DFX algorithm suffers from the problem of division by a small number, which is enhanced by the amplitude spectrum of the secondary acoustic path. Partially decorrelating the filtered input signal leads to the all-pass filtered-$x$ (APFX) algorithm presented in Section 4.8.2. The convergence properties of the different decorrelation methods are compared in Section 4.8.3.

#### 4.8.1 The Decorrelated Filtered-$X$ Algorithm (DFX)

The DFX algorithm is derived by decorrelating the conventional FX algorithm shown in Fig. 4.28. Since the decorrelation process is best explained in the frequency domain, the signals in Fig. 4.28 are transformed to the frequency domain, using the $F$ blocks, where filtering and adaptation operations are performed. The update block in this
Equation (4.47) also shows that the trajectory taken by $W$ from its initial value to its final solution is governed by the term $(I - 2\mu P_{X_f})$. The slope taken by each filter coefficient is dependent on the power contained in $X_f$ at the corresponding frequency bin. When the ratio $\max(P_{X_f}) / \min(P_{X_f})$ is high, some filter coefficients converge to their final values much slower than the others. This is known as the eigenvalue disparity problem [144]. This problem is enhanced in the FX case, since the spectrum of $X_f$ is colored not only by the input signal $X$ but also by the estimate of the secondary acoustic path $C$. This latter includes, apart from the acoustic response between the secondary source and the microphone, the response of the secondary loudspeaker, the microphone, the ADC and the DAC used in the identification process. The former two have band-pass characteristics and the latter two contain anti-aliasing and reconstruction filters, which make $C$ a band-pass process with high energy ratio between mid-band and out-of-band frequencies. It is those frequencies with low energy that degrade the convergence properties of the conventional FX algorithm.

To improve the convergence, this coloration must be removed, which is usually done by multiplying the second term in (4.45) by the inverse of a decorrelation matrix $P$, which leads to the update equation of the DFX

$$\tilde{W}_{new} = W_{old} - 2\mu P_{X_f} X_f E.$$  

Following the same steps as above, the dynamic convergence equation for the DFX can be obtained

$$\tilde{W}_{new} = (I - 2\mu P_{X_f}) W_{old} - 2\mu P_{X_f} (C^{-1} H).$$  

It follows from (4.49) that for all filter coefficients to approach their final values with the same speed, the decorrelation matrix should be chosen such that $P = P_{X_f}$ and (4.49) becomes

$$\tilde{W}_{new} = (I - 2\mu) \tilde{W}_{old} - 2\mu (C^{-1} H),$$

which shows a convergence behaviour that is dependent neither on the input signal $X$ nor on the secondary acoustic path $C$.

Although the DFX algorithm as given by (4.48) and (4.50) is completely decorrelated, it requires calculating and multiplying by the matrix $P_{X_f}$. In practice, this is done by maintaining the time average of a power vector $P_{X_f}$ and dividing the adaptation constant by that vector. In most practical cases, $P_{X_f}$ contains elements of small values due to the band-pass nature of $C$ as mentioned above. Since the power estimate is never accurate, dividing by such small numbers may apply large changes to the update constants and drive the update unstable. To avoid this instabiliy, the regularisation principle is often applied, a simple form of which is to divide by $(P_{X_f} + \delta I)$, where $\delta$ is a small scalar and $I$ is a vector of ones. This, however, limits the decorrelation of the slowest modes (corresponding to the low power frequencies) and degrades the convergence as will be shown in Section 4.8.3.
4.8.2 The All-Pass Filtered-X Algorithm (APFX)

Recognising that the slowest modes are due to the band-pass nature of \( C \), it is, therefore, preferred to handle the decorrelation of \( C \) differently. Assuming that the input signal is well behaved, it can still be decorrelated by multiplying by its inverse power matrix \( P_X^{-1} \). Decorrelation of \( C \) is then performed partly by multiplying by \( |C^*|^{-1} \) instead of \( |C^*C|^{-1} \), so that the decorrelation matrix \( P^{-1} \) in (4.48) now equals \( P_X^{-1} |C^*|^{-1} \) and the approximate decorrelated dynamic convergence equation becomes

\[
\mathbf{w}_{\text{new}} = (I - 2\mu P_X^{-1} |C^*|^{-1} P_{X_f}) \mathbf{w}_{\text{old}} - 2\mu P_X^{-1} |C^*|^{-1} P_{X_f} (C^{-1} \mathbf{h}).
\]

(4.51)

Substituting \( P_{X_f} = C^* X^* X C \) in (4.51) and simplifying, we obtain the dynamic convergence equation of the APFX algorithm

\[
\mathbf{w}_{\text{new}} = (I - 2\mu C^*_a C) \mathbf{w}_{\text{old}} - 2\mu C^*_a C (C^{-1} \mathbf{h}),
\]

(4.52)

in which \( C^*_a = |C^*|^{-1} C^* \) is an all-pass filter having a phase response equal to that of \( C^* \). Comparing (4.52) and (4.50) shows that the APFX is still colored by the amplitude response of \( C \) (\( C^*_a \) has a flat amplitude response and, therefore, no coloration effect). Despite this residual coloration, it is shown in section 4.8.3 that the convergence properties of the APFX algorithm are significantly improved compared to the conventional FX algorithm. The update equation of the APFX algorithm can be derived from (4.51) to be

\[
\mathbf{w}_{\text{new}} = \mathbf{w}_{\text{old}} - 2\mu P_X^{-1} X_{fa} e,
\]

(4.53)

where \( X_{fa} = C_a X \) is a new filtered-x signal, obtained by filtering the input signal \( X \) by the all-pass filter \( C_a \) and hence the algorithm's name. Thus, in the APFX, the partial decorrelation of \( C \) is performed by calculating the filtered-x signal using \( C_a \) instead of \( C \) and no division by the power spectrum of \( C \) is needed.

In many applications, it is known a priori that the input signal \( X \) has a flat amplitude spectrum. In these cases, it is sufficient to decorrelate \( C \) only by updating the filter coefficients using \( X_{fa} \) without multiplying by \( P_X^{-1} \). This improves the convergence properties significantly without any extra computations, since \( C_a \) can be estimated in the initialisation stage instead of \( C \) while decorrelation of the input spectrum costs calculating and multiplying by \( P_X^{-1} \) during the update process. Another special case of the APFX algorithm is when the input signal \( X \) contains the dominant slow modes or comparable slow modes to those in \( C \). In this case, decorrelation of \( X \) using \( P_X^{-1} \) suffers from the same problems discussed in Section 4.8.1 and the same partial decorrelation principle using phase information only may be performed.

4.8.3 Performance Comparison

The convergence properties of the above presented algorithms have been extensively tested by means of computer simulations. The simulations implemented the single-channel noise canceller shown in Fig. 4.28 using Block Frequency Domain Adaptive Filters (BFDAF) [55, 130]. The performance was studied using different acoustic transfer functions for \( \mathbf{g} \) and \( \mathbf{h} \). The results presented here were obtained using KEMAR Head-Related Impulse Responses (HRIR) in the horizontal plane (see Section 2.1.7). The impulse response \( \mathbf{g} \) was the HRIR of the left ear from azimuth angle \( \alpha = 35^\circ \), while \( \mathbf{h} \) was that from \( \alpha = 35^\circ \). Each impulse response was an FIR filter of 512 coefficients measured at 44.1 kHz sampling frequency. Fig. 4.29 shows the time, amplitude, and phase responses of the secondary acoustic path \( \mathbf{g} \). In the following simulations, the adaptive filter length was 1024, the new input samples in each block was 1025, and the Fourier transformation length was 2048. Each simulation updated the adaptive filter \( \mathbf{w} \), starting from zero initial condition, according to one of the above presented algorithms, and the mean squared error (time average of \( 10 \log_{10}(e^2) \)) was plotted against time. To perform honest comparison, the adaptation constant \( \mu \) was adjusted to give the same theoretical final misadjustment with all algorithms.

4.8.3.1 Performance for a White Noise Training Signal

With a white noise input signal \( X \) of zero mean and variance 1, the adaptive filter \( \mathbf{w} \) was updated according to each of the above presented algorithms. The learning curves for three specific cases are shown in Fig. 4.30. In this figure, curve 1 is the learning curve for the conventional FX algorithm given by (4.45). Curve 2 is that for the APFX given by (4.53). Since the input signal is white noise, multiplication by \( P_X^{-1} \) does not count, and was not performed. Curve 3 is the learning curve for the DFX update given by (4.48) with \( P_{X_f} + \delta I \) and \( \delta = 0.001 \).

From Fig. 4.30, the improvement achieved by the APFX over the conventional FX algorithm is evident. This is attributed to the considerable coloration added to the input signal by \( \mathbf{g} \) and partially removed by using the APFX algorithm. From experiments using other transfer functions [67], it is observed that the improvement
obtained by the APFX over FX increases as the coloration introduced by $e$ increases.

In this experiment, it was noticed that updating the filter according to the DFX without regularisation, or with $\delta$ smaller than about $10^{-5}$, results in an unstable update. This may be attributed to the pinna notch at about 8 kHz. At this frequency, the DFX divides the adaptation constant by about $4 \times 10^{-3}$, which causes the problem of dividing by a small number mentioned earlier. By limiting the smallest number the decorrelation process uses, the algorithm is stabilised and gives better performance than the APFX. This is expected since the APFX equivalently divides by the square root of the power of $e$ at each frequency bin while the DFX divides by the whole power. Increasing $\delta$ above a certain value scales the adaptation constant lower at other frequencies without performing the required decorrelation at the notch frequency. This may explain the slower initial convergence rate and the smaller final misadjustment of curve 3.

From Fig. 4.30, both the APFX and the DFX in this example decrease the error by 20 dB in about 120000 samples. This corresponds to 2.7 seconds at 44.1 kHz sampling frequency, which is about half the time required for the conventional FX algorithm to reduce the error to the same level. For tracking the listeners' movements using extra white noise (see Section 4.7.1), it is sufficient that the adaptive filter reaches a reasonable solution before the listener starts perceiving the error in the filters due to the movements. Since the adaptive filter in the tracking situation starts off from an initial condition that is close to the optimal solution, and only high frequency coefficients are expected to change, faster convergence time is expected. However, psychoacoustical experiments are necessary to determine whether tracking in 3D sound systems can be performed using adaptive filters.

### 4.8.3.2 Performance for an MA(2) Training Signal

To examine the performance of the different decorrelation methods for colored input signals, the above mentioned experiment was repeated after filtering the white noise signal by a moving average filter MA(2) with complex conjugate zeros at $\pm j 0.995$ before using it as the input signal $X$. The power spectrum of $C$, $X$, and $X_f$ show that the amount of coloration inserted by each component. The adaptive filter $W$ was updated according to each of the above presented algorithms. The learning curves obtained in the previous cases are shown in Fig. 4.32. Curve 1 is the learning curve for the conventional FX algorithm given by (4.45). Curves 2, 3, and 4 are the learning curves for the APFX update equation (4.53) without multiplying by $P_X^{-1}$, the APFX update equation (4.53), and the DFX update equation (4.48) with $P_{X_f} + \delta I$, $\delta = 0.001$, respectively.

From Fig. 4.32, the effect of decorrelating each component is clear. Since the notch in the input signal is spread over more frequencies than that in the secondary path, little improvement is observed by partial decorrelation of $e$ only (curve 2). However, the improvement is evident when the decorrelation of both the input signal and the secondary path is considered (curve 3). The scaling of the adaptation constant of the fast modes due to adding $\delta = 0.001$ is clear from the difference in the initial convergence rate between curve 3 and 4. Those results are important for tracking of listeners' movements using the control signals as mentioned in Section 4.7.2, where it is important to use a decorrelation method that considers both the input signal and the secondary path.

### 4.9 Summary

Several techniques have been introduced in this chapter to deal with moving listeners. Two main approaches have been presented: the first is enlarging the zones of equalisation to cope with small listener movements, and the second is tracking the listener when he moves beyond the equalisation zones.

Enlarging the equalisation zones is achieved by designing the control filters to have a valid solution in wider areas around the control points. Several methods have been presented to reach this goal. These methods include approximations to spatial derivative constraints, employing multiresolution control filters, increasing the number of reproduction loudspeakers, and loudspeaker positioning in the listening space. The improvement brought about by each of these methods has been shown by computer simulations in anechoic and reverberant environments.

Derivative constraints require analytical expressions for the acoustic transfer functions, which are difficult to obtain in reverberant environments. Approximations to derivative constraints in reverberant environments have been presented in this chapter. However, these approximations are not valid in severe reverberant conditions. Another approximation to derivative constraints is to control the sound field at discrete points in the proximity of the original required control points (difference constraints). This requires using a microphone at each newly introduced control point, which leads to considerable increase in the computational complexity. Furthermore,
Chapter 4: Robustness of 3D Sound Systems

the microphones have to be closely positioned to obtain non-splitting zones of equalisation. Although the spatial high-pass filters introduced in this chapter solve the problem of increasing complexity, they still require measuring several transfer functions at adjacent points.

Examining the behaviour of the above mentioned methods shows that the constraints force the control filters to assume solutions that increasingly deviate from the exact solutions with increasing frequency. Similar solutions may be obtained by employing filters having decreasing frequency resolution with increasing frequency. Such filters automatically ignore more spectral details with increasing frequency due to their coarser frequency resolution. The improvement in the robustness of 3D sound systems by using such multiresolution filters is also shown by computer simulations.

The zones of equalisation can also be improved by improving the system controllability and rearranging the acoustic transducers to obtain a well-conditioned system. Using computer simulations, it has been shown that increasing the number of reproduction loudspeakers improves the system robustness in both anechoic and reverberant environments. Positioning the reproduction loudspeakers in the listening space is also studied using simulations.

Tracking the listeners has been approached by updating the control filters to correct the errors introduced due to listeners' movements. Such on-line adaptation requires the adaptive filters to reach their new solutions, which correspond to the current listeners' positions, before the listeners perceive the errors. This requires improving the convergence speed of the conventional adaptive algorithms. Two modifications to the conventional filtered-x algorithm have been introduced for this purpose, the decorrelated filtered-x, and the all-pass filtered-x algorithms. The improvement in convergence speed due to these modifications compared to the conventional algorithm has been shown by computer simulations. On-line estimation of the electro-acoustic transfer functions, that is required for the convergence of the adaptive algorithms, has been also considered in this chapter. Several method for on-line measurements of those transfer functions with and without extra training signals have been discussed.

Chapter 5

Fast Multiresolution Analysis

5.1 Introduction

Many real-time audio applications perform convolution and correlation operations in the frequency domain for efficiency reasons [114, 134]. The FFT is usually used to transform the time signals back and forth to the frequency domain. Interpreting the FFT operation as a filterbank, the filters have constant bandwidth, and their centre frequencies are equally spaced. For audio signals, however, the FFT tends to give too little spectral resolution at low frequencies, and too much spectral resolution at high frequencies, due to the non-uniform spectral resolution of the human auditory system as mentioned in Section 2.1.8. In both models discussed in Section 2.1.8, the human auditory system is considered to perform a constant-Q (percentage bandwidth) analysis in the largest part of the audible frequency range.

Constant bandwidth analysis has several undesired effects in audio signal processing. Since all frequencies are analysed using the same bandwidth, this model is less suitable for representing signals of time-varying spectra [8, 120]. Constant resolution analysis also fails to estimate many of the important psychoacoustical phenomena such as spectral masking. Excessive high frequency resolution can also reveal spectral fine structure that is undetectable by the human ear. Most important for 3D audio systems, constant resolution results in control filters that are sensitive to spatial variations due to the short wavelength at high frequencies, as shown in Chapter 4. Therefore, a better signal representation resembling that used by the human auditory system is preferred in audio systems. It is the goal of this chapter to investigate such signal representations. To be of any use in real-time adaptive 3D sound systems, the new signal representation must be calculated as fast as the FFT. Therefore, the study in this chapter is restricted to fast algorithms that can be implemented in real-time.

The specifications a signal representation (transformation) must meet to be considered a candidate in the class of signal representations we are seeking are defined in Section 5.2. In these specifications, the signal transformation is required to support the convolution property, so that filtering operations may be performed efficiently by element-wise multiplication in the transformation domain. This convolution property
is explained in more details in Section 5.3.

Several multiresolution signal representations that are of fundamental importance and widely used in audio systems are briefly reviewed in Section 5.4. Most of those methods have been developed in other contexts such as optics, image processing, seismic exploration, and biomedical signal processing. The suitability of those transformations for audio applications in general and real-time adaptive 3D sound systems in particular is also discussed in Section 5.4.

Interpreting the output of an FFT algorithm for uniformly-spaced samples of a time signal as uniformly-spaced samples of the signal’s spectrum raises the question of what the correct interpretation of the output of an FFT algorithm for non-uniformly-spaced time samples should be. Intuitively, one tends to say that it must be non-uniformly-spaced samples of the signal's spectrum. If the intuitive answer is correct, this may represent an efficient method for calculating a large class of multiresolution transformations. Using a rigorous mathematical description of unitary axis warping, it is shown in Section 5.5 that the intuitive answer is indeed the correct answer. This axis warping transformations can be implemented simply by non-uniformly sampling the signal under consideration according to a warping function and is, therefore, attractive for real-time applications. Another advantage of this approach is that a family of signal representations is obtained, the characteristics of each member of this family is defined by the warping function. The theoretical aspects of unitary axis warping are discussed in Sections 5.5.1 through 5.5.4. The logarithmic warping function is of special interest for audio applications since it results in a signal representation that has coarser frequency resolution with increasing frequency. In the continuous-time domain, this logarithmic warping maps the Fourier transform onto the Mellin transformation that proved useful in many fields where scale-invariance is required. Due to this property, the logarithmic warping is studied in detail in Section 5.5.5. Implementation of the warping transformations using non-uniform sampling is presented in Section 5.5.6. Finally, an example of modifying a real-time filtering algorithm to accommodate the warping technique is discussed in Section 5.5.7.

5.2 Specifications

Signal representations (or transformations) relevant to real-time audio signal processing in general, and capable of improving the robustness of 3D sound systems in particular, must meet several requirements. The essential requirements are listed below.

1. **Multiresolution**: The transformation must perform multiresolution spectral analysis similar (but need not be identical) to that performed by the human auditory system. This increases the ability of an audio system to estimate important psychoacoustical phenomena and decreases the system’s sensitivity to spatial variations.

2. **Linearity**: Many audio signal processing applications, such as active sound control, rely on the superposition principle, which in turn assumes linearity. Therefore, the signal transformation is required to be linear.

3. **Invertibility**: The transformation must be invertible (non-singular). This is required for real-time signal processing operations such as filtering. Signals are transformed into a spectral domain where filtering is performed by element-wise multiplication and the result is transformed back to the time domain. Without an inverse transformation, no such operation is possible.

4. **Fast implementation**: The discovery of the FFT algorithm made it possible to perform the convolution and correlation operations faster in the frequency domain. Therefore, the chosen transformation is required to keep this speed advantage when those operations are performed in the new transformation domain.

5. **Support for the convolution property**: This is essential if filtering is required to be implemented by element-wise multiplications in the transformation domain. Due to its importance, this property is further explained in Section 5.3.

5.3 The Convolution Property

The linear convolution of two sequences \( x(n) \) and \( h(n) \) each of length \( N \) results in the sequence \( y_L(n) \) of length \( 2N - 1 \) given by

\[
y_L(n) = \sum_{i=0}^{2N-1} x(i) h(n-i), \quad n = 0, 1, \ldots, 2N - 1.
\]

On the other hand, the cyclic convolution is defined only for periodic signals. Therefore, the cyclic convolution of two sequences \( x(n) \) and \( h(n) \) each of length \( N \) can be calculated only if the two sequences are repeated to construct periodic sequences of period \( N \) samples. Alternatively, indexes may be calculated modulo \( N \) to obtain the length \( N \) sequence \( y_c(n) \) of cyclic convolution result,

\[
y_c(n) = \sum_{i=0}^{N-1} x((i))_N h(((n-i))_N), \quad n = 0, 1, \ldots, N - 1,
\]

where \((i)_N\) represents modulo \( N \) arithmetic.

A transformation \( T \) has the cyclic convolution property if the transform of the cyclic convolution of two sequences equals the product of their transforms,

\[
Y_C = X_C \otimes H_C^*,
\]
Chapter 5: Fast Multiresolution Analysis

where \( \mathbf{Y}_x = \mathbf{T} \mathbf{Y}_x, \mathbf{X}_x = \mathbf{T} \mathbf{x}, \mathbf{H}_x = \mathbf{Th}, \) and \( \odot \) denotes element-wise multiplication. This definition assumes that the sequences are periodically extended with period \( N \) or the indexes are evaluated modulo \( N \). For the \( N \times N \) transformation matrix \( \mathbf{T} \) to satisfy the cyclic convolution property, the elements of \( \mathbf{T} \) have to satisfy the following three conditions [7]:

1. All \( t_{k,m} \) elements of \( \mathbf{T} \) must be \( N^{th} \) roots of unity, i.e. \( (t_{k,m})^N = 1 \).
2. All the \( t_{k,1} \) elements (the second column of \( \mathbf{T} \)) must be distinct for the matrix to be non-singular. Since there are only \( N \) distinct \( N^{th} \) roots of unity, \( t_{k,1} \) must be those \( N \) distinct roots.
3. The rows of \( \mathbf{T} \) are constructed such that \( t_{k,m} = (t_{k,1})^m \), i.e. the \( m^{th} \) element in any row must be the second element in the same row raised to the \( m^{th} \) power.

If \( t_{1,1} = \alpha, (t_{1,1})^N = 1 \) and the rows are arranged such that \( t_{k,m} = (t_{1,1})^m \), then the elements of \( \mathbf{T} \) can be written as \( t_{k,m} = \alpha^m, k, m = 0, 1, \ldots, N - 1 \), and the matrix \( \mathbf{T} \) has the following structure:

\[
\mathbf{T} = \begin{bmatrix}
\alpha^0 & \alpha^0 & \alpha^0 & \cdots & \alpha^0 \\
\alpha^0 & \alpha^1 & \alpha^2 & \cdots & \alpha^{N-1} \\
\alpha^0 & \alpha^2 & \alpha^4 & \cdots & \alpha^{2(N-1)} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\alpha^0 & \alpha^{N-1} & \alpha^{2(N-1)} & \cdots & \alpha^{(N-1)(N-1)}
\end{bmatrix} \quad (5.4)
\]

This structure makes the transformation \( \mathbf{T} \) orthogonal and the elements of the inverse transformation \( \mathbf{T}^{-1} \) are given by \( N^{-1} (t_{k,m})^{-1} \). The structure in (5.4) is the only one supporting the cyclic convolution property. The DFT with \( \alpha = e^{-j2\pi/N} \) is the only transformation supporting the cyclic convolution property in the complex number field [7]. Furthermore, any transformation supporting the cyclic convolution property and, therefore, having the structure of (5.4), has a fast computational algorithm similar to the FFT, provided that \( N \) is highly composite. Therefore, the DFT is the only transformation in the complex number field that supports the cyclic convolution property and has a fast computational algorithm [7].

It is worthwhile to mention, however, that transformations supporting the cyclic convolution property in other fields do exist. Examples of these transformations are the Fermat Number Transformation (FNT) [7, 138] and the Mersenne Number Transformation (MNT) [122]. These transformations have similar properties to those known for the DFT and can be used to implement the convolution operation very efficiently [7]. The drawback of these transformations is the lack of a physical meaning coupled to them as the spectral analysis interpretation for the DFT. Since we are not only interested in fast convolution in the transformation domain but also aim at utilising the human auditory system behaviour, which is well understood in the frequency domain, only transformations in the complex number field will be considered here.

The Linear convolution of two finite length sequences can be implemented using cyclic convolution by padding the two sequences with zeros [114]. If support for linear convolution only is desired, a more general class of transformations exists, which does not restrict \( t_{k,1} \) to be the \( N^{th} \) roots of unity but still requires that \( t_{k,m} = (t_{k,1})^m \). For this class of transformations to be invertible, \( t_{k,1} \) must also be distinct. The general structure of the transformation matrix \( \mathbf{T} \) of this type is given by

\[
\mathbf{T} = \begin{bmatrix}
\alpha_0^0 & \alpha_0^0 & \alpha_0^2 & \cdots & \alpha_0^{(N-1)} \\
\alpha_1^0 & \alpha_1^1 & \alpha_1^2 & \cdots & \alpha_1^{(N-1)} \\
\alpha_2^0 & \alpha_2^2 & \alpha_2^4 & \cdots & \alpha_2^{2(N-1)} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\alpha_{N-1}^0 & \alpha_{N-1}^{N-1} & \alpha_{N-1}^{2(N-1)} & \cdots & \alpha_{N-1}^{(N-1)^2}
\end{bmatrix} \quad (5.5)
\]

The inverse transformation in this case does not have a simple structure and, in general, a fast algorithm for calculating the transformation does not exist. An example of this class of transformations in the complex number field is the non-uniform discrete Fourier transformation discussed in Section 5.4.3.

5.4 Related Work

In this section, some important transformations are briefly mentioned. Starting with transformations that are of fundamental importance, the short time Fourier transformation and the wavelet transformation are discussed in Sections 5.4.1 and 5.4.2, respectively. The non-uniform discrete Fourier transformation is discussed in Section 5.4.3. The digital frequency warping technique has recently found great interest in audio applications, therefore, a detailed discussion of this technique is given in Section 5.4.4. Section 5.4.5 summarises several other published methods to perform non-uniform spectral analysis.

5.4.1 The Short Time Fourier Transformation

A widely used time-frequency representation of signals with time-varying spectra is the Short Time Fourier Transformation (STFT) defined in the continuous-time domain as [8, 119, 120]

\[
F(t,f) = \int_{-\infty}^{+\infty} x(\tau) h^*(\tau-t) e^{-j2\pi f(\tau-t)} d\tau, \quad (5.6)
\]

which is the Fourier transform of \( x(\tau) \) seen through the window function \( h(\tau) \) centred at time \( t \). The STFT can be considered as a similarity measure between the signal \( x(\tau) \) and the transformation kernel \( h(\tau-t) e^{-j2\pi f(\tau-t)} \). The latter is a constant envelope with increasing number of oscillations as the analysis frequency increases.
In the frequency domain, this amounts to frequency shifted (modulated) versions of a prototype low-pass filter given by the Fourier transform of the window \( h(t) \). Therefore, the STFT may be interpreted as a constant bandwidth filterbank with \( F(t, f_c) \) being the output of the band-pass filter centred at \( f_c \) as a function of time. Provided that the window \( h(t) \) is of finite energy, many exact inversion formula exist. Although efficient implementations of the STFT using the FFT algorithm has been previously published \[49, 118\], the constant bandwidth analysis does not result in a multiresolution spectra that are preferred in audio applications.

### 5.4.2 The Wavelet Transformation

The Wavelet Transformation (WT) replaces the frequency shifting operation of the STFT by a scaling operation and is, therefore, considered as a time-scale representation rather than time-frequency one. The continuous-time WT is defined as \[40\]

\[
W(t, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(\tau) h^* \left( \frac{\tau - t}{a} \right) \, d\tau,
\]

where the mother wavelet \( h(t) \) is localised in time. Similar to the STFT, the WT can be considered as a similarity measure between the signal \( x(\tau) \) and the transformation kernel \( (1/\sqrt{a}) h \left( \frac{\tau - t}{a} \right) \). The latter is a time-shifted and scaled version of the mother wavelet. The envelope is narrowed as higher frequencies are analysed, while the number of oscillations remains constant. In the frequency domain, this amounts to using a dilated or compressed version of a band-pass filter whose relative bandwidths are constant (constant-Q filterbank).

Wavelet transforms can be calculated using a fast method, the Fast Discrete Wavelet Transformation (FDWT). The FDWT is performed by successively filtering the signal through a pair of Quadrature Mirror Filters (QMF). This divides the signal into two parts, each covering half the original bandwidth. Both signals are then downsampled by a factor of two. The high frequency part is taken as the highest order wavelet coefficients, while the low frequency part goes through the same process of filtering and downsampling, which yields one more set of wavelet coefficients. This process is repeated until only one sample is left, which is taken as the lowest order coefficient. However, the FDWT produces much coarser resolution at high frequencies than needed in most audio applications. Better resolution may be obtained by decomposing the output of both QMF filters, which is called wavelet packets. This increases the transformation complexity and destroys the constant-Q structure. For this reason, warped wavelets \[62, 63\] that are developed using the digital frequency warping technique (see Section 5.4.4) have found their applications in audio signal processing.

### 5.4.3 The Non-Uniform Discrete Fourier Transformation

The Non-uniform Discrete Fourier Transformation (NDFT) \[101\] may be considered as a generalisation of the DFT. The DFT is obtained by sampling the Z-transform of a sequence \( x(n) \) of length \( N \) at \( N \) equally spaced points in the unit circle in the Z-plane. The NDFT of the same sequence is defined as samples of the Z-transform at \( N \) distinct points located arbitrarily in the Z-plane

\[
X[z_k] = \sum_{n=0}^{N-1} x(n) z_k^{-n}, \quad k = 0, 1, \ldots, N - 1,
\]

where \( X[z_k] \) is the Z-transform of \( x(n) \) evaluated at \( z = z_k \). The set of points \( \{z_k \mid k = 0, 1, \ldots, N - 1\} \) are distinct points arbitrarily chosen in the Z-plane. This can be expressed in a matrix notation as

\[
X = D_z x,
\]

where

\[
D = \begin{bmatrix}
z_0^0 & z_0^{-1} & z_0^{-2} & \cdots & z_0^{-N+1} \\
z_1^0 & z_1^{-1} & z_1^{-2} & \cdots & z_1^{-N+1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
z_{N-1}^0 & z_{N-1}^{-1} & z_{N-1}^{-2} & \cdots & z_{N-1}^{-N+1}
\end{bmatrix}
\]

The transformation matrix \( D \) is of the type given in equation 5.5, therefore, the NDFT supports the linear convolution property only, as mentioned in Section 5.3. This also suggests that the NDFT does not have a fast computation algorithm, which is essential for real-time implementation. Unlike the inverse DFT, which is obtained by calculating the Hermitian transpose of the forward transformation, the inverse NDFT has to be obtained by inverting the matrix \( D \). Since \( D \) is a Vandermonde matrix, the inverse \( D^{-1} \) exists provided that the \( N \) sampling points are distinct. Furthermore, if the sampling points are chosen in complex conjugate pairs, then the transformation shows similar symmetry properties as those known for the DFT with odd, even or real sequences. Finally, the NDFT reduces to the DFT when the \( N \) sampling points are chosen at equally spaced angles on the unit circle.

Since the NDFT is missing the special symmetry of \( \mathbb{Z}_N \), which is exploited in deriving the different FFT algorithms, it has to be evaluated using matrix multiplication as in (5.9). A straightforward implementation of (5.9) requires \( N^2 \) complex multiplications and \( N^2 - N \) complex additions. For real signals and when the sampling
It has been mentioned in Section 5.3 that the DFT is the only transformation in the complex number field that supports the cyclic convolution property and, therefore, the DFT is the only transformation in this field having a fast computational algorithm [7]. To obtain a non-uniform spectral transformation that has a fast computational algorithm, one may search for a linear transformation $W$ that transforms a uniformly sampled time sequence $f(n)$ to another time sequence $g(m)$ such that the frequency axis $\Omega$ defined for $f(n)$ is nonlinearly mapped to the frequency axis $\omega$ defined for $g(m)$ [26, 32, 78, 112, 113]. Evaluating the DFT of $g(m)$ results in a uniform spectral sequence $G(\omega)$, which also corresponds to a non-uniform spectral sequence on the $\Omega$ axis.

The above mentioned warping principle can be explained from the point of view of the NDFT discussed in Section 5.4.3. The NDFT matrix $D$ in (5.9) can be factored into the two matrices $W$ and $F$ such that $D = WF$, where $F$ is the DFT matrix and $W$ is the warping matrix as shown in Fig. 5.1. The warping matrix $W$ can, therefore, be calculated from $W = F^{-1} D$. The required spectral coefficients are obtained by choosing the sampling points $\{z_0, z_1, \ldots, z_{N-1}\}$ in the NDFT on the unit circle. Furthermore, if the symmetry properties of the DFT must be maintained, the sampling points must be chosen in complex conjugate pairs as discussed in Section 5.4.3.

Since the sampling points in the NDFT are allowed to be arbitrarily chosen, the transformation $W$ is, in general, not orthogonal. If the transformation $W$ is restricted to be orthogonal, and the mapping is restricted such that $\omega$ changes by $2\pi$ when $\Omega$ changes by $2\pi$, the warping transformation may be implemented by the network shown in Fig. 5.2 [113]. The inverse warping may also be implemented by the same network after replacing each occurrence of $\lambda$ by $-\lambda$. The sequence $g(m)$ is obtained in Fig. 5.2 by sampling the output of the network at $n = 0$, i.e. $g(m) = \hat{g}(m,0)$. For a running evaluation of the spectrum of $f(n)$, the output is sampled every $N$ samples, and the DFT is calculated for the sequence $g(m)$ every $N$ samples. With $\lambda$ real and between $0$ and $1$, the warping results in a decreasing frequency resolution with increasing frequency, which resembles the analysis performed by the human auditory system. For $\lambda$ negative and between $-1$ and $0$, the warping results in an increasing frequency resolution with decreasing frequency. For complex values of $\lambda$, higher spectral resolution at an intermediate frequency is obtained.

Using the network shown in Fig. 5.2 to generate the warped time sequence $\hat{g}(m,n)$ requires two real additions and two real multiplications for each $\hat{g}(m,n)$, except for the first two output points. For $\hat{g}(0,n)$ only one real addition and real multiplication are required, while one real addition and two real multiplications are required for $\hat{g}(1,n)$. If the circuit consists of $M$ sections, the total number of real multiplications and additions per sample is then $2M - 1$ and $2M - 2$, respectively. Since an FFT is calculated every $N$ samples, the total number of real multiplications and additions per FFT is then $2MN - N$ and $2MN - 2N$, respectively, in addition to the operations required for performing the FFT. For $M = N$, the computation complexity is proportional to $N^2$, which limits the usability of this technique to off-line applications. However, digital frequency warping has found several applications such as modelling HRTFs [84, 85, 86], linear prediction [96], and audio coding [80].

### 5.4.5 Other Methods

Several authors have published methods to perform non-uniform spectral analysis using block processing algorithms. In 1976, Harris [81] processed the FFT output with spectral windows of constant time duration but adjustable bandwidths centred at the nearest FFT bin to the required analysis frequency. FFT pruning [22, 99, 128, 133] also falls into this category of processing. The pruning is done by discarding samples of the output of the FFT. These methods do not alter the time resolution of the signal and are, therefore, information destructive, which means that they have no inverse transformations.
Another group of authors concentrated on calculating a constant-Q spectral analysis and synthesis. In 1978 Youngberg and Boll [145] presented a constant-Q integral transformation, which is a generalisation of the STFT, by letting the analysis window be a variable in the product of time and frequency. In 1971 and 1972 [74, 75], Gambardella showed that if a signal undergoes a short time spectral analysis via a continuous set of constant-Q band-pass filters, this process can be mathematically represented through an integral transformation that can be inverted by means of the Mellin transformation. The above methods must be implemented as constant-Q filterbanks, which are computationally expensive, making them less attractive for real-time applications. In 1978 and 1983, Kates [88, 89] used an exponentially decaying window whose argument is a constant times the product of time and frequency to calculate the constant-Q integral. For this specific window, the integral can be evaluated using the chirp Z-transformation. Once again, the chirp Z-transformation is too computationally expensive for use in real-time filtering. In 1991 and 1992, Brown [34, 35] has developed a 1/24 octave discrete constant-Q transformation for analysing music signals. This method has no inverse and is, therefore, only suitable for spectral analysis applications.

In 1980, Toaney et al. [135] exploited the perfect fifth numerical coincidence \( \frac{3}{2} \approx 2.15 \) to develop what they called the tempered Fourier transformation. They used four analogue-to-digital converters working at sampling rates separated by \( \frac{2}{7} \approx \frac{31}{12} \) to generate a 1/3 octave analysis. The complexity of using four different sampling rates and the sequence of computations are the difficulties of this method. Complexity and invertibility are not mentioned in [135].

### 5.5 Unitary Coordinate Transformations

The concept of coordinate transformation, better known as warping, has been used extensively in many applications such as coding, adaptive filters, and dynamic time warping of speech signals. Recently, Baraniuk and Jones [13, 14, 16, 18] have extended the application domain of unitary axis warping to joint signal representations of arbitrary variables including time-frequency and time-scale distributions. The main goal behind coordinate transformation is to represent signals and systems in terms of new basis functions that better suit the application at hand [19, 21, 142]. Audio signal processing is a good example of such applications, since it is known that the human auditory system performs non-uniform spectral analysis on sound waves (see Section 2.1.8). Therefore, audio signal processing systems that interact with the human auditory system should use an internal signal representation similar to that of the human ear. In this section, the fundamentals of unitary warped transformations are introduced. Efficient implementation of warping transformations using non-uniform sampling, and the extension of their applications to real-time filtering are also considered.

Section 5.5.1 introduces the basic principles of unitary operators needed in this section. Duality and unitary equivalence principles, that are introduced in Section 5.5.2 and 5.5.3, respectively, form the tools to deriving the relationships between variables and operators in the time and frequency domains and their warped counterparts. Systems that perform their signal processing in a warped domain are discussed in Section 5.5.4. The multiscale properties achieved by axis warping transformations are demonstrated for the logarithmic warping in Section 5.5.5. An efficient implementation of the coordinate warping transformations using non-uniform sampling is derived in Section 5.5.6. And finally, an example of tailoring the overlap-save algorithm to accommodate the warping technique is discussed in Section 5.5.7.

In addition to the conventions mentioned in Section 1.5, the following conventions will be used throughout this section. Unless otherwise explicitly stated, all signals are considered to be elements of the Hilbert space of square-integrable functions \( L^2(\mathbb{R}) \), which has inner product \( \langle s, h \rangle = \int_{\mathbb{R}} s(t)h^*(t)dt \) for \( s, h \in L^2(\mathbb{R}) \) and a second norm \( \|h\|_2 = \langle h, h \rangle \). Operators are expressed using boldface capital letters, and the notation \( (Us)(x) \) is used to denote processing the signal \( s \) by the operator \( U \) and evaluating the result at \( x \). Throughout this section, the time domain is taken as the default domain of signal representation, therefore, operators are defined in the time domain. The same treatment may be considered for signals and operators defined in the frequency domain with the appropriate interpretation. The symbol \( F \) is used to denote the Fourier operator, which is defined in the continuous-time domain as

\[
S(f) = (F s)(f) = \int_{\mathbb{R}} s(t) e^{-j2\pi ft} dt. \tag{5.11}
\]

#### 5.5.1 Unitary Operators

A unitary operator \( U \) is a linear transformation that maps the Hilbert space onto itself. Unitary operators preserve energy: \( \|Us\|_2^2 = \|s\|_2^2 \) and inner products: \( \langle Us, Uh \rangle = \langle s, h \rangle \). As a consequence, a unitary operator maps a complete set of orthonormal bases in \( L^2(\mathbb{R}) \) into another complete set of orthonormal bases in \( L^2(\mathbb{R}) \). Unitary operators in \( L^2(\mathbb{R}) \) satisfy the relationship \( UU^{-1} = U^{-1}U = I \), where \( I \) is the identity operator. In the following development, three classes of unitary operators are needed. The first is the class of unitary axis warping operators defined by variable substitutions. The second is the class of operators associated with physical quantities such as time, frequency and scale variables. Projecting a signal onto the eigensystems of the latter class of operators defines the third and last class of unitary operators considered here. A familiar example of this projection is the Fourier transformation which is obtained by projecting a signal onto the eigensystems of the time-shift operator. These three classes of unitary operators are examined in more details in this section.

##### 5.5.1.1 Axis Warping Operators

Given a function \( s(t) \) of one independent variable \( t \), performing the variable substitution \( t = \gamma(x) \) results in a new function \( \tilde{s}(x) = s(t) \) expressed in terms of the variable
The effect of axis warping transformations on the signal’s spectrum can be seen from orthonormal bases in the $x$-domain.

The weighting factor does not preserve the signal energy as can be seen from Fig. 5.3 and is, therefore, not a unitary transformation. A unitary version of the original Fourier transformation given by (5.11) and the transformation given by (5.14) can be interpreted as measures of similarity between the signal $s(t)$ and the transformation kernel. While the Fourier bases $e^{j2\pi ft}$ are linearly frequency modulated versions of each other, the bases in (5.14) are amplitude modulated (AM) as well as nonlinearly frequency modulated (FM) versions of each other. Frequency modulation defines the analysis values of the new variable $y$, which is related to the instantaneous frequency at time $t$ ($f_x$) by [45]

$$ (f)_t = y \zeta'(t). $$

On the other hand, both types of modulation contribute to the analysis bandwidth of the kernel. The contribution of each modulation type in the analysis bandwidth for a real mapping function $\zeta(t)$ is given by [45]

$$ B_{AM}^2 = \frac{1}{4\pi^2} \int \left( \frac{d}{dt} \sqrt{|\zeta'(t)|} \right)^2 dt, $$

$$ B_{FM}^2 = \int (y \zeta'(t) - \langle f \rangle)^2 \left| \zeta'(t) \right| dt, $$

where $\langle f \rangle$ is the global average frequency given by

$$ \langle f \rangle = \int \zeta'(t) \left| \zeta'(t) \right| dt. $$

From these relationships, it is clear that both the instantaneous frequency and the analysis bandwidth are dependent on the derivative of the mapping function $\zeta'(t)$. Therefore, the key point is the choice of $\zeta(t)$ to achieve the required spectral resolution. Alternatively, starting in the $x$-domain, and performing the variable substitution $t = \gamma(x)$ results in similar relationships between $\langle y \rangle_x$, $f$, and $\gamma(x)$. The relationships between the original time variable $t$, its dual $f$ and the warped versions $x$ and $y$ are best derived through the duality and unitary equivalence principles.
Table 5.1: Time domain definitions of the time-shift, frequency-shift and dilation families of operators.

<table>
<thead>
<tr>
<th>Operator</th>
<th>Definition in the time domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-shift</td>
<td>((T_\mu s)(t) = s(t - \mu), \quad s \in L^2(\mathbb{R}))</td>
</tr>
<tr>
<td>Frequency-shift</td>
<td>((F_\nu s)(t) = e^{-j2\pi \nu t} s(t), \quad s \in L^2(\mathbb{R}))</td>
</tr>
<tr>
<td>Dilation</td>
<td>((D_\sigma s)(t) = e^{-\sigma/2} s(e^{-\sigma} t), \quad s \in L^2(\mathbb{R}))</td>
</tr>
</tbody>
</table>

introduced in Section 5.5.2 and 5.5.3, respectively. These principles make use of the idea of associating unitary operators to physical variables. Such unitary operators and the expansion onto their eigenfunctions are summarised in Sections 5.5.1.2 and 5.5.1.3, respectively.

5.5.1.2 Operators Associated with Physical Quantities

The idea of associating a variable (such as time or frequency) with an operator is fundamental to the theory of signal representations. Historically, variables have been associated with Hermitian operators [43, 44]. Recently, parameterised unitary operators have also been used [14, 17, 18, 19]. For dual variables, the two representations are equivalent through Stone's theorem [18, 126]. Therefore, only unitary operators have also been used [14, 17, 18, 19]. For dual variables, the two representations associated with Hermitian operators [43, 44]. Recently, parameterised unitary operators have also been used [14, 17, 18, 19]. For dual variables, the two representations are equivalent through Stone's theorem [18, 126]. Therefore, only unitary operators have also been used [14, 17, 18, 19].

5.5.1.3 Eigenfunctions Expansion Operators

The family of unitary operators \(\{X_\xi: \mathcal{H} \rightarrow \mathcal{H}\}\), where \(\mathcal{H}\) is a closed subset of \(L^2(\mathbb{R})\), which corresponds to the variable \(x\) possesses the common set of eigenfunctions \(\{e_{x}(y, \cdot): y \in \mathbb{R}\}\) (which is independent of \(\xi\)), and the corresponding set of eigenvalues \(\{e^{-j2\pi \xi y}\}\). The expansion of a signal \(s(t)\) onto these eigenfunctions defines another unitary transformation \(S_X: \mathcal{H} \rightarrow L^2(\mathbb{R})\) given by [21, 127]

\[
S(y) = (S_X s)(y) = \langle s, e_x(y, t) \rangle = \int_{\mathcal{H}} s(t) e_x^*(y, t) \, dt, \quad (5.20)
\]

which is the signal representation with respect to the eigenvectors of \(\{X_\xi\}\). The inverse transformation \(S_X^{-1}: L^2(\mathbb{R}) \rightarrow \mathcal{H}\) is given by

\[
s(t) = (S_X^{-1} S)(t) = \langle S, e_x^*(x, t) \rangle = \int_{\mathcal{H}} S(y) e_x(y, t) \, dy. \quad (5.21)
\]

The operator \(X_\xi\) is then given by

\[
(X_\xi s)(t) = \int_{\mathcal{H}} e_x(y, t) e^{-j2\pi \xi y} S(y) \, dy. \quad (5.22)
\]

Using (5.20) and (5.21), we can write (5.22) more compactly as

\[
X_\xi = S_X^{-1} A_\xi S_X, \quad (5.23)
\]

where the diagonal operator \(A_\xi\) is defined as

\[
(A_\xi s)(x) = e^{-j2\pi \xi x} s(x), \quad x \in \mathbb{R}. \quad (5.24)
\]

Taking the Fourier transformation of both sides of (5.24) results in the shift property of the Fourier transformation

\[
A_\xi = F \Gamma_\xi F^{-1}, \quad (5.25)
\]

where the diagonal operator \(\Gamma_\xi\) is defined as

\[
(\Gamma_\xi s)(x) = s(x - \xi), \quad x \in \mathbb{R}. \quad (5.26)
\]

Alternatively, taking the inverse Fourier transformation of both sides of (5.24) results in the relationship

\[
A_\xi = F \Gamma_\xi F^{-1}. \quad (5.27)
\]

Therefore, depending on the domain of signal representation, either (5.25) or (5.27) holds. The transformation \(S_X\) given by (5.20) is known to be invariant (up to a phase shift) to the transformation of the signal by the operator \(X_\xi\), for all \(\xi\). That is

\[
(S_X X_\xi s)(y) = \langle A_\xi S_X s, y \rangle = e^{-j2\pi \xi y} (S_X s)(y), \quad (5.28)
\]

which can be readily obtained from (5.23) and (5.24). Therefore, the transformation \(S_X\) is referred to as the \(X\)-invariant transformation [18, 127], since from (5.28) \(\|S_X X_\xi s\| = \|S_X s\|\). The signal representation obtained by \(S_X\) is independent of changes in the signal corresponding to changes in the variable \(x\), in other words, \(S_X\) ignores the transformations of the signals by \(X_\xi\).
The eigenfunctions and the corresponding eigenvalues of the time-shift, frequency-shift and dilation families of operators defined in Table 5.1 are given in Table 5.2.

<table>
<thead>
<tr>
<th>Operator</th>
<th>Eigenfunctions</th>
<th>Eigenvalues</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_\mu )</td>
<td>( { e_\tau (f, t) = e^{j2\pi ft} : (f, t) \in \mathbb{R}^2 } )</td>
<td>( { e^{-j2\pi \rho \mu} : \rho \in \mathbb{R} } )</td>
</tr>
<tr>
<td>( F_\nu )</td>
<td>( { e_\delta (f, t) = \delta(t - \delta) : (f, t) \in \mathbb{R}^2 } )</td>
<td>( { e^{-j2\pi \nu \xi} : \xi \in \mathbb{R} } )</td>
</tr>
<tr>
<td>( D_\sigma )</td>
<td>( { e_\sigma (c, t) = \frac{e^{j2\pi \sigma (k,t)}}{\sqrt{t}} : (c, t) \in \mathbb{R} \times \mathbb{R}_+ } )</td>
<td>( { e^{-j2\pi \sigma c} : c \in \mathbb{R} } )</td>
</tr>
</tbody>
</table>

Table 5.2: The eigenfunctions and the corresponding eigenvalues of the time-shift, frequency-shift and dilation families of operators defined in Table 5.1.

The shift property of the Fourier transformation (5.25) states that a time signal \( s(t) \) can be modulated (frequency shifted) either by multiplying it by \( e^{j2\pi \rho \omega t} \) or by translating its Fourier transform: \( F_\omega = F^{-1} \ T_\omega F \). Similarly, (5.27) results in \( T_\omega = F^{-1} \ F_\omega F \), which expresses time-shift being equivalent to multiplying \( F s \) by the phase factor \( e^{j2\pi \rho \omega} \). This principle of operator equivalence modulo a unitary transformation can be generalised to any unitary operator, which leads to the unitary equivalence principle [14, 18, 20].

The shift property of the Fourier transformation (5.25) states that a time signal \( s(t) \) can be modulated (frequency shifted) either by multiplying it by \( e^{j2\pi \rho \omega t} \) or by translating its Fourier transform: \( F_\omega = F^{-1} \ T_\omega F \). Similarly, (5.27) results in \( T_\omega = F^{-1} \ F_\omega F \), which expresses time-shift being equivalent to multiplying \( F s \) by the phase factor \( e^{j2\pi \rho \omega} \). This principle of operator equivalence modulo a unitary transformation can be generalised to any unitary operator, which leads to the unitary equivalence principle [14, 18, 20].

**Definition 5.1** Two operators \( \tilde{A} \) and \( A \) are unitary equivalent if \( \tilde{A} = U^{-1} AU \), with \( U \) a unitary transformation.
In this section, the above unitary equivalence principle is used to establish the relationships between two variables connected by a unitary axis warping transformation of the type \((5.12)\). For this, consider the unitary operator \(A = A_\alpha\) to be a parameterised unitary operator associated with a physical quantity \(\alpha\) (Section 5.5.1.2), and the unitary operator \(U\) a unitary axis warping operator (Section 5.5.1.1). From (5.23), the warped operator \(A_\alpha\) can be expressed as

\[
\tilde{A}_\alpha = U^{-1}A_\alpha U = U^{-1}S_A^{-1}A_\alpha S_A U = S_{A'}^{-1}A_\alpha S_{A'},
\]

(5.37)

where

\[
S_{A'} = S_A U
\]

(5.38)

is the \(A\)-invariant transformation that is obtained by projecting the signal \(s(t)\) onto the eigenfunctions of the warping operator \(A_\alpha\) (see Section 5.5.1.3). The set of eigenfunctions of \(\tilde{A}_\alpha\), \(\{\epsilon_{\tilde{A}}(y, \cdot)\}\), can be obtained from the definition of the invariant transformation (5.20)

\[
(S_{A'} \psi)(y) = (S_A U \psi)(y) = (U \psi, \epsilon_{A}(y, \cdot)) = \langle \psi, U^{-1} \epsilon_{A}(y, \cdot) \rangle,
\]

(5.39)

from which

\[
\epsilon_{\tilde{A}}(y, \cdot) = U^{-1} \epsilon_{A}(y, \cdot).
\]

(5.40)

Moreover, since \(U\) is a unitary operator, it preserves the inner product in \(L^2(\mathbb{R})\) such that \((A, B) = (A, B)\) for any two operators \(A\) and \(B\). Therefore, the transformation \(U\) can be interpreted as mapping the family of unitary shift operators \(\{A_\alpha\}\) associated with the variable \(\alpha\) onto another family of unitary shift operators \(\{A_\beta\}\) associated with the variable \(\beta\). The transformation \(S_A\) that is \(A\)-invariant is mapped to \(S_{A'}\) that can be shown to be \(A\)-invariant. However, the relative angles between operators remain unchanged [18].

Application of the unitary equivalence principle to the time and frequency shift operators \(T_\mu\) and \(F_\nu\) defined in Table 5.1 results in an alternative definition of the duality principle introduced in Section 5.5.2. Noting that the operators \(T_\mu\) and \(F_\nu\) are equivalent to the operators \(T_\alpha\) and \(F_\beta\), respectively, and using (5.27), the warped time-shift and warped frequency-shift operators \(X_\mu\) and \(Y_\nu\), respectively, can be expressed as

\[
X_\mu = U^{-1}T_\mu U = U^{-1}T_\alpha U = U^{-1}F_\beta U = S_X^{-1}A_\mu S_X
\]

(5.41)

\[
Y_\nu = U^{-1}F_\nu U = U^{-1}A_\mu U = S_Y^{-1}A_\nu S_Y
\]

(5.42)

where

\[
S_X = F U, \text{ and } S_Y = U.
\]

(5.43)

According to (5.33), since \(S_Y = F^{-1} S_X\), then the two warped families of shift operators \(\{X_\mu\}\) and \(\{Y_\nu\}\) are also dual families of unitary operators, the same way \(\{T_\mu\}\) and \(\{F_\nu\}\) are. The results obtained from the duality and unitary equivalence principles when \(X_\mu\) is unitary equivalent to \(T_\mu\) and \(Y_\nu\) is unitary equivalent to \(F_\nu\) are summarised in Fig. 5.4. Note that in this figure, the relationship between the spectral variables \(f\) and \(y\) are given by an operator \(U\) that is defined in the time domain according to (5.42). Therefore, a signal \(S(f)\) is warped to \(\tilde{S}(y)\) through the unitary warping transformation \(F U F^{-1}\).

Similarly, using (5.25) in place of (5.27) results in \(S_X = F^{-1} S_X\), \(S_Y = U\), and \(S_Y = F S_X\). That is, \(X_\mu\) is unitary equivalent to \(F_\nu\) and \(Y_\nu\) is unitary equivalent to \(T_\mu\). The above result is summarised in the following Theorem.

**Theorem 5.2** Two families of unitary operators \(\{X_\mu\}\) and \(\{Y_\nu\}\), both satisfying \((5.19)\) and admitting the spectral representation \((5.23)\) are dual if and only if they are unitary equivalent to time and frequency operators \(\{T_\mu\}\) and \(\{F_\nu\}\), respectively, via \(U = S_Y\) or \(\{F_\nu\}\) and \(\{T_\mu\}\), respectively, via \(U = S_X\).
time systems, such as adaptive 3D sound systems. The warping technique can be used to implement the multiresolution adaptive filters required for improving the robustness of 3D sound systems as mentioned in Section 4.3. This is achieved by pre-processing all input signals of a conventional adaptive 3D sound system by a unitary warping transformation $U$ as shown in Fig. 5.5 [18]. This pre-processing maps the fundamental coordinates of $P$ (the time and the frequency) to the new concepts $x$ and $y$ defined through the mappings $T \mapsto X$, and $F \mapsto Y$, as discussed in Section 5.5.3. This is shown by expressing the input and output signals of $P$ as functions of the warped time variable $x$ instead of the original time variable $t$ (see e.g. Fig. 5.3). The system output $y(x) = (PU)_{s}(x)$ may be interpreted as the response of the warped-LTC (WLTC) system given by $PU$ to the input signal $s(t)$. Since an LTC system is characterised by its covariance to time-shifts, the original system $P$ is covariant to $T$: $(PT_{\mu}s)(t) = (Ps)(t-\mu)$. However, the WLTC system $PU$ is covariant by translation not to $T$ but to $X$: $(PUX_{\mu}s)(x) = (PU)_{s}(x-\mu)$. That is, the unitary axis warping transformation $U$ maps the $T$-covariant system $P$ into the $X$-covariant system $PU$. For instance, the warping function $\gamma(x) = e^{d}x$ in (5.12) maps a linear time covariant system to a linear scale covariant one as will be shortly shown in Section 5.5.5. The power of this approach lies in the following two observations:

- The warped time $x$ is completely defined by the warping operator $U$, therefore, infinitely many internal signal representations inside $P$ may be obtained, since infinitely many transformations $U$ exist.
- This infinite spectrum of signal representations is realized without requiring a computationally expensive processing unit $P$ such as time-varying or time-frequency filters. Well-understood and efficient classical signal processing tools such as the FFT may be used.

These two features make the unitary axis warping transformations attractive for real-time systems, such as adaptive 3D sound systems. The warping technique can be used to implement the multiresolution adaptive filters required for improving the robustness of 3D sound systems as mentioned in Section 4.3. This is achieved by pre-processing all input signals of a conventional adaptive 3D sound system by a unitary warping transformation $U$. The warped signals are processed exactly in the same way as before, while in the warped domain. Since the processed signals must be presented to the users as a function of the original time and frequency coordinates, the inverse axis warping $U^{-1}$ in Fig. 5.5 must be performed. This allows making full use of existing 3D sound systems, while processing the signals in the warped domain maps the time-covariant system to a time-varying one. This enables the system to cope with signals of time-varying spectra, and improves the system ability to predict many of the auditory phenomena that are related to the non-uniform frequency selectivity of the human ears.

5.5.5 Logarithmic Warping and the Mellin Transform

In this section, the mapping function $t = \gamma(d) = e^{d}t$ is considered to illustrate the application and properties of the warping concept discussed in the previous sections. Note that for this mapping function, while $d \in \mathbb{R}$, the variable $t \in \mathbb{R}_+$. Substituting in (5.12), we obtain the unitary warping operator for this warping function,

$$\tilde{s}(d) = (Us)(d) = e^{d/2} s(e^{d}).$$

(5.44)

From the unitary equivalence to the time variable $t$, the warped eigenfunctions can be obtained from (5.40)

$$e_{o}^{*}(c,t) = U^{-1} e_{c}(c,t) = \frac{e^{\frac{j2\pi c \log(t)}{\sqrt{t}}}}{\sqrt{t}}.$$  

(5.45)

The expansion onto these eigenfunctions defines the transformation $S_{D} : L^{2}(\mathbb{R}_{+}) \mapsto L^{2}(\mathbb{R})$ as in (5.20)

$$(S_{D} s)(c) = \langle s, e_{c}(c,t) \rangle = \int_{\mathbb{R}_{+}} s(t) \frac{e^{-j2\pi c \log(t)}}{\sqrt{t}} dt,$$  

(5.46)

which is the Mellin transform of $s(t)$ [9, 28, 29, 97, 146, 147], that is known to be invariant to dilation as will be soon verified. The shift operator associated with the variable $d$ is then obtained from (5.23)

$$(D_{d} s)(t) = (S_{D}^{-1} A_{d} S_{D} s)(t) = e^{-s/2} s(e^{-s}t), \quad s \in L^{2}(\mathbb{R}_{+}),$$  

(5.47)

which exactly the dilation family of operators $\{D_{s} \}$ defined in Table 5.1. Note that $(S_{D} D_{d} s)(c) = e^{-j2\pi c} (S_{D} s)(c)$, which verifies the Mellin transformation to be $D$-invariant. The dual family of operators is obtained by first deriving the eigenfunctions $e_{c}^{*}(d,t)$ from (5.35)

$$e_{c}^{*}(d,t) = \int_{\mathbb{R}} \frac{e^{-j2\pi \log(t)}}{e^{d/2}} d\delta \log(t) - d \rangle \frac{d\delta}{e^{d/2}}.$$  

(5.48)

\footnote{For correct dimensions, a mapping function of the form $t = d_{0} e^{d/d_{0}}$ should be used, where $d_{0}$ is an arbitrary reference value. However, $t = e^{d}$ will be used here for simplicity.}
Using (5.34) and the inverse warping relationship $d = \log(t)$, the transformation $S_{C}: L^2(\mathbb{R}_+) \rightarrow L^2(\mathbb{R})$, which is the expansion of the signal onto the eigenfunctions $e_{c}(d, t)$, is obtained

$$
(S_{C} s)(d) = \int_{\mathbb{R}_+} s(t) \frac{\delta(\log(t) - d)}{\sqrt{t}} dt = e^{d/2}s(e^d), \quad (5.49)
$$

which is exactly the unitary warping transformation $U$ we started off with in (5.44), a result which could be directly obtained from (5.43). Note that $S_{C}$ is covariant to dilation $(S_{C}D_{d} s)(d) = e^{(d-d)/2}s(e^{d-\sigma}) = (S_{C} s)(d - \sigma)$ and, therefore, measures scale contents in the signal $s(t)$. The dual family of unitary operators $\{C_{d}\}$ is then given by

$$
(C_{d}s)(t) = (S_{C}^{-1} A_{d} S_{C} s)(t) = e^{-2\pi i d \log(t)} s(t), \quad t \in \mathbb{R}_+, \tag{5.50}
$$

which corresponds to logarithmic modulation [15, 17, 18, 115].

To demonstrate the multiresolution properties of warped signals, we consider applying the unitary warping transformation (5.44) to a simple signal $s(d)$ composed of three pure tones. Starting in the $d$-domain, the spectrogram (the amplitude squared of the STFT) of this signal in the interval $d = [-0.5, 1.5]$ is shown in Fig. 5.6-[a]. Applying the unitary warping transformation (5.44) on $s(d)$ results in the warped signal, the spectrogram of which is shown in Fig. 5.6-[b]. The logarithmic warping mapped the pure tones onto $d$-varying spectra that have increasing values of $c$ and increasing bandwidth as $d$ increases. A comparison between the Fourier transform of each signal, calculated at a specific $d$ as in Fig. 5.6-[c] and Fig. 5.6-[d], shows that the Fourier transform of the warped signal has increasing analysis bandwidth with increasing $c$.

The key to understanding the frequency resolution properties of the warped signal in Fig. 5.6-[b] is the instantaneous frequency law given by (5.15). For the specific warping in hand, the instantaneous Mellin variable $(c)_{d}$ is related to the frequency variable $f$ by

$$
(c)_{d} = f \gamma'(d) = f e^{d}. \quad (5.51)
$$

This relationship is shown in Fig. 5.7-[a] for the frequency components of the signal $s(d)$ shown in Fig. 5.6-[a] in the interval $d = [-0.5, 1.5]$. A comparison between Fig. 5.7-[a] and Fig. 5.6-[b] in the interval $d = [0.3, 0.7]$ suggests that $(c)_{d}$ is the average of the values of the Mellin variable (the warped frequency) that exist in a signal at a specific $d$, just as the instantaneous frequency $(f)_{t}$ is the average of the frequencies that exist in a signal at a specific time $t$. For this specific warping, the instantaneous Mellin variable $(c)_{d}$ as well as its spread (local bandwidth) increase with increasing time. Fortunately, in most practical applications, especially in real-time...
filtering in a transformation domain, we are interested in processing short segments of signals, which limits this increase. To illustrate, consider that the segment to be analysed is localised at its spread can be achieved. The effects of changing the parameters interval, a given value of the spectrum of the signal more frequencies, which results in wider analysis bandwidths at higher values of the instantaneous Mellin variable coincides with the frequency large values of the normalised c, as shown in Fig. 5.7-[a] (see also Section 5.5.6).

\[ \gamma(nT_x) \]

Figure 5.8: The spectrograms of the logarithmically warped version of the signal shown in Fig. 5.6-[a] for \( d_0 = 0 \) and \( \Delta = 0.2 \) [a], \( d_0 = 1 \) and \( \Delta = 0.2 \) [b], \( d_0 = 0.5 \) and \( \Delta = 0.1 \) [c], and \( d_0 = 0.5 \) and \( \Delta = 0.3 \) [d].

filtering in a transformation domain, we are interested in processing short segments of signals, which limits this increase. To illustrate, consider that the segment to be analysed is localised at \( d = d_0 \) and has a duration of \( 2\Delta \), which correspond to \( t_0 = e^{d_0} \) and an interval \( T = e^{d_0+\Delta} - e^{d_0-\Delta} \) in the \( t \)-domain, as shown in Fig. 5.7-[b]. In this interval, a given value of \( c \) spans a band of frequencies, with higher values of \( c \) span more frequencies, which results in wider analysis bandwidths at higher values of the variable \( c \). By controlling the parameters \( d_0 \) and \( \Delta \), sufficient control on \( (c)_d \) and its spread can be achieved. The effects of changing the parameters \( d_0 \) and \( \Delta \) on the spectrum of the signal \( s(d) \) are shown in Fig. 5.8. For \( d_0 = 0 \) (Fig. 5.8-[a]), the instantaneous Mellin variable coincides with the frequency \( (c)_d = f \) at \( d = 0 \). For large values of \( d_0 \), the warping may lead to aliasing, as in Fig. 5.8-[b]. This is expected since the frequency component \( f_2 \) hits the 0.5 value of the normalised \( c \), as shown in Fig. 5.7-[a] (see also Section 5.5.6).

5.5.6 Axis Warping Using Non-Uniform Sampling

To meet the real-time requirement, an efficient implementation of the unitary warping transformation (5.12) is required to put the warping principle into application. Since DSP systems operate on samples of signals, the starting point is discretising the unitary warping transformation (5.12). Assuming that the sampling process is performed after warping, i.e. in the \( x \)-domain, the discrete-time warped signal \( \tilde{s}(nT_x) \) (see Fig. 5.3) can be expressed as

\[ \tilde{s}(nT_x) = (Us)(nT_x) = |\gamma(nT_x)|^{1/2} s(\gamma(nT_x)), \]  

(5.52)

where \( T_x \) is the sampling period in the \( x \)-domain that is considered to comply with the classical sampling theorem, i.e. \( T_x \leq 1/(2y_{\text{max}}) \), and \( y_{\text{max}} \) is the maximum warped frequency. Implementation of the weighting factor \( |\gamma(nT_x)|^{1/2} \) in (5.52) forms no difficulties, since \( \gamma(x) \) must be chosen a smooth, monotonic, and one-to-one function. In most cases, implementation of this weighting factor can be done off-line, and the result stored in a lookup table to be accessed during real-time processing. The second factor of the right hand side of (5.52) can be interpreted as sampling the signal \( s(t) \) at sampling moments given by \( t_n = \gamma(nT_x) \). That is, the discrete-time warped signal \( \tilde{s}(nT_x) \) is a weighted and non-uniformly sampled version of \( s(t) \). This can be seen as a generalisation of the discrete Mellin transformation [28, 29], which deals only with exponential sampling. However, such non-uniform sampling may lead to aliasing as in the case encountered in Fig. 5.8-[b], and a method for avoiding this aliasing is required.

Since \( s(nT_x) \) is uniformly sampled in the \( x \)-domain without aliasing, then the continuous-time signal \( s(x) \) can be exactly recovered from the uniformly spaced samples, and so can also the original signal \( s(t) \) from its non-uniformly spaced samples. This result is clearly stated in the following Theorem [42]:

**Theorem 5.3** Let \( s(t) \) be a band limited function of one variable that is sampled at sampling moments \( \{t_n\} \) not necessarily equally spaced. If a one-to-one continuous mapping \( x = \zeta(t) \) exists such that \( \zeta(t_n) = nT_x \), and if \( \tilde{s}(x) = s(\zeta^{-1}(x)) \) is band limited to \( \omega_0 = \pi/T_x \), then the function \( s(t) \) can be reconstructed exactly from its samples using the interpolation formula

\[ s(t) = \sum_{n=-\infty}^{\infty} s(t_n) \frac{\sin \left[ \frac{\pi}{T_x} (\zeta(t) - nT_x) \right]}{\left[ \frac{\pi}{T_x} \right] \zeta(t) - nT_x)}, \]  

(5.53)

Therefore, for the non-uniform sampling to be a valid implementation of the warping transformation, two conditions have to be satisfied:

1. The variable substitution \( x = \zeta(t) \) must be chosen such that the mapping function \( \zeta(t) \) is smooth, monotonic and one-to-one and, therefore, invertible with its inverse being \( \gamma(x) \).
Chapter 5: Fast Multiresolution Analysis

Figure 5.9: Hardware implementation of the warping operation and its inverse warping, sc = start conversion.

2. The warped signal is band-limited to $y_{\text{max}}$ and the uniform sampling period $T_x$ must be chosen such that $T_x \leq 1/(2y_{\text{max}})$. The value of $y_{\text{max}}$ for a specific mapping can be calculated from (5.15).

The non-uniform sampling suggested by (5.52) may be implemented in software using any of the well known interpolation algorithms. In this case, the transformation complexity depends on the complexity of the interpolation algorithm in use. A more attractive solution for real-time implementation is using a hardware Analogue-to-Digital Converters (ADC) and Digital-to-Analogue Converters (DAC) to perform the warping and inverse warping (reconstruction) operations, respectively. In this case both the ADC and the DAC must be clocked at the non-uniformly spaced sampling moments $t_n = \gamma(nT_x)$ as shown in Fig. 5.9. When the signal to be warped is already in the discrete-time form, an ADC clocked at $nT_x$ is used to convert the signal to continuous-time and then an ADC clocked at $t_n = \gamma(nT_x)$ is used for warping as in Fig. 5.9.

5.5.7 Real-Time Filtering in Warped Domains

As mentioned in Section 5.5.4, a linear time covariant system may be warped by simply warping its input signals as shown in Fig. 5.5. However, since the mapping function is restricted to be smooth and monotonic, the warped time axis must be stretched (or compressed). This leads to compressing (stretching) the frequency axis, and the warped frequencies decrease (increase) without limit, as mentioned in Section 5.5.5. Therefore, warping long signals becomes impractical, and another arrangement has to be found for real-time filtering in warped domains. In this section, the necessary modifications to the overlap-save algorithm to allow real-time filtering in warped domains are discussed.

Real-time frequency domain filtering algorithms successively process short segments of signals. The FFT algorithm is used to transform a block of samples to the frequency domain where convolution is performed by element-wise multiplication. The multiplication result is then transformed back to the time domain using the inverse FFT. As mentioned in Section 5.5.1.1, calculating the Fourier transformation of a time warped signal amounts to evaluating a new spectral transformation given by

\[ \mathbf{F} U \]

Since the FFT considers its input signal to be periodic of a period equal to the block length, only one period need to be warped prior to calculating the Fourier transformation. The inverse warping is then performed on the block of data after calculating the inverse Fourier transformation. This is illustrated in Fig. 5.10 for the overlap-save algorithm [114, 130]. Similar modifications may be applied to other real-time algorithms such as overlap-add and frequency domain adaptive filters.

In Fig. 5.10, the convolution between an infinitely long input signal $x(k)$ and a finite length filter coefficients $\mathbf{w}$ is calculated in the warped spectral domain. In this Figure, $M$ is the filter length, $L$ is the number of new input samples per block and $B = M + L - 1$ is the block length in the $k$-domain. $M$, $L$ and $B$ are the corresponding variables in the warped domain. The uniformly sampled signal $x(k)$ is stored in an input buffer of length $B$ from which $L$ samples are new, while $M - 1$ samples are taken from the previous block. Every $L$ samples, the input buffer is warped using non-uniform sampling as discussed in Section 5.5.6, producing a non-uniformly sampled sequence of length $B$. This length $B$ vector is transformed to the frequency domain using an FFT algorithm of length $B$ (the $\mathbf{F}_B$ block). The filter coefficients
samples. This vector is finally unbuffered (block clock cycle, giving the linear convolution (or correlation) result. The last of the convolution, the complex conjugate of one of these two vectors is calculated of the two warped frequency vectors.

The usability of many of the well known and frequently used transformations in real-time adaptive 3D sound systems has been investigated. This study shows that those methods are not suitable for the application in hand due to their high computational complexity, unsuitable frequency resolution, or non-invertibility, which motivated the development of the non-uniform sampling method presented in this chapter.

It has been shown in this chapter that by sampling a signal at non-uniformly spaced time moments and calculating the traditional FFT of those samples, a non-uniformly spaced samples of the spectrum of that signal is obtained. The spectral resolution along the frequency axis of the obtained transform is determined by the non-uniform sampling pattern. This can be seen as an efficient implementation of unitary axis warping transformations and a generalisation of the exponential sampling used to calculate the discrete Mellin transformation. In addition to its efficiency, this method requires minor modifications to convert existing constant resolution systems (such as conventional and adaptive 3D sound systems) to multiresolution ones. This has been demonstrated by modifying the overlap-save algorithm to accommodate the warping technique. The non-uniform spectral resolution of non-uniformly sampled signals has been demonstrated for the logarithmic warping function, and the effects of choosing the centre and duration of the analysis window along the time axis have been presented by examples.

5.6 Summary

The realisability of real-time multiresolution (adaptive) filters, that are required for improving the robustness of 3D sound systems, has been addressed in this chapter. Since implementation of an (adaptive) filter in a multiresolution transformation domain forces the filter to have the same multiresolution characteristics, the problem reduces to selecting a transformation that possesses the required spectral resolution. The main forces the filter to have the same multiresolution characteristics, the problem reduces to selecting a transformation that possesses the required spectral resolution.

The objective of the present work is to investigate the application of adaptive filters in creating robust virtual sound images in real-time through loudspeakers. In the previous chapters, the basic principles of adaptive 3D sound systems have been introduced. Efficient implementations, improving robustness, and listeners tracking have been discussed. This chapter summarises the obtained results and highlights the contributions of the present work.

3D sound systems achieve their objectives by filtering monophonic sound signals through a matrix of digital filters. The problem of designing a 3D sound system, therefore, reduces to calculating and implementing those filters. Throughout this dissertation, adaptive filters are used to implement the system filters. In such adaptive 3D sound systems, adaptation of the filters’ coefficients is complicated by the matrix of acoustic transfer functions between the reproduction loudspeakers and the listeners’ ears (the secondary paths). Due to those transfer functions, the traditional Least Mean Square (LMS) adaptive algorithm becomes unstable and can not be directly used. A modified version, the multiple error filtered-x LMS, must be employed. This algorithm is similar to the LMS except that it correlates the error signals with the input signals filtered by the above mentioned matrix of secondary paths. The disadvantages of the filtered-x algorithm is twofold. First, the system complexity increases rapidly as the number of reproduction loudspeakers, listeners, or audio signals increases. Second, the convergence speed is much degraded due to the coloration of the input signals by the secondary paths.

An efficient alternative to the filtered-x LMS algorithm is the adjoint LMS algorithm. This latter achieves stability by filtering the error signals (instead of the input signals) by the secondary paths, which results in a considerable computational saving for multichannel systems. Further computational saving is gained by implementing the filtering and adaptation operations in the frequency domain. This solves the computational complexity problem of the filtered-x algorithm but does not improve the convergence speed.

To improve the convergence speed of the above mentioned adaptive algorithms, the all-pass filtered-x algorithm has been developed. This algorithm limits the coloration and, therefore, speeds the adaptation process by filtering the input signals (or error
signals in the case of the adjoint LMS) through all-pass filters having phase responses equal to those of the secondary paths. This introduces the time delay required for the algorithm stability while keeping the coloration limited without requiring any additional computations.

The matrix of secondary paths not only affects the convergence of the adaptive algorithm, but may also lead to filters that are sensitive to listeners' movements. This is true for adaptive as well as for non-adaptive 3D sound systems. Since the system filters must, partially, implement the inverse of the matrix of secondary paths at all frequencies, the matrix must be non-singular for a solution to exist. If a solution exists, this solution is sensitive to small changes in the matrix if the matrix is ill-conditioned. By examining the condition number of the matrix at each frequency in the frequency range of interest, the optimum positions of the reproduction loudspeakers in a specific listening situation may be investigated. Simulation results in an anechoic environment show that closely spaced stereo loudspeakers result in a matrix that is ill-conditioned at low frequencies and well-conditioned at higher frequencies. As the separation between the two loudspeakers increases, the condition number decreases at low frequencies and spatial aliasing appears at higher frequencies. The decrease in the condition number at low frequencies is transferred equally to the newly introduced peaks due to aliasing. The number of frequencies at which aliasing appears also increases with increasing loudspeaker separation.

Adding a third reproduction loudspeaker to the two-loudspeaker system mentioned above improves the system robustness considerably. However, positioning the third loudspeaker near any of the two original loudspeakers reduces the system to an equivalent two-loudspeaker system.

Similar simulations in reverberant environments show that increasing reverberation increases the system sensitivity to movements. Although similar traces as those obtained in the anechoic simulations could be identified, those traces have higher condition numbers and spread in a wider frequency bands than in the anechoic case. Adding a third loudspeaker to the stereo system in a reverberant environment improves the system robustness considerably. From the above mentioned experiments, it may be concluded that in an anechoic environment, closely spaced loudspeakers result in a more robust system at higher frequencies with ill-conditioned filters at low frequencies. On the other hand, widely spaced loudspeakers result in a system that is robust only in narrow frequency bands between the peaks of the condition number resulting from aliasing. The bandwidths of these frequency bands decrease as the separation between the loudspeakers increases. However, no such conclusions can be drawn in reverberant environments. Better system robustness in both anechoic and reverberant environments may be obtained by increasing the number of reproduction loudspeakers.

Since the coefficients of the adaptive filters are dependent on the listeners' positions, the filters become in error when a listener moves. Traditional 3D sound systems use head trackers to correct the introduced errors. In adaptive 3D sound systems, head trackers may not be needed provided that the directional information to be perceived by the listeners is already included in the audio signals so that the system operates as a cross-talk canceller. It is sufficient to update the filters to their new solutions that correct the errors introduced due to movements provided that the adaptive filters converge sufficiently fast. Measurements of the secondary paths that are required by the adaptive algorithm may also be performed on-line (during system operation) using an adaptive identification process. On-line identification and on-line adaptation of the control filters are only possible if microphones are placed inside the listeners' ears.

System robustness to small listeners' movements may be improved by enlarging the zones of equalization created around the listeners' ears. Two methods are known in the literature that perform this task. The first is the method of derivative constraints, where the spatial derivatives of the sound field in the vicinity of the listeners' ears are constrained to be zero. This method requires analytic expressions of the acoustic transfer functions to calculate the spatial derivatives. Although approximations of such analytical expressions may be obtained in anechoic environments (with the listeners absent) from the geometry of the listening situation, this becomes more difficult as the reverberation in the listening space increases. In the present work, the method of derivative constraints has been extended by developing approximations to the derivatives which are derived from the measured transfer functions. Although simulations prove the validity of those approximations in moderately reverberant rooms, the approximations fail in severe reverberation situations.

The second method is the method of difference constraints, which controls the sound field at discrete points in the vicinity of each ear to be the same as that required at the eardrum. This is achieved by adding extra microphones at those discrete points and adding a control equation for each microphone. This increases the system complexity considerably and requires more reproduction loudspeakers to better control the sound at the newly introduced microphones. To contain this increase in complexity, the spatial filtering method has been developed. This method reduces the number of control equations by controlling linear combinations of the sound field at discrete points in the vicinity of the listeners' ears. This is on the contrary to explicitly controlling the sound field at each individual point as in the difference constraints method. However, spatial filtering still requires knowledge of the acoustic transfer functions at those individual points to calculate the required linear combinations. This requires inserting microphones at those points and measuring the transfer functions, which is not convenient in many applications. This problem is solved by using multiresolution filters.

Observation of the behaviour of the above mentioned constrained filters shows that the constraints force the filters to assume solutions that are increasingly deviating from the optimum solutions with increasing frequency. This observation is in full agreement with the physics of the problem, since the difference between transfer functions measured at adjacent points in space increases with increasing frequency. An average solution that is valid (but not exact) within small movements (or in a
small area in space) may be obtained by ignoring specific spectral details in individual transfer functions. Since specific details increasingly appear with increasing frequency, the average solution may be obtained by employing filters having decreasing frequency resolution with increasing frequency. Such filters automatically ignore more spectral details with increasing frequency due to their coarser frequency resolution. This motivates replacing the traditional constant resolution filters in 3D sound systems by multiresolution ones. Furthermore, the above mentioned physical motivation for using multiresolution filters is also supported by a psychoacoustical motivation. The human auditory system performs a similar non-uniform spectral analysis on sound waves collected by the outer ears. Although ignoring specific details in individual transfer functions has been previously proposed by other authors, to the best of the author’s knowledge, multiresolution filters have not been previously suggested.

Non-adaptive multiresolution filters may be calculated by smoothing the acoustic transfer functions using constant-Q (percentage bandwidth) averaging filters prior to using them in calculating the control filters. For adaptive 3D sound systems, the non-uniform frequency resolution may be obtained by updating the filters in a transformation domain that has the required frequency resolution. To maintain the low computational complexity achieved by implementing the filters in the frequency domain, the multiresolution transformation must be calculated as fast as the Fast Fourier Transformation (FFT). An efficient implementation of the unitary warping transformation using non-uniform sampling has been developed for this purpose. By non-uniformly sampling a segment of a signal and calculating the traditional FFT of those samples, non-uniform samples of the spectrum of that segment are obtained. The spectral resolution along the frequency axis of the obtained transform is dependent on the non-uniform sampling pattern. The main advantage of this method is that the architecture of the system is maintained. All that is needed to replace the constant resolution filters by multiresolution ones is to insert a resampling operation before each FFT in the original system. The resampling operation may be implemented in hardware using analogue-to-digital and digital-to-analogue converters.

Bibliography


Bibliography


Bibliography


[133] H.V. Sorensen and C.S. Burrus, Efficient Computation of the DFT with only a Subset of Input or Output Points, IEEE Trans. on Sig. Proc., March 1993, pp. 1184-1200.


Bibliography

Samenvatting

3D geluids systemen bereiken hun doel door audiosignalen te filteren met een matrix van digitale filters. De oplossing voor die filters bevat meestal twee componenten. De eerste is het “cross-talk cancellation” netwerk welke de matrix van akoestische overdracht functies tussen de reproduktieluidsprekers en de oren van de luisteraars inverteert. De tweede is de matrix van de gewenste overdracht functies welke de richtingsinformatie bevat die de luisteraars zouden moeten waarnemen. Traditionele 3D geluidssystemen gebruiken vaste filters om beide componenten te implementeren. De coëfficiënten van de filters worden van de al gemeten overdracht functies (vaak HRTF genoemd) afgeleid. Vanwege de verschillen in grootte en de fijne geometrische details van het menselijk oor varieert HRTFs veel tussen mensen onderling. Omdat het vrijwel onmogelijk is om de HRTFs van elke luisteraar te meten en op te slaan worden de vaste filters met behulp van algemene HRTFs berekend. De algemene HRTFs zijn vaak gemiddelden van HRTFs van verschillende mensen of de gemeten HRTFs van een kunstmatig hoofd. Tenzij de algemene HRTFs op die van de luisteraar lijken, leidt het berekenen van de filters met behulp van algemene HRTFs tot een slecht waargenomen geluidsbeeld.

De bovengenoemde vaste filters gelden alleen voor één luisterpositie; namelijk de positie waarin de HRTFs zijn gemeten welke in de berekening van de filters zijn gebruikt. De filters gelden niet meer wanneer de luisteraar afwijkt van deze ideale positie. Traditionele 3D geluidssystemen gebruiken “headtrackers” om verschillende filters in te schakelen bij verschillende luister posities. Dat vergt het meten en opslaan van HRTFs van verschillende posities rondom de luisteraar.

Een alternatieve aanpak voor 3D geluidssystemen is de vervanging van de vaste filters door adaptieve filters. In dit geval is het meten en opslaan van HRTFs niet nodig. Bovendien worden de HRTFs van elke individuele luisteraar gebruikt in het ontwerpen (adapteren) van de filters. Door het on-line instellen van de coëfficiënten van de adaptieve filters kunnen fouten, die door bewegende luisteraars worden veroorzaakt, gecorrigeerd worden. Voor zo’n on-line adaptatie van de filters is er een methode nodig om de fouten te bepalen. Dit is mogelijk door kleine microfoons in (of in de nabijheid van) de oren van de luisteraars te plaatsen.

In galm kamers kan de impulsresponsie tussen twee punten een paar honderd milliseconden duren. Met een bemonsteringsfrequentie van 44.1 kHz (de standaard voor audio compact diske) zijn er duizenden FIR filter coëfficiënten nodig om deze impuls responsie te modelleren. Om die reden eisen adaptieve 3D geluidssystemen enorme
Samenvatting

hoeveelheden berekeningen. Dit maakt de real-time software-implementatie van het systeem op beperkte hardware tot een moeilijke taak. Efficiënt implementatie van lange adaptieve filters is dus essentieel. De implementatie van de convolutie en de correlatie in het frequentie domein bespaart veel berekeningen. In meerkanaalsystemen leverd de vervanging van het traditionele “filtered-x” LMS adaptatie algoritme door het efficiëntere “adjoint” LMS algoritme een verdere rekenbesparing op.

Voor het succesvol volgen van bewegende luisteraars met behulp van adaptieve filters moet de convergentiesnelheid van het adaptieve algoritme sneller zijn dan de bewegingen van de luisteraars. Beide boven genoemde algoritmen (de “filtered-x LMS” en de “adjoint LMS”) zijn bekend om hun slechte convergentie eigenschappen vanwege de kleuring die door filtering met de HRTFs is geïntroduceerd. Door die kleuring te verminderen is het mogelijk om de convergentiesnelheid van de adaptieve algoritmen te verbeteren.

De robustheid van het systeem tegen kleine bewegingen kan worden verbeterd door de equalisatiezones in de buurt van de oren van de luisteraars te vergroten. Door het gebruik van meer reproduktieluidsprekers en de optimalisatie van de posities van deze luidsprekers in de luisterruimte is het mogelijk om de robuustheid te verbeteren. Het gedrag van de filters kan ook door het gebruik van “spatial derivative constraints” worden gecontroleerd. Vergelijkbare resultaten kunnen behaald worden met behulp van multi-resolutie filters met een steeds grovere resolutie bij hogere frequenties. Het vervangen van de filters met een constante resolutie door multi-resolutie filters is ook gemotiveerd door het feit dat het menselijke oor een soortgelijke multi-resolutie spectrale analyse verricht op geluidsgolven.

Multi-resolutie adaptieve filters zijn mogelijk wanneer de adaptieve filters in een transformatiedomein (welke over de gewenste frequentie resolutie beschikt) te implementeren. Om de snelheidswinst behaald door het gebruik van frequentie domein adaptieve filters te bewaren moet de multi-resolutie transformatie net zo snel als de Fast Fourier Transformatie (FFT) berekend kunnen worden. Een efficiënte implementatie van de unitaire warping transformatie met behulp van niet-uniform bemonstering is om die reden in dit proefontwerp ontwikkeld. Door het niet-uniform bemonsteren van een gedeelte van een signaal en het berekenen van de gewone FFT van deze monsters wordt een multi-resolutie transformatie van het signaal bereikt.

Biography

John Garas was born on November 6, 1966 in Souhag, a small and quiet city about 500 km to the south of Cairo, Egypt. He attended the primary and preparatory schools in Souhag, where he passed the province final examination with total marks of 99%. Because of this result, he could move to Cairo at age 15 to join a special secondary school that accepts only the first five students from each province. Having successfully finished his secondary school, he joined Ain-Shams University in 1984, where he graduated in June 1989 with honour in the Faculty of Electrical Engineering, Electronics and Communications Department.

After the obligatory military service he joined the Electronics Research Institute in Cairo, where he was involved in research projects in the field of solar energy systems. In January 1993 he received a scholarship to complete his Master degree in The Netherlands. In May 1995 he received the Master degree in Electronics from Eindhoven International Institute, Eindhoven University of Technology, The Netherlands with distinction. He further pursued his studies towards the PhD degree in Eindhoven. In 1996 he received the title of Master of Technological Design from Stan Ackermans Institute, Eindhoven University of Technology, after completing the two-year post-master programme in Information and Communication Technology. This two-year programme has been extended with two-and-half years to carry out the research documented in this dissertation.
STELLINGEN

behorende bij het proefontwerp

Adaptive 3D Sound Systems

door

John Garas
1. Sampling a segment of a signal at nonequally spaced time moments and calculating the discrete Fourier transformation of those samples results in nonequally spaced samples of the spectrum of that segment [this dissertation, Chapter 5].

2. Using adaptive filters in 3D sound systems to track listeners' movements requires inserting microphones inside the listeners' ear canals [this dissertation].

3. Sound quality is not always the main objective when positioning the loudspeakers in the living room.

4. A direct approach to increasing the number of electrical engineering students at the universities is to promise them a future salary comparable to that of their future managers.

5. Simplicity is the top of smartness is a valid proverb, also in high-tech projects.

6. Interculture communication education of all parties is essential for the success of any cooperation between people with different backgrounds.

7. Free software products are not always the most popular, even if they are known for their high quality. Shipments of Microsoft desktop operating systems came in 1998 to about 80 million, compared to half a million copies of Linux [IEEE Spectrum, May 1999].


9. The world knows much more about ancient Egyptians than it does about modern ones.

10. The millennium problem does not seem very scary to the thousands of women who are planning to spend the new millennium eve in the hospital for delivery.