Subjective evaluation and electroacoustic theoretical validation of a new approach to audio upmixing.

John S. Usher

Schulich School of Music, McGill University, Montreal

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Doctor of Philosophy.

September 29, 2006

Your file Votre référence
ISBN: 978-0-494-27853-6
Our file Notre référence
ISBN: 978-0-494-27853-6

Summary

Audio signal processing systems for converting two-channel (stereo) recordings to four or five channels are increasingly relevant. These audio upmixers can be used with conventional stereo sound recordings and reproduced with multichannel home theatre or automotive loudspeaker audio systems to create a more engaging and natural-sounding listening experience. This dissertation discusses existing approaches to audio upmixing for recordings of musical performances and presents specific design criteria for a system to enhance spatial sound quality. A new upmixing system is proposed and evaluated according to these criteria and a theoretical model for its behavior is validated using empirical measurements.

The new system removes short-term correlated components from two electronic audio signals using a pair of adaptive filters, updated according to a frequency domain implementation of the normalized-least-means-square algorithm. The major difference of the new system with all extant audio upmixers is that unsupervised time-alignment of the input signals (typically, by up to ± 10 ms) as a function of frequency (typically, using a 1024-band equalizer) is accomplished due to the non-minimum phase adaptive filter. Two new signals are created from the weighted difference of the inputs, and are then radiated with two loudspeakers behind the listener. According to the consensus in the literature on the effect of interaural correlation on auditory image formation, the self-orthogonalizing properties of the algorithm ensure minimal distortion of the frontal source imagery and natural-sounding, enveloping reverberance (ambiance) imagery.

Performance evaluation of the new upmix system was accomplished in two ways: Firstly, using empirical electroacoustic measurements which validate a theoretical model of the system; and secondly, with formal listening tests which investigated auditory spatial imagery with a graphical mapping tool and a preference experiment. Both electroacoustic and subjective methods investigated system performance with a variety of test stimuli for solo musical performances reproduced using a loudspeaker in an orchestral concert-hall and recorded using different microphone techniques.

The objective and subjective evaluations combined with a comparative study with two commercial systems demonstrate that the proposed system provides a new, computationally practical, high sound quality solution to upmixing.

Abrégé

Les systèmes audio de traitement des signaux pour convertir les enregistrements (stéréophoniques) à deux voies en quatre ou cinq canaux sont de plus en plus en demande. Ces <<démixeurs>> peuvent être employés, avec des enregistrements stéréophoniques conventionnels, pour le cinéma-maison multicanal ou pour les systèmes audio de voitures afin de créer une expérience d'écoute plus naturelle et engageante. Cette thèse discute des approches existantes au <<démixage>> audio pour des enregistrements de musique et présente des critères spécifiques de conception pour qu'un tel système augmente la qualité spatiale du son qui peut être évaluée par des essais d'écoute et des mesures électroacoustiques.

Un nouveau système de <<démixage>> basé sur des filtres adaptatifs est alors proposé et évalué selon ces critères. Le nouveau système retire les composantes corrélées des deux signaux d'entrée en utilisant une paire de filtres adaptatifs basés sur un algorithme NLMS (<<Normalised Least-Mean-Square>>) dans le domaine fréquentiel. Les deux signaux de sortie sont alors projetés par des haut-parleurs disposés derrière l'auditeur. Les propriétés d'orthogonalisation de l'algorithme assurent une déformation minimale de l'image frontale et une image réverbérée naturelle et enveloppante.

L'amélioration de la qualité spatiale du son est définie par rapport à l'image produite lors d'une expérience d'écoute avec une paire stéréophonique conventionnelle de haut-parleurs. Cette amélioration est mesurée par les descriptions qu'un auditeur peut fournir en terme d'image spatiale perçue. En outre, la nature de l'impression spatiale sonore est étudiée en termes d'images qui contribuent à l'impression spatiale d'un instrument musical enregistré et à l'impression spatiale de l'environnement d'enregistrement.

L'évaluation des performances est accomplie de deux façons. Premièrement, par des mesures électroacoustiques empiriques qui valident une description théorique du système avec un nouveau modèle électroacoustique pour des enregistrements à deux microphones d'exécutions musicales en solo dans une salle de concert. Deuxièmement, par des essais formels d'écoute qui étudient l'image spatiale auditive avec un outil graphique interactif pour visualiser les propriétés géométriques de l'image spatiale auditive perçue par les auditeurs. Les deux techniques d'évaluation ont évalué l'efficacité du système de <<démixage>> avec une variété de configurations de microphones pour des exécutions musicales en solo reproduites à l'aide d'un haut-parleur dans une salle de concert. À partir des résultats des évaluations objectives et subjectives et d'une étude comparative avec d'autres systèmes de <<démixage>> existants, il est démontré que le système proposé dans cette thèse fournit une nouvelle solution de <<démixage>> à haute qualité sonore et pratique en terme de charge de calcul.

Acknowledgements

I wish to express my gratitude; first to Professor William Martens, for his help with the final draft of the dissertation and many interesting talks over the last few years which have explained fundamental concepts in method and analysis techniques for audio experiments in a clear and thought-provoking way. Prof. Wieslaw Woszczyk helped with initial development of the graphical mapping system and has been a constant source of support throughout my time at McGill. Prof. Jacob Benesty has been greatly helpful in the development of the theoretical model for the signal processing system; ensuring a disciplined rigour which would have been impossible without his detailed comments and advice.

Through many hours of friendly discussion and questioning, Professor Al Bregman and Dr. Eliot Handelman have both influenced my views and understanding of auditory perception and approaches to audio signal processing which can possibly solve some of the most interesting problems in science today: that of machine hearing. Their advice have together been the most influential factor in my academic history and their teachings will be with me for a long time.

Financial help has been provided by Bang and Olufsen, where the initial motivation for work in spatial audio signal processing was provided by Jan Abildgaard Pedersen (now of Lyngdorf audio). Financial support has also been provided by grants from the National Sciences and Engineering Research Council of Canada, and Valorisation-Recherche Quebec.

Some experimental evaluation was undertaken at the Banff Centre for the Performing Arts, thanks to the kind assistance from Theresa Leonard and Steve Bellamy. Thanks also to the audio engineers and musicians there who took part in the listening tests. Some exploratory testing of the system was undertaken at Philips Research, and many useful suggestions and comments were received thanks to the excellent knowledge of members of the DSP group; thanks especially to Dr. R. Aarts, Dr. D. Schöbben and Dr. W. de Brujn.

Many thanks also to Philippe-Aubert Gauthier for help with the French translation of the thesis summary and many interesting conversations over the past few years.

Table of Contents

List of Abbreviations

ASI	Auditory Spatial Impression
ASUS	Adaptive Sound Upmixing System
ASW	Auditory Source Width
CD	Critical Band
D	only Dry channel active
D+W	Dry and Wet channel active
DA	Descriptive Analysis
DPLII	Dolby Pro Logic II
ER	Early Reflections
HATS	Head And Torso Simulator
IAC	InterAural Coherence
IACC	InterAural Cross-correlation Coefficient
ICC	InterChannel Coherence
ICCC	InterChannel Cross-correlation Coefficient
IR	Impulse Response
LEV	Listener Envelopment
LMS	Least Mean Square
NLMS	Normalized Least Mean Square
ORTF	space angled pair of microphone (Office de Radiodiffusion Télévision Française)
PHIRSF	Perceptually Horizontally Isotropic Reverberant Sound Field
PoO	Principle of Orthogonality
PoS	Principle of Superposition (defines a linear system)
PWAP	PairWise Amplitude Panning
R image	Reverberance Image
RGT	Repertory Grid Technique
RT	Reverberation Time
S image	Source image
SD	Standard Deviation
SRT	Speech Reception Threshold
VBAP	Vector-Based Amplitude Panning
W	only Wet (artificial reverberation) channel active
WFS	Wave Field Synthesis
XY	coincident pair of microphones

Chapter 1

Introduction

The quality of loudspeaker audio has been increasing at a steady rate for over a century. In terms of timbre, there is a strong argument for saying recreation of a recorded sound is as good as it is going to get. However, in spite of recent advances in audio processing hardware "spatial quality... has some way to go before the curve could be said to be asymptotic to some ideal" (Rumsey, 2006). This discrepancy is due to the relatively new arrival of multichannel audio systems in our homes and cars, providing the means to reproduce sound in a way which seems both engaging and aesthetically "natural". And yet the vast majority of our musical recordings are stored with a two-channel "stereo" format which we are forced to listen to with a two-loudspeaker electroacoustic system. This thesis presents a new audio signal processing device which enables reproduction of two-channel recordings with four surrounding loudspeakers; presenting sound in a way which seems both engaging and natural. Such an upmixing system can be classed as a *spatial audio enhancer.*

Spatial audio is a topic which is being discussed by audio engineers at an increasing rate. For example, out of the 185 articles published at either conferences of the Audio Engineering Society (AES) or in the AES journal which include the two-word term *"spatial audio"*; 136 have been published since the turn of the millennium.[1] So what is a spatial audio system? Taking *audio* to mean *the reproduction of sound* and an audio system to be a means for accomplishing this, then if the perception of sound is inherently spatial (Blauert, 1997, pg. 3) then any audio system is *spatial* in the same way that it is *timbral.* Therefore, spatial audio systems do not necessarily mean multichannel (or multi-loudspeaker) systems: Even with music reproduced from a single loudspeaker, spatial properties of the perceived auditory image for the musical instrument can still be consistently described (for example, from the relative loudness of the perceived reverberation in a recording we have an impression of how far away the perceived auditory image is; Begault, 1992).

Idiomatically, however, a spatial audio system means a sound reproduction system which is designed to *enhance* a listening experience in terms of spatial sound quality (Rumsey, 2001, pg. ix). Enhancement of a listening experience may be measured according to a number of criteria; for example, as an enhanced speech intelligibility (Shinn-Cunningham et al., 2001). Specifically in this thesis, *enhancement* means changes in the spatial aspects of auditory images; that is, changes in *auditory spatial imagery* (Martens, 2001; Usher and Woszczyk, 2004). In audio-engineering research, studies under the name of *spatial enhancement* are not restricted to sound reproduction but can, of course, also apply to sound recording techniques (e.g. Woszczyk, 1990; Fukada et al., 1997; Theile, 2000; Berg and Rumsey, 2002) and these will be commented on, but in the context of loudspeaker audio system design

[1] Up until May 2005.

and evaluation for music reproduction.

1.1 Aims of the thesis

This thesis will introduce a new audio signal processing system for converting two-channel recordings to four channels for reproduction with four loudspeakers around the listener. This device- an *audio upmixer*- is intended to be used with conventional sound recordings of musical performances for reproduction with multichannel home theatre or automotive loudspeaker systems (these "surround sound" systems typically have four or five loudspeakers).

The goal of a commercial loudspeaker spatial audio system for music reproduction is generally to increase the *enjoyment* of the listening experience in a way which the listener can describe in terms of spatial aspects of the perceived sound. Likewise, the new system enhances the perceived spatial sound *quality* of the conventional method of reproducing these two audio signals (i.e. with a pair of loudspeakers).

The new upmixer introduced to accomplish this task is specifically developed and its theoretical model validated for two-microphone recordings of single sound sources in a medium-sized concert hall. This is not to say that the new system is restricted to these recording configurations- indeed, performance of the upmixer with conventional "off-the-shelf" commercial recordings of musical ensembles is considered- but simple test signals ensure a solid foundation for understanding the new system, allowing the study to be conducted in a rigorous and controlled manner.

The underlying imperative for the design of the new upmixer is that

the evoked auditory imagery be consistent with that in a conventional two-loudspeaker sound scene created using the same recording. This comes from the general maxim that the mixing intentions of the sound engineer are to be *respected.* As will be shown, this general imperative is translated into meaning that the spatial imagery associated with the recorded musical instrument (the *source* image) remains the same (undistorted) in the upmixed sound scene. The enhancement is therefore in terms of the imagery which contributes to the listeners' sense of the recording space; the *ambiance* or *reverberance* imagery.

The enhancement is qualified in two ways: Firstly, using electroacoustic measurements which relate to parameters known to affect imagery in audio (such as interaural correlation). And secondly, with two subjective listening tests; one concerning a descriptive comparison of the imagery in the upmixed sound scenes and a conventional two-loudspeaker audio scene, using a new computer mapping tool to visualize the perceived location and extent of the auditory images; and another experiment where overall *preference* was investigated.

So to summarize; the new system introduced in this thesis is designed to achieve two general goals, both of which relate to modifying spatial aspects of the auditory imagery experienced with a conventional loudspeaker-pair reproduction of a musical recording:

1. So as to create a natural-sounding auditory image which maintains the sound character of the recording environment and the mixing intentions of the recording engineer.

2. To create a listening experience which people would prefer over the

2/0 listening experience (or at least, a 2/0 audio scene should not be preferred to an upmixed scene).

The second goal is subservient to the first; the new system is not intended as a "special effect" which reinterprets the creative mixing intentions of the recording engineer, but rather as a system to compliment these intentions in ways which are consistent with sound in the natural environment. Nor is this thesis about the design of a *product*; a rigorous understanding of the system at a fundamental level is the aim rather than a broad exploration of the systems' traits.

1.2 Thesis structure

The thesis is presented in six chapters, which are summarized thus:

1. Review of the literature:
   - Introduction to the concept of perceptual auditory imagery in audio as adopted in the thesis, specifically relating to those aspects of imagery which can be described in spatial terms: this particular aspect of imagery is defined as *auditory spatial imagery.*
   - An overview of experimental techniques used in loudspeaker audio research for describing auditory spatial imagery.
   - An overview of extant methods for the enhancement of spatial sound quality for reproduced musical sound recordings using loudspeakers.
2. Exploratory investigation of imagery in audio:
   - Design of a graphical mapping system for visualizing auditory imagery relating to the recorded sound source and auditory imagery relating to

the recording environment (i.e. the *source* and *reverberance* image).

- A report of an exploratory experiment investigating the subjective interaction of source and reverberance imagery in multichannel loudspeaker audio systems.
- Subjective design criteria for the new audio upmixing system, based on results from the exploratory experiment and a review of the literature.

3. Electronic description of the new system:
   - Translation of the subjective design criteria to criteria which can be measured using electroacoustic methods.
   - A model for predicting the system performance with a two-microphone recording of a single sound source in a concert hall.

4. Electroacoustic validation of theoretical system model, using empirical data from recordings of a single sound source in a concert hall.

5. Subjective evaluation of performance; a comparative study of auditory spatial imagery created by the new system and imagery in a conventional loudspeaker-pair audio scene as well as a preference experiment are reported here.

6. Conclusions: A review of the findings from each chapter, limitations of the system and suggestions for further work.

Chapter 2

Overview of Methods for Enhancement of Auditory Spatial Imagery in Reproduced Sound Scenes

rationalize ... to put a thin veneer of reason over thought and actions that are in fact emotional.
Sir Ernest Gowers, in *Fowler's Modern English Usage*, second edition.

"Yeah well, you know, that's just like, your *opinion* man."
The Dude, in *The Big Lebowski.*

2.1 Auditory Spatial Imagery

The term *imagery* has many different meanings in studies in both music and the auditory sciences and in this section it will be defined as it is used throughout the dissertation. It will be discussed why it is an appropriate term to use and how it relates to other terms used to describe perceived sound. The discussion will start in a general sense of various representations of sound and will then progress to a more specific description of perceived sound experienced in loudspeaker audio scenes.

The first quotation at the beginning of this chapter is intended as a reminder that whilst many of the theories of sound perception discussed here can be modeled as a simple causal relationship between a stimulus and listener behavior, listening to *musical* sounds involves mental faculties encompassing realms of consciousness far beyond what is generally called the "auditory system", and that to fully understand how the *nature* of music presentation (i.e. the transduction mechanism) affects a listening experience necessitates an understanding of all these aspects of consciousness (incidently, it is the authors opinion that such a task is beyond the means of humans).

The second quotation puts this rather more colloquially; all opinions about music are just that: opinions. Though some researchers in the perception of reproduced sound believe "opinion can be turned into fact" (Toole, 1982); there is a growing consensus in the auditory sciences that this is precisely what we can not do with opinions.

2.1.1 Perceptual representations of sound

Air-borne sound and perceptual sound are two different things; the first is a collection of vibrating molecules caused by a vibrating body (for example, vibrating air molecules caused by a vibrating guitar string) and the second is a mental phenomenon caused as a result of the vibrating air molecules impinging on the ear-drum.[1] The choice of words to distinguish acoustic and perceptual sound is varied. An interesting on-line discussion resulted in various suggestions such as simply "sound 1" and "sound 2" or "acoustic wave" and "sound" (Warren, 1998). Two other commonly used phrases for the "perceptual world" event is *auditory object* (Griffiths and Warren, 2004) or *auditory event* (Blauert, 1997, pg. 102); the latter defined by Scheirer as "*a short, undifferentiated sound stimulus*" (Scheirer, 2000). However, "object" intuitively seems to be a better word for a higher level representation of a collection of auditory events which has a direct correlate with a physical body, such as an entire piano; the word "event" seems like a transitory occurrence of perceptual sound, such as a single note created by a piano.

To reflect the active process which brings about the perception of the sound, Bregman uses the word "stream": "*the perceptual unit that represents a single happening ... The stream serves the purpose of clustering related qualities*" (Bregman, 1990, pg. 10). The word *object* is often used synonymously with *stream* (Kubovy and Valkenburg, 2001) but the word "stream" seems better used when the ontological relationship between acoustic and perceptual event are more complex than a simple one-to-one mapping. A good example of this is musical voices; for instance, in a fugue for solo pi-

[1] As Yost (1991) mentions- an auditory image may be formed in the absence of a physical source; an auditory illusion (or "inner singing" Marin and Perry, 1999). Such perceptual sound is not considered here.

ano we may hear three or four musical "voices" (i.e. *streams*) but only a single piano *object*. Put another way, an auditory stream generally refers to an abstract representation of a collection of auditory events (i.e. with related qualities) and one which is not usually described using words relating to physical properties such as direction or size. For representing an auditory object in terms of a limited number of aspects, such as timbre and space, the word (auditory) *image* is often used (e.g. McAdams, 1983; Letowsky, 1989; Yost, 1991; Scheirer, 2000; Griffiths and Warren, 2004). In this sense, it is tautological to say "*perceived* image(ry)" (though it helps to remind us that we are talking about a *perceptual* rather than a *physical* phenomenon).

The word *image* is particularly useful for describing the perception of music reproduced with loudspeakers; not least because it is the accepted idiom in this field of research (e.g. Chernyak and Dubrovsky, 1968; Toole, 1983; Letowsky, 1989). The etymology of the word reflects the complicated nature of the perception of reproduced music, which is inherently abstract because the local sound source is a loudspeaker yet the percept is, for example, of a piano. The origin of the word *image* comes from three different Greek words; *eikon*, *eidolon* and *phantasma*.[2] *Eikon* means resemblance, re-expressing reality or the things of the mind or dreams; *eidolon* means looking at an appearance or form; and the word *phantasma* means vision, dream and phantom. Likewise, in research concerning loudspeaker audio a distinction is made between a *phantom*[3] image and a *real* auditory image (Pulkki and Hirvonen, 2005; Usher and Woszczyk, 2005): A real image is perceived to exist at the same location as the sound-creating transducer whereas a phantom image is perceived to exist somewhere else (when the sound is reproduced with two or more loudspeakers, this is generally- though not always (Queen,

[2]Oxford English Dictionary, second edition, 1989.

[3]The word "virtual" is used interchangeably with "phantom".

1979)- between them).

The use of the word *image* for a perceptual sound event which is not as verifiable as, say, a sound producing object which the listening can see as well as hear is also shared by Yost (1991), who calls an auditory image any sound event which is a *candidate* for a real-world sound creating body: "*each auditory image indicates a possible sound source*". This implies that an auditory image is a lower level representation of an auditory object; a representation which is less inclusive. Letowsky (1989) defines an auditory image as a representation of an auditory object in terms of timbre and space. Griffiths and Warren (2004), on the other hand, use only the timbral dimensions of frequency and time.[4] The distinction between these two definitions reflects the theory that space in audition is not an *indispensable attribute*. "*An attribute (or dimension) is defined as indispensable if and only if it is a prerequisite of perceptual numerosity*"(Kubovy and Valkenburg, 2001). A simple example supporting this theory is familiar: that we can segregate (and count the number of) different musical instruments in, say, a piano concerto even when the recording is reproduced with a single loudspeaker.

2.1.2 Global descriptions of auditory imagery

The combined effects of the timbral and spatial aspects of auditory images is often called *basic audio quality* (ITU-R BS 1116, 1994). Rumsey et al. (2005a) conducted a series of listening tests with 21 expert listeners to look at how

[4]According to The American Standards Association 1960 definition, timbre is defined as "...that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar." Strictly speaking, therefore, timbre is affected by not just *spectral* changes of the stimulus, but also by *spatial* changes. However, as with many previous works (e.g. Letowsky, 1989; Bech, 1998) we shall consider timbral and spatial aspects of reproduced sound to be different things.

a single basic audio quality rating of a sound scene relates to separate timbral and spatial descriptions of the same sound scene (the spatial description was split into a frontal and surround image rating). The timbral and spatial descriptions were assessed in terms of *fidelity*; a measure of the perceived similarity between the test scene and a reference scene. The sound scenes were created using five loudspeakers spread around the listener as shown in figure 2.10 (i.e. the conventional 3/2 ITU-R BS 775-1 arrangement) and used a variety of musical and non-musical stimuli which had been processed so as to reduce the audio quality, e.g. with low-pass filtering. The timbral fidelity ratings were highly correlated with the basic audio quality ratings (Pearson correlation $r = 0.93$) and these timbral ratings were less (but significantly) correlated with the spatial fidelity ratings ($r = 0.33$ for the front images and $r = 0.19$ for the rear images). This supports the two-component model of Letowsky (1989) for the representation of an auditory image with approximately orthogonal perceptual axes of spatial and timbral dimensions (if the spatial and timbral ratings were highly correlated in the study by Rumsey et al. (2005a), it could be interpreted that the terms meant a similar thing to the listener). The spatial fidelity ratings were also significantly correlated with basic audio quality ($r = 0.63$ for the front and and $r = 0.43$ for rear imagery). Rumsey et al. (2005a) suggests that the reason surround (rear image) information did not seem to affect the ratings of basic audio quality as much as frontal image ratings could be due to the relative unfamiliarity of the subjects with surround-sound audio scenes compared with conventional two-loudspeaker scenes (i.e. 2/0 scenes).

Rumsey (1999) found in a previous study that upmixed surround-sound scenes using five loudspeakers were rarely preferred over two-loudspeaker (2/0) scenes (though the results with the new upmixer system in this thesis in chapter 6 show that experienced listeners generally preferred the upmixed

scene to the original 2/0 scene). Another explanation is that the subjects simply considered any changes to the spatial fidelity less important than changes in timbral fidelity. The subjective evaluation of the new upmixer in this thesis is primarily in terms of spatial imagery because the physical modification provided by the system is a spatial one; that is, radiation using four surrounding loudspeakers rather than just two front ones.

Just the spatial aspects of auditory imagery are generally associated with the holistic feature called (Auditory) Spatial Impression (ASI; e.g. Barron, 1971; Blauert, 1997; Morimoto, 2001); "the three-dimensional nature of sound sources and their environments" (Rumsey, 2006). ASI is often associated with *spaciousness*, though this is also used as another word for describing the reverberance imagery or ambience. Letowsky (1989) defines spaciousness as "*...that attribute of an auditory image in terms of which the listener judges the distribution of sound sources and the size of acoustical space*". This definition seems very similar to the idea of ASI; a holistic judgement relating to the properties of both the sound-creating object and the room the object is in.

2.1.3 Source and reverberance imagery

When specifically considering those parts of an auditory image which can be described in terms relating to geometric (coordinate) space, the phrase *auditory spatial imagery* is used (Kendall, 1995; Martens, 2001).[5] The definition used in this thesis is:

[5]Auditory spatial imagery is not abbreviated to ASI as ASI is more commonly used to refer to Auditory Spatial Impression (e.g. Morimoto, 2001). The two terms are obviously closely related, but *impression* is a more general term than *imagery*. For instance, a perceptual *impression* of a sound may relate to a general sense of size- large or small- but not to specific measures such as an absolute judgement of an auditory image size.

Auditory spatial imagery means those parts of a perceptual sound image which can be described in terms of physical space.

The word *stereophony* is also used to describe the spatial aspects of an auditory image (Snow, 1953; Steinke, 1996). The word stereo comes from the Greek root *stereos*, meaning *solid*, so in other words a stereo auditory image is one which can be talked about using all the words we use to describe the spatial properties of solid objects in the real world. The *phonic* part reminds us we are talking about perceptual images which have been created as a result of hearing. However, in this thesis the word *stereophony* will be avoided due to the unfortunate idiomatic association of stereo and two-channel audio (for example, to many people a *stereo* audio signal means two electronic or acoustic signals and a *stereo* image means the perceptual image evoked by reproducing these two signals with a pair of transducers). An example of an audio system with weak stereo properties is reproduction with a single loudspeaker (which is why such systems are generally not called spatial audio systems).

As mentioned by Rumsey, the subjective evaluation of loudspeaker audio systems has tended to concentrate on describing the auditory imagery evoked by the recorded sound source(s) (Rumsey, 1999). The auditory image corresponding to the original sound-creating object in the recording is called the *source image*. A source image may be spatially discontinuous, so a single auditory *object* may be described with more than one auditory *image*. For example, the perceived regions of space which contain the source object may be separated by regions of space which do not. An example of this in nature is the sound of a large but partly occluded sound-creating thing, such as the sea heard from behind a sand-dune. For loudspeaker audio this is even more likely to occur due to the spatially separated transducers used to create the

sound.

When listeners are not told to specifically describe the source image, commonly used adjectives are related to the reflected sound content in the recording (words such as *envelopment*, *depth* and *presence*; e.g. Berg and Rumsey, 1999; Zacharov and Koivuniemi, 2001a; Guastavino and Katz, 2004). In physical terms, reflected sound can be thought of as consisting of two parts: early reflections (ER's) and reverberation. ER's are defined as "*those reflections which arrive at the ear via a predictable, non-stochastic directional path, generally within 80 ms of the direct sound*" (Beranek, 1996) whereas reverberation is generally considered to be sound reflections impinging on a point (e.g. a microphone or ear-drum) which can be modeled as a stochastic ergodic function, like random noise (Schroeder, 1987; Jot and Chaigne, 1997). This distinction will be expressed mathematically in chapter 4, as it has important consequences for the new audio system introduced in this thesis which tries to remove those sound components from a recording which contain information about the spatial properties of the recorded sound source. Put another way, the new system tries to remove all those components from a recording of a live musical performance which enable a listener of the recording to describe the spatial properties of the source image. The corresponding psychological term to describe the perception of reverberation is *reverberance* (Meyer and Schodder, 1952; Marshal and Barron, 2001; Morimoto and Asaoka, 2004). Especially by the French, this is also called *ambience* (e.g. Tak, 1958; Letowsky, 1989; Steinke, 1996; Avendano and Jot, 2004), though the word reverberance is preferred as ambience has other meanings in English associated with emotive feelings about the listening experience such as *mood*.

When listening to live or recorded musical instrument performances, according to Griesinger (1996b): "*The brain processes incoming sound into a*

foreground stream - the part which holds the information content of the signal - and a background stream... In a reverberant environment the background is the reverberation". This two-component perceptual representation of imagery in audition is related to the figure-ground analogy of what an auditory object is (Moore, 1997, pg. 295;Kubovy and Valkenburg, 2001), where the "figure" in this case is the source (or foreground) stream and the "ground" is the perceived reverberation (or background) stream. In the ideal analytical listening case: "*Attention selects one putative object (or a small set of them) to become figure ... and relegates all other information to ground*" (Kubovy and Valkenburg, 2001). However, it is not attention alone that controls which stream the listener chooses to attend to; Griesinger (1997) says that the reverberance stream is completely inhibited (i.e. can't be attended to) for 50 ms after the end of a sound segment; a kind of *forward-masking*,[6] whereby the early-arriving source-related information masks the reverberance. A simple example of this is to record speech or a live musical performance in an ordinary room and to replay the passage backwards (Houtsma et al., 1987): The reverberance seems louder than when the recording is played forwards as now the reverberation *precedes* the direct sound and early reflections and the reverberance is unmasked. This principle explains why the rear loudspeaker channels of many spatial audio systems are delayed and also why such simple spatial audio systems as the Madsen system works; some examples are described in section 2.2.

[6] *Masking* means the amount (or process) by which the threshold of audibility for a target sound is raised by the presence of another masking sound. Forward masking occurs when the masking sound precedes the target sound (Moore, 1997).

2.1.4 Factors affecting source and reverberance imagery

Morimoto (2002) investigated temporal factors of the reflected sound which affected the perceptual distinction between source and reverberance images. He showed that the factors were related to the precedence effect: early, high-level reflections are fused with the direct sound to create the source stream contributing to Auditory Source Width (ASW) and later low-level reflections (i.e. reverberation) are involved in the formation of the reverberance stream (which is often called listener envelopment, LEV, or spatial impression).

Using loudspeaker reproduction of music with delays from various directions in an anechoic chamber, three factors were identified which affect LEV (Bradley and Soulodre, 1995; Soulodre et al., 2003):

- Side (i.e. lateral) reflections contributed to a greater sense of LEV.
- LEV and reproduction level were positively correlated.
- The ratio of reflected energy arriving after a time boundary to the total sound energy was correlated with perceived LEV; this time boundary was frequency dependant; it was larger for low frequencies, e.g. 160 ms at 125 Hz, and shorter for high frequencies, e.g. 45 ms at 4 kHz.

However, when "listening to reverberation", due to the strong cognitive "perceptual completion" mechanisms which the auditory system employs it is often difficult to tell whether the acoustic energy itself (i.e. of live or reproduced reverberation) was perceived or the brain just "imagined" the sound in the absence of the acoustic cues themselves. This quandry is almost philosophical in terms of its complexity, but the strength of the cognitive

phenomenon can be demonstrated well with the *continuity illusion* (Bregman, 1990, pg. 344): part of an acoustic signal, such as a spoken sentence, is deleted and replaced with a louder sound (e.g. a noise burst) yet the sentence is perceived to continue uninterrupted 'behind' the sound (this particular example is called *phonemic restoration*).

Morimoto and Asaoka (2004) conducted a dissimilarity judgement experiment to rate simple sound stimuli in terms of the adjective "reverberance". A solo violin was reproduced from a centre (i.e. 0°) loudspeaker in an anechoic chamber and two uncorrelated reverberation channels were reproduced from the left and right with loudspeakers at various angles from the centre-speaker (±7°, ±35°, and ±80°). The reverberation time (T60) was also varied (1.0, 1.4, and 2.0 seconds). A multidimensional scaling analysis of the dissimilarity judgement ratings for the different scene configurations revealed that these two variables were represented approximately orthogonally on a two-dimensional space. In other words, the 9 stimuli presented (the permutations of the 3 spatial variables and 3 temporal ones) could be arranged geometrically in terms of perceived similarity using markers on a 2D map (where each marker on the map represents one of the nine stimuli, and the distance between markers represents the perceptual difference in terms of reverberance). As only two physical variables affected the stimuli, then it can be concluded that one of the perceptual axis is related to the temporal manipulation of the stimuli (i.e. RT) and the other related to the spatial variable.

The spatial distribution of reproduced reverberation needed for the perception of a diffuse sound field was investigated by Hiyama et al. (2002), who found that the local acoustic wavefield around a listener does not have to satisfy the acoustic definition of reverberation. In their study, it was found that regardless of frequency the perceived spatial homogeneity of reverberance

using 24 loudspeakers equally spaced around the listener (in the horizontal plane) could be achieved using only 12, and that with only 4 loudspeakers (conveniently arranged according to the 2/2 ITU-R BS 775-1 format, with rear loudspeakers at $\pm 120°$) it is possible to create a soundfield which is perceived to be nearly identical to the 24-speaker arrangement. Similar findings were arrived at by Sonke (2000) in an investigation with a wave field synthesis system (this system is explained later in this chapter). The aim of his experiment was to see how many plane waves needed to be reproduced for a perceptually, horizontally isotropic reverberant sound field (PHIRSF) to be achieved. This was done in an elegant way: "*the property of isotropy corresponds to the property of rotation invariance... if a sound field can be rotated over an arbitrary angle without changing its properties, it is isotropic*" (Sonke, 2000, pg. 102). It was found that the number of plane waves needed for a PHIRSF was about four (median) and the most discriminating listener needed only eleven plane waves to be reproduced for a PHIRSF. So it seems that a reasonably natural-sounding reverberance image aesthetic can be achieved using only four loudspeakers; a hypothesis which is investigated later by looking at how uniformly spread the distribution of reverberance images are around a listener.

So in conclusion; the formation of a reverberance stream *is influenced* by the perceived spatial distribution of reverberation but it is the *temporal* nature of this stream which distinguishes it from the source stream. Some researchers have posited that reverberance imagery and source imagery (or ASW and LEV) are really the same thing; it is just a matter of extent. For example, Morimoto (2001) says: "a small degree of spatial impression could be termed as ASW and a large one as LEV, leaving the border between them fuzzy." The term "LEV" is deliberately avoided in this dissertation because its name is ambiguous, as both source and reverberance images can sometimes

be considered *enveloping*.

2.1.5 Image definition

When recorded music is reproduced with two loudspeakers the listener will often have a sense of location for the recorded instrument (i.e. the source image). To describe how strong this sense of location is, Blauert (1997) uses the term *locatedness*. That part of a source image which has the strongest sense of location is often called the image *focus* (e.g. Martin et al., 1999; Ford, 2005). If the sense of direction is very weak, the locatedness is said to be *diffuse*, if it is strong, then the locatedness is *sharp*, *concise*, *clear* or *defined* (e.g. Toole, 1983; Rasch and Plomp, 1982; ITU-R BS 1116, 1994; Usher and Woszczyk, 2004). Lee and Rumsey (2005) showed that when describing an audio image in terms of width and locatedness, the two terms were generally highly correlated but with significant exceptions when analysed for different stimuli. Therefore, an auditory image can have a wide image extent *and* a high degree of locatedness.

Gabrielsson et al. (1974) and Letowsky (1989) also use a term called *distinctness* to describe the perceived spatial separation between auditory images. Lund (2000) discusses a similar metric called the "consistency score" which is composed of three subjective attributes called "certainty", "robustness" and "diffusion". Though *robustness* is not defined, *certainty* relates to the confidence a listener has in reporting an image direction and *diffusion* relates to how wide the image is. Other studies have also asked listeners to rate their confidence in reporting image direction (Segar and Rumsey, 2001; Corey and Woszczyk, 2002).

The term *definition* is used as a more general term which encompasses all these image descriptions; defined by Toole (1983) as "*the extent that different sources of sound are spatially separated and positionally defined*". In the sane vein, Rasch and Plomp (1982) define *definition* as "*The ability to distinguish and to recognize sounds*". In an exploratory experiment (Usher and Woszczyk, 2003) it was wondered whether listeners could report the definition of an auditory image directly. A five-point system was used to describe the definition of a image, with the lowest category corresponding to a weak definition. In this experiment, the definition categorization system was explained to the subjects with the analogy of temperature- with the highest image category corresponding to a "hot-spot" which the subject should use to indicate a region of high definition (the image was described using a graphical mapping system similar to that introduced in chapter 3). However, subjects generally reported not understanding how to differentiate between the categories and tended to use just three of the five image categories (i.e. the upper, lower and middle categories). This reflects the nature of image definition being a complicated multidimensional construct. To simplify the description of an image definition, it was decided to represent one of the attributes of definition which is more intuitive and reasonably unaffected by emotive feelings of the listener towards their listening experience: image temporal stability. With the graphical mapping system which was developed the source image could be described as being either *stable* or *unstable*[7].

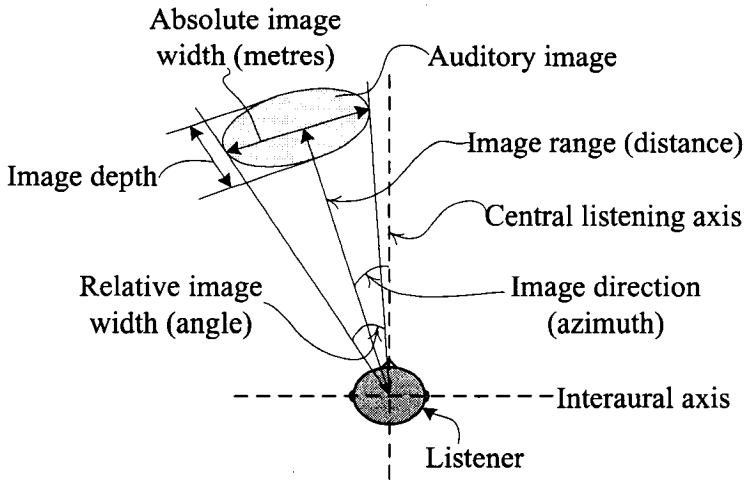


Figure 2.1: Top-down view of geometric auditory image attributes showing terminology used in this thesis.

2.1.6 Methods for describing auditory spatial imagery

Ideally, we could measure the acoustic wavefield created by a new spatial audio system, or just the electronic output signals, and predict how a listener would perceive the spatial imagery in the sound scene if they were exposed to it. This would save long and complicated experiments where the listeners must firstly be trained how to grade the magnitude of their auditory sensations (e.g. in terms of how far away or loud the image seems) and how to report this. Secondly, in most listening tests the experimenter must relate a sensory judgment to the measured stimulus intensity (e.g. relate the perceived image width to the sound pressure level or electronic correlation between the loudspeaker signals). This ideal is at the heart of psychophysics: *"the relation between the physical world stimulus objects and the psychological domain of conscious awareness... motivated by a desire to predict the out-*

[7]Other studies have noted this distinction and have called the unstable images *fuzzy* images (e.g. Corey, 2002, pg. 60-64).

come of experiments..." (Baird, 2001). Psychoacoustics is the psychophysics of hearing; *"the quantitative relation between acoustical stimuli and hearing sensations"* (Zwicker and Fastl, 1999, pg. VII). Such classical psychophysical motivations for experiments in audition are contrasted with what Rumsey (2002) calls "product evaluation". The latter motivation applies more to the listening tests undertaken in this thesis for the design and development of a new audio system, and evaluation of it in terms of developed criteria and for validation of the new electroacoustic theory.

To describe the auditory spatial imagery of a sound creating object is a very familiar and everyday task (e.g. how far away a bus is), but the problem with describing the sound character of a *reverberance* image in spatial terms is that we are dealing with an inherently abstract listening experience: listening to (reproduced) reverberation and describing this sensation in terms of spatial imagery. This is not an entirely foreign task; people, especially the blind, are used to "listening to reverberation" to estimate the size and contents of a room (McGrath et al., 1999). However, asking a listener to report the extent and location of a reverberance image created with loudspeakers is not the same thing. Along similar lines, this point was made by Helmholtz (quoted by Moore, 1997, pg. 133):

> "... we are exceedingly well trained in finding out by our sensations the objective nature of the objects around us, but that we are completely unskilled in observing these sensations *per se*; and that the practice of associating them with things outside of us actually prevents us from being distinctly conscious of the pure sensations."

The way in which a listener describes their listening experience can be

considered to be either a direct or indirect judgement. A direct judgement is generally considered to be when the subject is asked (by the experimenter) to scale the magnitude of a sensation, such as to give an absolute value as to how far away a sound source is perceived to be (e.g. in metres). Mershon (1997) criticises such a paradigm for allowing cognitive biasing to influence a listeners judgement of auditory spatial imagery. An example of this is that in the real-world, when a familiar sound creating object moves away from us we expect to hear certain changes from our previous experiences with this object, such as a persons voice getting quieter. Therefore, when we hear a whispered voice reproduced with a loudspeaker, "high-level" cognitive centres of the auditory system may influence our judgement to underestimate the distance to the loudspeaker because the whispering voice is *expected* to be close (we generally can't hear someone if they are whispering far away from us). Mershon (1997) reported a study where recorded speech was reproduced with loudspeakers and the listeners' task was to report the image distance. Recordings of shouted speech were consistently reported farther than speech at a whispered or "conversational" level, even though the shouted speech was presented at a level about 6 dB greater . This cognitive biasing makes describing the spatial properties of an auditory image an inherently problematic task, especially for reproduced music where the listener has a strong association with the sound of an instrument and its size, such as a piano or piccolo. Nevertheless, various methods which require the subjects to directly scale and report the magnitude of an auditory sensation will be discussed, as these techniques can provide a large amount of data in a very useful way which can be subjected to established statistical analyses.

Direct and indirect psychophysical methods for investigating auditory spatial imagery

Direct methods.

Direct methods apply to tasks where the listener has to directly report the magnitude of a sensation on a scale. These techniques are generally used when the stimuli used in the experiment encompass a large range of variation in terms of the perceptual attribute(s) under investigation, i.e. direct methods are generally used for *global* rather than *local* comparisons.

o Magnitude estimation.

In this experimental paradigm, the subject is asked to rate the strength of the sensation in terms of a particular attribute. A simple example of magnitude estimation is without any reference and the listener is asked to simply report the magnitude in "real-world" units, such as an estimate of the sound source distance in metres (Brungart, 1993). This task may also be performed with reference to a fixed stimulus (an "anchor sound"; Zwicker and Fastl, 1999, pg. 9) to which the experimenter assigns an arbitrary magnitude (which the subject is told). For example, in a distance estimation task by Zahorik (1997), listeners were provided with three reference sounds which they were told corresponded to distance magnitudes of 1, 10 and 100 (the stimuli were actually recordings of sounds at distances of 0.3, 2.5 and 19 metres). Their task was to report the distance of another stimulus in proportion to the references; a method called *ratio scaling*.

o Method of adjustment (MoA).

The MoA (Zwicker and Fastl, 1999, pg. 9) is different from magnitude estimation in that the listener has control over manipulating the stimuli. For example, the listener may be presented with two stimuli, one of which is a

fixed reference, and the listener must then adjust the variable stimulus until it matches the reference (e.g. in terms of image width; Mason et al., 2004; or moving a sound marker- also called a "template sound" by Evans (1998)- such as a loudspeaker, until it is perceived to have the same direction as the reference stimulus; Ratliff, 1974; Theile and Plenge, 1977). Another variation in the MoA is *magnitude production*, in which the subject changes the variable stimulus until it is, say, half the magnitude of a particular attribute relative to the fixed stimulus (a method used by S. S. Stevens in his classic work on loudness perception; Moore, 1997, pg. 131). Ratio scaling can only be applied to attributes which have scales with an origin represented by zero (Gescheider, 1997, pg. 188); such as ego-centric distance or image width, but not azimuth where negative azimuths often mean the direction is to the left of the central listening axis. It could be argued that when the task is simply to match two stimuli (in terms of a particular subjective parameter) the task is an indirect method.

o Rank-order experiments.

Here, a listener is presented with different stimuli and asked to rate each using a scale in terms of a particular attribute. For instance, Neher et al. (2002) conducted a study where listeners were asked to rate each of five different sound scenes on a scale in terms of source width.

o Global dissimilarity judgments.

This class of magnitude estimation is similar to Thurstonian scaling (described later) except that the listener gives an estimation of the perceived dissimilarity between the two stimuli presented in terms of a particular attribute, such as image width (Martens, 2001), or image distance (Zahorik, 1997). The experimenter can then use multi-dimensional scaling techniques to reveal the underlying perceptual structure of the listeners' perception of

the stimuli.

Indirect methods.

Indirect methods for describing auditory imagery can help to remove cognitive biasing effects. This is especially useful when the difference in the sensorial strength between stimuli (in terms of the attribute under investigation) is expected to be small; i.e. for *local* comparisons.

o Measure of difference threshold.

Introduced by Fechner in the late nineteenth century, the premise for this method for determining sensory magnitude uses the Just Noticeable Difference (JND) as a criteria for calculating perceptual units of equal magnitude (Blauert, 1997). Here, the subject is not asked to scale the strength of the sensation, but to just report whether a difference in the sensation is heard (a kind of binary scaling). From the data, an *interval scale* can be constructed for the sensation magnitude as a function of stimulus intensity (Gescheider, 1997).

o Thurstonian scaling.

This method presents two stimuli, and the subject has to report which of the stimuli is greater with respect to a particular attribute (Gescheider, 1997, pg. 185); the underlying assumption being that stimuli which are frequently confused with one another are assumed to be psychologically similar. The stimuli can be ordered on an interval scale by the proportion of times a stimulus is judged to be greater (with respect to a particular attribute) than another stimulus (using Thurstone's case V model; Torgerson, 1958;Gescheider, 1997, pg. 203). For example, Zahorik (1997) investigated perceived auditory image distance using this method.

- Method of constant response.

This approach can be used to estimate the point of subjective equality of two stimuli in terms of a particular attribute. For example, Martens (2004) conducted a study with this method and presented listeners with three stimuli; two of which were the same (the so-called two-alternative-forced-choice paradigm) and the subject was asked "Which of the stimuli seems closer?" The stimuli were changed until the subjects were shown to be selecting the answer by chance; when this occurred it could be inferred that the image distance (range) of the two stimuli were psychologically identical.

Methods for investigating auditory spatial imagery in loudspeaker audio

Auditory images can be described in two ways; in terms of either the sound *quality* or sound *character*. Letowsky (1989) defines a *quality* description as one where the listener expresses a degree of satisfaction or dissatisfaction; a rating of how well *liked* or *pleasant* the auditory experience is (i.e. a hedonic description), whereas a description of the sound *character* attribute is purely descriptive (Rumsey, 2002). The distinction between quality and character is similar to the difference between *sentiments* and *judgements*. Nunally and Bernstein define a judgement as an opinion (i.e. unproven belief) which can be subject to veridical evaluation; but "veridicality does not apply to sentiments" (Nunally and Bernstein, 1994). An example of a sentiment or quality description is *preference* or *interest*, where a subject is neither right nor wrong for finding a particular sound scene more agreeable or interesting than another (this emphasizes the point made in the introductory quotations; that hedonic judgments or opinions have no veridicality and therefore can not be evaluated in terms of being *factual*).

For evaluating multichannel audio systems, such as the spatial audio systems reviewed in the next section, ITU-R BS 1116 (1994) recommends that the following three aspects of sound quality are evaluated using a five point scale in terms of the perceived difference between a test stimulus (the "object") and a reference audio scene:

- Basic audio quality - This single, global attribute is used to judge any and all detected differences between the reference and the object.
- Front image quality - This attribute is related to the localization of the frontal sound sources. It includes stereophonic image quality and losses of definition.
- Impression of surround quality - This attribute is related to spatial impression, ambience, or special directional surround effects.

Alternatively, overall sound quality may be rated with a single mean opinion score "which conflates all aspects of sound quality, including preferences and descriptive characteristics, into a single rating" (Rumsey, 2002). The problem with this approach is that sentiments relating to sound quality are reported inconsistently due to their emotional loading. Spatial sound quality ratings of audio systems are therefore generally in terms of judgements of attributes which are known to relate to overall quality rather than holistic quality attributes as in the above list. Bech (2001) calls such sound characters which contribute to total auditory spatial impression the "primary attributes" and outlines three general methods for investigating them:

- Descriptive analysis.
- Repertory grid technique.

- Nonverbal graphical technique.

Descriptive analysis.

Descriptive Analysis (DA), as it applies to the investigation of auditory spatial imagery, assumes that a listeners' auditory spatial impression of a sound scene is constructed by the sum of a number of perceptual attributes each defined in terms of a sensorial strength (Bech, 1999). People undergoing a descriptive analysis test rate these attributes on scales of perceived intensity, which could be accomplished with either direct or indirect techniques. As Bech (1999) points out "*the subjects should be encouraged not to use preference or preference related words, such as good, bad etc.*" So descriptive analysis is about sound *character* rather than *quality* descriptions.

These attributes are generally provided by the experimenter. An example of some attributes for describing spatial auditory imagery is given by Toole (1983):

- Definition of sound sources ("definition" here means the same thing as image definition discussed in section 2.1.5).
- Continuity of sound stage.
- Width of sound stage.
- Impression of distance or depth.
- Abnormal effects (this could be considered a sound *quality* issue) .

Alternatively, the list of attributes could be arrived at by the listener themselves using a free description process (Guastavino and Katz, 2004)

or repertory grid technique (Berg and Rumsey, 1999) (both methods are described under the next heading). Zacharov and Koivuniemi (2001a) undertook an extensive investigation with twelve listeners and generated 532 different words to describe 104 different sound scenes (created with musical and non-musical sound recordings). Following a discussion amongst the listeners, this list was reduced to eight attributes (which are translated here from the original Finnish):

- Sense of direction.
- Sense of depth.
- Sense of space.
- Sense of movement.
- Penetration (if the imagery is located in the head, this is considered a *penetrating* attribute, i.e. the opposite of *externalization*).
- Distance to events.
- Broadness (width).
- Naturalness.

The attributes can be considered axes which define the perceptual space describing the total auditory impression. To reduce the number of attributes which the listener must rate (and therefore reduce the experiment time and task burden for the subject) these axes should be *orthogonal* so that the experimenter is not "asking the same question twice". Zacharov and Koivuniemi (2001a) showed how principal component analysis can be used to look

at the correlation between attributes and found that the data could be described using two axis; one of which was loaded (i.e. affected) by the "sense of movement" and "sense of space" attributes and the other by "broadness" and "penetration" whilst negatively correlated by "direction" and "distance" (the negative correlation between *penetration* and *distance* is, of course, expected).

Repertory grid technique.

Listening tests using descriptive analysis (DA) and the repertory grid technique (RGT) strive at the same thing: to investigate the structure of perceptual features of a perceived sound scene. As summarized by Bech (2001): *"The basic difference between DA and RGT is that whilst DA forces all subjects in the experiment to use the same set of words for their evaluations, RGT allows each subject to use their own word set".* The advantages of this are two-fold: firstly, biasing from the experimenter for how the listening experience is to be described can be removed. Secondly, with RGT it is not necessary to train subjects in the use and interpretation of the words used in test, unlike DA.

Once the stimuli have been generated, there are two stages to the RGT method:

1. Elicitation of attributes from free verbal descriptions of perceived similarity and dissimilarity between different audio scenes.
   In the pioneering study by Berg and Rumsey (1999), three sound stimuli (or "elements") were presented at a time and listeners were asked to describe a way in which two were alike and thereby different from the

third. Four common constructs were found relating to spatial features of the reproduced sound:

- Naturalness.
- Lateral image direction (i.e. azimuth) and image size.
- Envelopment.
- Depth.

The last three in this list relate directly to auditory spatial imagery and can be investigated using either a descriptive analysis approach or with a graphical mapping technique. The problems of investigating an impression of naturalness of a reproduced sound scene is more difficult, and is discussed later in section 2.1.7.

A grid (matrix) is then created with columns defined by the stimuli, and rows defined by these constructs. For each point on the grid (i.e. intersection of construct and stimulus), the stimulus is rated on a five-point scale in terms of that particular construct.

2. Matching of scenes and attributes- the experimenter can analyse the elicited attributes in terms of character and quality judgements, and look at the relationships between these judgements (for example, the experimenter can look at which sound character descriptions relate to, for example, listener preference).
   The grid can then be analysed with a number of techniques; principal component analysis (PCA); cluster analysis; or rank-order correlation. The result is that the relationships between the constructs (e.g. "Does *presence* mean the same things as *distance*?); the relationships between the stimuli (e.g. "Is sound scene A generally described in the same way as scene B?"); and the relationships between constructs and stimuli

(e.g. "Is there a high sense of image presence in scene A?") can be quantitatively measured.

Guastavino and Katz (2004) undertook a study which is similar to the RGT method, but the method for evaluating the constructs and the relationships between them were slightly different to that used by Berg and Rumsey (1999). Four different sound scenes were presented to 27 listeners and each subject could freely describe the scene using any adjective and qualifier they wished (e.g the adjective *presence* and the qualifier *outside*); a technique called "free verbalizations". The rate of occurrence of each adjective and the qualifier was then used to compare each sound scene. For example, the attribute *distance* could be described as being either *close* or *distant* and the number of times a particular scene was described as *distant* was counted and compared with comments from other scene descriptions. The judgements for each scene were then compared using a PCA-based technique to visualize the scene description on the axes of each attribute. The attributes elicited by this work were very similar to that found by Berg and Rumsey (1999) and Zacharov and Koivuniemi (2001a).

Graphical mapping techniques.

A pictorial representation is a useful method for describing auditory spatial imagery using the terms shown in figure 2.1. This is not to say that the aspects of auditory imagery such as perceived image size and location can not be reported using verbal techniques; but there is an obviously higher degree of isomorphism between a pictorial representation of geometric space than with a verbal representation. Ideally this map would have an ego-centric

perspective; that is, the listener would not have to do any translation from their real-world perception of space to the map. However, to do so in a way that ego-centric *distance* (or range) can be represented would be very difficult. Therefore, an exo-centric map such as the top-down or plan-view perspective shown in figure 2.1 is generally used, with the subject shown as the absolute reference for the distance dimension. Such a mapping system requires the listener to project the three-dimensional space of their physical reality on to a two-dimensional map; "*translating egocentric spatial perceptions into external representations*" (Mason et al., 2001).

To help the listener report the perceived image direction (and sometimes distance), pointing methods are often used. This intuitive approach has been adopted in a number of studies: With either a hand or a stick (Thurlow and Runge, 1967; Mershon, 1997); with the listeners head using a head-tracker to measure the direction (Middlebrooks, 1992; Shinn-Cunningham et al., 1998); or using a hand-held optical pointing device such as a laser-pointer (Oldfield and Parker, 1984; Choisel and Zimmer, 2003). Visual markers are also used to help the listener map the perceived image direction onto the graphical representation. A commonly used marker system is numbered cards on a screen in front of the listener (Queen, 1979; Nelson et al., 1997) or labeled loudspeakers (which are not all reproducing the sound stimulus) (Nielsen, 1993; Møller et al., 1996; Usher, 2001). The problem with using visual markers is that the reported direction can be influenced (biased) by the markers; a kind of spatial quantization.

The visual dominance over auditory cues for source localization is demonstrated well in an experiment reported by Shinn-Cunningham et al. (1998): A series of clicks was reproduced over headphones and filtered in a way so that when the listeners' eyes were closed, the clicks seemed to originate from

a particular direction (which is called the *auditory direction*). A light was also flashed from another direction (this is the *visual direction*). When the visual and auditory direction were not the same; at first the listener reported hearing the source somewhere between the auditory and visual direction, but after repeated presentations they reported hearing the sound in the direction of the flashing light. Furthermore, this "re-mapping effect" lasted even when the visual and auditory direction coincided- as if the mapping function in the auditory system which relates interaural acoustic cues to a sound source direction had been (at least temporarily) altered.

Free-drawing systems for mapping auditory images perceived in loudspeaker audio scenes have been used in many studies. With these systems, the perceived image location and extent (in the horizontal plane) is drawn either by hand (e.g. Chernyak and Dubrovsky, 1968; Wagener, 1971; Blauert and Lindemann, 1986; Woszczyk, 1993; Ford et al., 2001; Neher, 2004) or with a computer: (e.g. Mason, 2002; Usher and Woszczyk, 2003; Hanyu and Sekiguchi, 2004; Merimaa and Hess, 2004; Ford, 2005). Free-drawing schemes suffer from the problems of emotive bias, e.g. the listener may draw jagged image to represent a "jagged" sound, such as a trumpet (as was found by Woszczyk, 1993). Furthermore, it can be difficult to statistically interpret freely drawn shapes, e.g. to find the geometric centre of the object. As shall be seen in chapter 3 for the graphical mapping system developed to investigate spatial imagery in the new audio system, ellipses are used for representing images because the centre, width and depth can be easily calculated.

Ford (2005) has designed a "Universal Graphical Language" system for evaluating spatial auditory imagery experienced with loudspeaker audio systems in cars. In her system, the spatial extent and perceived origin of a

source image can be drawn using either hand-sketches or with a computer. In addition, the focus of the source image is marked with either a letter or another shape, and thirdly; a shape to represent the spatial envelope of the reverberance image (which she called "a feeling of space"). Regarding statistical interpretations of the elicited maps; for the hand-drawn maps (Ford et al., 2002) image width was measured manually (as a percentage width of the listening environment, which was a car) and the elicited image focal point was used to calculate the image direction. We shall see in chapter 3 how the new mapping system in this thesis can computationally measure all of the attributes (i.e. elicited image width, distance and azimuth) shown in figure 2.1 for both source and reverberance images.

2.1.7 Naturalness of reproduced sound

The term *naturalness* is often used in experiments to evaluate imagery in audio (e.g. Berg and Rumsey, 1999; Zacharov and Koivuniemi, 2001a). Natural sounding imagery in reproduced music is a tricky proposition; Theile (1991) asks: "How can the naturalness of the sound image be defined?" In audio, the reference is not necessarily a recollection of a perceived acoustic wavefield in a *live* concert-hall musical performance; "*there is no natural environment to imply or recreate and one is dealing with an artificial creation that has no natural reference or perceptual anchor*" (Rumsey, 2006). Evaluations of reproduced sound scenes are therefore taken in reference to another sound scene (e.g. created with a different electroacoustic system or stimulus) and perhaps a more ecologically valid word than *naturalness* is the word *fidelity*; *"fidelity implies a trueness of reproduction quality to that of the original"* (Rumsey et al., 2005a).

Theile (1991) concludes on the problem of naturalness in audio that to recreate the timbral and spatial properties of the imagery conjured when experiencing the original live performance is in most situations contradictory. For example, Mason and Rumsey investigated what the effect of different rear-microphone techniques were on "stereo" (i.e. source) image quality and reverberance image quality (naturalness was part of this quality measure). The results of their study highlight the paradox that, simply put, a high source image quality often means a poor reverberance image quality (Mason and Rumsey, 1999). The audio engineer must therefore make a compromise between capturing and reproducing the imagery in the recording environment and maintaining an aesthetic balance that is appropriate for reproduced concert-hall music in the absence of both musicians and a concert hall. The design criteria for the new system are principally governed by respect for the artistic mixing intentions of the audio engineer inasmuch as the spatial imagery of the recorded instrument (i.e. the source imagery) is maintained. This means that the source image for the recorded musical instrument in a sound scene created with the new upmix system should have the same spatial characteristics as if the recording was auditioned using a conventional loudspeaker pair arrangement, as the audio engineer intended the mix to be heard. In other words, the *frontal spatial fidelity* (Rumsey et al., 2005a) of the sound scene created by the new system must be maintained with regard to the original audio scene.

In two independent studies by Berg and Rumsey (1999) and Guastavino and Katz (2004), the terms *presence* and *naturalness* were very highly correlated. The sense of presence when listening to reproduced music can mean one of two things; the musician being present in the listening room or (more commonly) the feeling of presence of the listener in the recording environment. It is generally the latter aim that the recording and mixing engineers

try to achieve (Rumsey, 2002); likewise, the sound processing system introduced in this thesis is designed with the intention of reproducing sound in a way that gives the listener a convincing sense of presence in the original recording room. The importance of the auditory senses in the feeling of presence within a virtual environment is receiving increasing attention due to the advanced capabilities of teleconferencing enabling near-instantaneous full-duplex transmission of high quality multichannel audio across entire continents (de Bruijn, 2004; Woszczyk et al., 2005). Whilst this is not a motivation for the thesis, it is certainly a further direction for development of the new technology.

2.1.8 Panning sound around a listener in audio

In loudspeaker audio, time (delay) and amplitude panning refer to the positioning of a virtual auditory image in the horizontal plane by changing the relative time delay and/ or amplitude of the electronic signals fed to the loudspeakers. Panning generally refers to the placement of only the source image, however we shall see in an experiment in chapter 3 that some of the same methods for affecting the perceived direction of a *source* image can also be applied to a *reverberance* image as well.

Role of interaural cross-correlation (IACC) in panning.

In loudspeaker audio, *summing localization* is achieved by radiating coherent acoustical energy from a pair of loudspeakers (Blauert, 1997, pg. 203). What is meant by *coherent* is that the electronic signals feeding each loudspeaker have a cross-correlation which is close to either 1 or -1; in psychoacoustics,

the coherence between two signals is generally understood to mean the maximum absolute cross-correlation within a certain time limit (Blauert, 1997; Culling et al., 2001; de Vries et al., 2001; Faller and Merimaa, 2004). The correlation between the acoustic pressure measured at contralateral ear-drums of a listener is often used for investigating imagery in audio, because it has been shown to be strongly correlated with aspects of auditory imagery such as the width (Chernyak and Dubrovsky, 1968; Keet, 1968; Mason et al., 2005) and distance (Kurozumi and Ohgushi, 1983; Martens, 1999) of a source image; and the degree of listener envelopment (i.e. surrounding reverberance imagery) (Morimoto, 2001; Soulodre et al., 2003). The pressure can be measured either using small probe microphones in the listeners' ear canal or a dummy head with plastic pinna (often with an artificial, armless upper body called a head and torso simulator; HATS), and the microphone diaphragms located where the ear-drums would be (or at the entrance to the ear-canal). The correlation between the two electrical mike signals $m_l(t)$ and $m_r(t)$ is called the interaural cross-correlation coefficient (**IACC**) (e.g. Schroeder et al., 1974; Ando, 1985; de Vries et al., 2001), as defined in (2.1):

$$\mathbf{IACC} = \rho_{lr}(\tau) = \frac{\int_{t1}^{t2} m_l(t) m_r(t+\tau) dt}{\sqrt{\int_{t1}^{t2} m_l(t)^2 dt \int_{t1}^{t2} m_r(t)^2 dt}}. \tag{2.1}$$

The **IACC** is a vector (hence it is in **bold** text) of time index τ, and (as mentioned) the maximum or minimum (whichever absolute value is larger) value within the time limit $t2 - t1$ is the interaural coherence (IAC). Commonly used time integration intervals are ± 2 ms (Culling et al., 2001); ± 1.5 ms (Braasch et al., 2004); or ± 1 ms (Schroeder et al., 1974; Faller and Merimaa, 2004). The correlation between a pair of electronic signals from two microphones can be calculated in the same way (using (2.1), where $m_l(t)$ and $m_r(t)$

are the two signals) to give the interchannel cross correlation **ICCC**, and the maximum absolute value is called the interchannel coherence (ICC).

The relationship between the IAC and the ICC is that ICC's close to zero give slightly higher IAC's due to the acoustic cross-talk between the two ears (or microphones) from diffraction around the head and low-order sound reflections. ICC's close to 1 or -1 give IAC's closer to zero due to the decorrelating affect of high-order sound reflections in the room. A frequency-dependant model to map the **ICCC** to **IACC** for a loudspeaker pair at $\pm 30°$ (i.e. the ITU-R BS 775-1 2/0 configuration- discussed later) was given by Kim et al. (2005), though this did not include any room reflections. An empirical measurement was made for neighboring loudspeaker combinations for the 3/2 ITU-R BS 775-1 configuration in the MARLAB (see appendix B.1), as shown in figure 2.2 (see figure caption for measurement details). The lowest IAC was measured for the rear loudspeaker pair (LS-RS), which is convenient as the rear loudspeakers are generally used for radiating sound which contributes to reverberance imagery (i.e. for sound recordings mixed for reproduction with loudspeakers arranged to the 3/2 or 2/2 ITU configuration), and these sound components should create a low interaural correlation in order to give a diffuse and natural sounding reverberance image (the interaural coherence in a reverberant field is close to zero; Jacobsen and Roisin, 2000). It is also interesting to note that the IAC-to-ICC mapping was not symmetrical- i.e. the curve for the L and LS loudspeaker pair is different than R-RS pair. This is probably due to inter-speaker non-linearities, asymmetric sound reflecting objects in the listening room and absorption properties of the room boundaries.

Especially for pop-music recordings, low frequency musical instruments are often mixed out of phase (i.e. with an ICC of -1), so a negative IAC

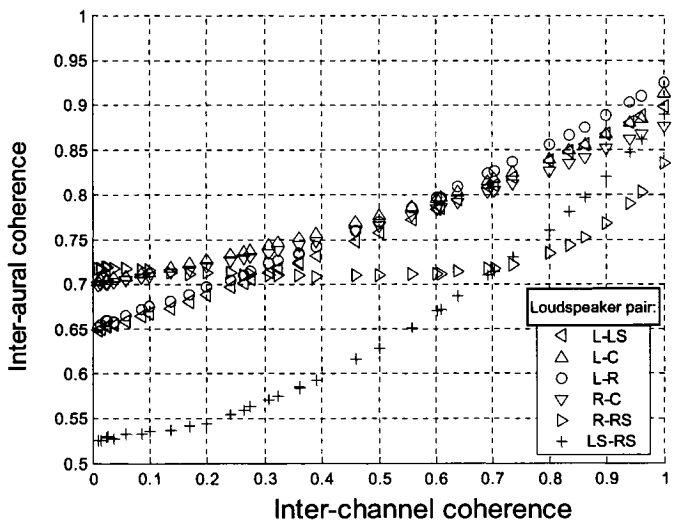


Figure 2.2: Relationship between interchannel coherence (ICC; i.e. the maximum interchannel cross-correlation in a ± 1 ms window) between white-noise signals fed to loudspeaker pairs and interaural coherence (measured using dummy head-plus torso with pinna at the listening position). Loudspeakers were arranged according to 3/2 ITU-R BS 775-1 configuration (rear loudspeakers at $\pm 120°$) in the MARLAB. ICC was manipulated using the superposition technique (Blauert, 1997).

at low frequency is not uncommon in loudspeaker audio. The sign of the IAC is important and has been shown to affect spatial imagery in audio; generally speaking, when the IAC is positive, auditory images are heard further away than for negative IAC's (Kurozumi and Ohgushi, 1983; Kendall, 1995; Martens, 2001). This applies especially for low frequencies but is not so relevant to high frequency IAC due to the lack of phase locking for high frequency stimuli by hair cells on the basilar membrane (Moore, 1997), which is why the **IACC** is often calculated with the high frequencies attenuated (de Vries et al., 2001).

The interaural cross-correlation can be used as a predictor as to how likely it is that a single image percept between the loudspeakers radiating the signals may occur (e.g. Damaske and Ando, 1973; Okano, 2000). The psychophysical model by Faller and Merimaa (2004) predicts that the interaural time and level difference cues (Moore, 1997) are used for source direction estimation within a critical band (CB) only when the correlation in that CB is above a certain threshold. This model does not assume a fixed "hard" threshold, but one related to the standard deviation of the interaural time and level differences. Chernyak and Dubrovsky showed that using noise presented with headphones with an IAC of 0.4, a single, spatially unified image was generally heard (Chernyak and Dubrovsky, 1968).

Variation in the **IACC** over time has been shown to affect the width of source images (Mason et al., 2005). This can occur in concert halls due to "beating" from interference between strong (low order) sound reflections and the direct sound (or other reflections), and the ASW of an auditory image increases as the modulation rate and depth increase (Mason et al., 2005). This makes measurement of **IACC** with a single pair of impulse responses or from a time-averaged measurement difficult and could give a misleading

conclusions about how the source image may be perceived.

Because of the usefulness of the **IACC** in predicting auditory spatial imagery and the linear relationship with **ICCC**, the electronic signal correlations between the outputs of the new sound processing system are used to evaluate the specific subjective design requirements with the electroacoustic design criteria introduced in chapter 4.

Amplitude panning in front of the listener

When a single audio signal is fed to a loudspeaker pair with a signal gain g_1 and g_2 for each loudspeaker and a single image is heard between the two loudspeakers, the process is called *summing localization* (Blauert, 1997), *intensity stereophony* (Theile and Plenge, 1977) or (vector based) *amplitude panning* (VBAP) (Pulkki, 1997). When these loudspeakers are in front of and at the same height as the listener, the perceived direction of the image can be predicted for a variety of signals with an accuracy of a few degrees according to the *tangent panning law* and predicts image direction better than Blumleins' classic stereophonic law of sines for mobile-head listeners (Bernfeld, 1973; Pulkki, 1997). This is summarized in (2.2):

$$\frac{\tan \varphi}{\tan \varphi_0} = \frac{g_1 - g_2}{g_1 + g_2}, \tag{2.2}$$

where the angle between the loudspeakers is $0° < \varphi_0 < 90°$ and the predicted image direction φ is $-\varphi_0 \leq \varphi \leq \varphi_0$, with speaker gains $g_1, g_2 \in [0, 1]$.

The within-subject and between-subject consistency for reporting virtual image direction (which in the literature applies only to source images) is related to the "localization blur" (Blauert, 1997; Corey and Woszczyk, 2002)

and is expressed in terms of a standard deviation (SD) or inter-quartile range for different direction judgements from the same or different people. In a study involving both virtual images created with VBAP and real images using 400 ms octave-band-limited noise bursts reproduced from loudspeakers, Pulkki and Hirvonen (2005) found the localization blur for a real source at 0° to be less then 2° (SD) and for a virtual image panned at 15° (with channel gain coefficients according to the tangent-law) the localization blur SD was <4°, regardless of frequency. In a similar study using a variety of anechoic and noise stimuli, Choisel and Zimmer (2003) found the between-subject variation in reported image azimuth to be between to 1° and 2° (SD) for both real and virtual sources located from 0° to 30°. The localization blur is distinct from auditory source width (ASW); localization blur is generally taken as the variation in reported azimuth for auditory images for a given stimulus, whereas ASW is a measure (generally in degrees) of the perceived extent of an auditory image in the horizontal (i.e. lateral) plane from the perspective of the listener (Blauert, 1997). When the source image is located between the front left and right loudspeakers, ASW is dependant on the reproduced signal (Merimaa and Hess, 2004; Usher and Woszczyk, 2003) and is positively correlated with loudness (Keet, 1968; Usher and Woszczyk, 2003). Also, timbral artifacts are introduced by amplitude panning and it has been found that for VBAP with loudspeakers at ±30°, colouration is related to localization blur, being maximal for images panned at ±15° (Pulkki, 1999a).

Amplitude Panning to the side of the listener

When the loudspeaker pair is located to the side of the listener, such as the ITU-R BS 775-1 recommended front-right and rear-right loudspeaker locations (see figure 2.10), the relationship between perceived image direction

and inter-speaker signal gain (in dB) is not a smooth function. Theile and Plenge (1977) used a loudspeaker pair located to the left of the listener, with the front and rear speakers at 50° and 110° to the central axis- i.e. as if the front loudspeaker pair had been rotated by 80° to the left. The localization blur was largest when the interchannel level difference was small; for a level difference of 0 to -6 dB (that is, with the front speaker softer than the rear), the inter-quartile range was approximately 50°. Various "angles of rotation" were investigated: 40°, 60°, 80° and 90°, and the results are summarized in table 2.1.

The pulling of an auditory image away from the direction predicted by the tangent panning law is called *detent*; (Gerzon, 1992b). For the study by Theile and Plenge (1977), the degree of detent was greater for side images; as shown in table 2.1. Similarly, Ratliff (1974) found that for a quadraphonic "square" array (loudspeakers at ±45° and ±135°), when panning between a side pair of loudspeakers with an interchannel level difference of 0 dB the image direction was reported at approximately ±60 ° and described as "very diffuse" and "very jumpy".

Using the ITU-R BS 775-1 loudspeaker arrangement (rear loudspeakers at ±120°) Corey (2002) found a similar forward-pulling effect as that found by Theile and Plenge (1977): Corey investigated pair-wise panning of an anechoic source at a variety of intended directions between the front and rear loudspeakers; 45°, 65°, 80° and 100°. The 45° and 65° sources were reported about 5° and 10° towards the median plane, whilst the 80° and 100° sources were pulled towards the rear loudspeakers by a similar amount. Also, the direction of side images in ITU-R BS 775-1 systems are reported with less *certainty*, with the lowest certainty for images reported in the direction of 80° (Corey and Woszczyk, 2002) and 90° (Lund, 2000).

Rotation δ	Reported image direction φ	Detent (δ - φ)	Blur
0	0	0	0
40	38	2	8
60	55	5	15
80	68	12	40
90	77	13	45

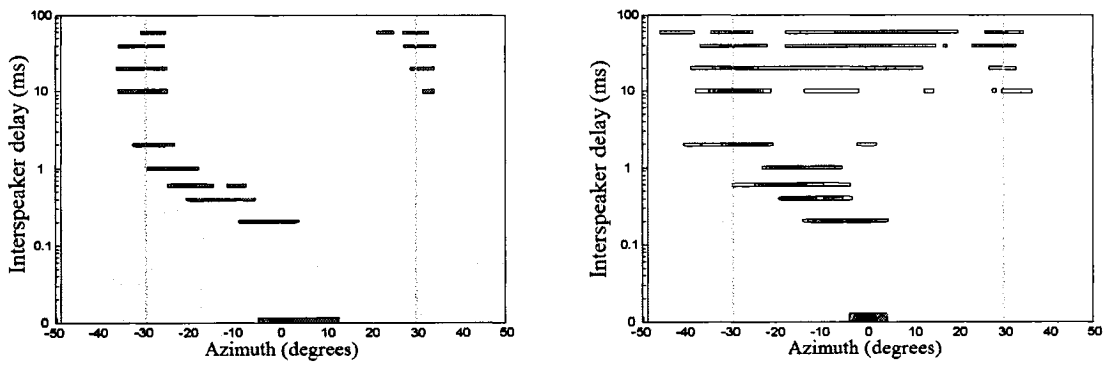
Table 2.1: Data from the Theile and Plenge (1977) study (units for all values is degrees) showing how a source image is pulled towards the "straight-ahead" 0° bearing as a speaker-pair is rotated to the side of the listener and how localization blur increases (blur is the inter-quartile range of the reported image direction, in degrees). The inter-speaker angle was 60° for all five loudspeaker configurations, so the configurations can be considered as a rotation around the listener of the front left and right loudspeakers in the 2/0 ITU-R BS 775-1 arrangement. The stimuli were coherent white noise and anechoic speech and each loudspeaker signal had the same gain. δ is the rotation angle and φ is the reported image azimuth (median). In this thesis, the same polar zero-degree reference is used; defined as that point half way between the front left and right loudspeakers.

Time-delay panning in front of the listener

When a single audio channel is fed to a front loudspeaker pair and one of these channels is delayed (i.e. with an inter-aural coherence close to unity at lag $\tau \neq 0$), then a single auditory image will appear at a location in the direction of the non-delayed channel as long as the absolute value of the lag of this peak is below about 1.5 ms; an effect called the *law of the first wavefront* or the *precedence effect* (Blauert, 1997). As the delay is increased, a second image is heard localized at the lagging loudspeaker as well. For a 2/0 loudspeaker system (i.e. a front loudspeaker pair), the relationship between interchannel delay and the perceived image direction is shown in figure 2.3.

Time-delay panning can be implemented either electronically during mix-

ing or "naturally" using spaced microphone pairs and placing the sound source (e.g. musical instrument) closer to one microphone than the other. We will look at how the new upmix system behaves when time delay panning occurs, and how the microphone configuration and sound source location affect this time delay.



(a) Anechoic speech.

(b) Anechoic double bass.

Figure 2.3: Reported image direction and extent (in horizontal plane) for time-delay panning of two stimuli between front loudspeaker pair (loudspeaker pair at $\pm 30°$) for various interchannel delay times. Response from 8 subjects with 3 test runs. Responses were drawn using a computer GUI similar to that developed in chapter 3 and overlayed to create density plots. The response for each delay shows a cross-section through the density plot ignoring the reported image distance. The darker parts show directions where listeners reported source images more often. (Adapted from Usher and Woszczyk, 2003.)

2.2 Loudspeaker Spatial Audio Systems

2.2.1 Historical overview

Probably the first example of what s generally called a "loudspeaker spatial audio system" is that demonstrated by Clément Ader in 1881 at the Paris International Exhibition of Electricity (Hertz, 1981). Ten transducers, each consisting of ten pencils with a thin wooden diaphragm, were used to convert the sound of a musical performance at the Paris Grand Opera into an electrical signal that was reproduced with a pair of telephone receivers to listeners three kilometres away. As the signal to the headphone receivers were different for each ear, a reasonable stereophonic image could be heard reflecting the spatial balance of the live music (the balance would be affected by which microphone pairs were being listened to). Since then, in terms of business the driving force for loudspeaker spatial audio system design has been the film ("movie") industry. Indeed, the first major multichannel loudspeaker system was for the Disney film *Fantasia* in 1938. Using three separate audio tracks steered to five loudspeakers around the listener, it was a major coup for a new art form and continues to surprise us in its startlingly creative interpretation of major works of classical music.

For reproduction of recorded music, spatial audio systems have (historically speaking) been exploited primarily for artistic rather than commercial endeavours. One of the earliest examples of music composed specifically for reproduction with a multichannel loudspeaker system is still one of the most technically impressive; the performance of Edgard Varèses "*Poème Electronique*" in the futuristic Philips Pavilion at the 1958 Brussels World's Fair. The recorded sound was reproduced from a three-track tape machine, amplified

with 20 amplifiers and relayed using a 15-track control steering signal from another tape machine to 350 loudspeakers (Tak, 1958).

Since the introduction of affordable home theatre systems over the last decade, generally comprising of a DVD media playback for video and audio and five "surround" loudspeakers, commercial releases of various genres of music mixed explicitly for reproduction with five loudspeakers has increased and spatial audio systems are a common feature in homes and cars throughout the world (Holman, 2000a): These systems are discussed shortly under the heading of *discrete audio systems.* Furthermore, the owners of home theatre systems now have a means for converting their existing two-channel audio collection (e.g. their mp3's, CD's or LP records) for reproduction with four or five loudspeakers. This thesis presents such a system; an *upmixer*, and these systems are discussed under the heading *Blind Adaptive Audio Upmixers.*

2.2.2 Classes of spatial audio systems

An extensive review of all approaches to spatial audio systems would have to include the huge diversity of systems for electroacoustic music compositions. Most of these systems are an inherent part of the composition itself (such as the *Poème Electronique*) and as there are constantly new compositions for spatial audio the list would be huge; a review of some pieces since the 1960's is given by Trochimczyk (2001).

This discussion of spatial audio systems is restricted to music reproduction with loudspeakers (as this is the context of this thesis). However, there are many uses of spatial audio systems for communication. Such an exam-

ple for human-computer interaction is spatial auditory displays, where spatially separating different verbal messages (e.g. computer-generated alerts) for headphone signal presentation can significantly improve the intelligibility compared with reproducing the messages in a way that is perceived as originating from the same location (Begault and Erbe, 1994); this is an example of *spatial unmasking* and is discussed more thoroughly in the next chapter. Another fast-growing use of spatial audio is for *telepresence*; evoking an illusion on someone so they perceive they are located in another environment by presenting a variety of stimuli to the persons' senses (Fisher, 1990); sound (including structure-borne vibrations) and vision being the main ones (Martens and Woszczyk, 2003; Woszczyk et al., 2005), which may be for a variety of purposes using a variety of methods. In the foregoing discussion, classes of spatial audio systems are grouped by implementation technique rather than the motivation (or application) for the particular system. In other words, *motivation* is nested within *method.*

Wave field and wave front reconstruction systems

Functionally speaking, spatial audio systems can be categorized by their approach to inflicting the auditory percept on the listener; a distinction summarized by Gautier et al. (2004) as "simulation of sound fields" and "simulation of the perception".

Systems striving for "simulation of sound fields" (and their recording methods) are characterized by attempting to recreate the recording environments' acoustic wave field or wave*front*; the distinction being that wavefront simulation just concerns the propagating direct sound wave from the source (here, a source could apply to a sound-reflecting object) whereas wave

field synthesis recreates both the propagating and stationary parts such as room modes. True wavefield synthesis is difficult to achieve due to acoustic interactions with the reproduction room and the transducers themselves. This recreation may be at just a point in the room (such as Ambisonics), a number of points (cross-talk canceler) or over an area (wave field synthesis); these three systems are described shortly. On the other hand, "simulation of the perception" means attempting to create any wave field that is necessary to give just the same *perceived* impression of the original wave field. The new system presented in this thesis (and most extant spatial audio systems) fall into this category, but methods for creating an acoustic field in a room will also be discussed to give the reader an idea of the state of the art for loudspeaker spatial audio.

•Wave Field Synthesis (WFS).

The basic premise for WFS is from a 17^{th} Century theory developed by the Dutch scientist Huygen. The theory is: that an acoustic wave field created by a single sound source (such as a musical instrument) can be recreated by an infinite number of secondary sound sources spread out on the surface of the wavefront. An everyday example of Huygen's principle is when a sound-source (e.g. musical instrument) is obstructed by a wall with holes in. Here, for an observer on the other side of the wall each hole can be treated as a separate sound source and even if the occluding object is only 10% transparent then due to diffraction from the holes the transmitted wave has the same angle of direction as the incident wave (Cremer and Müller, 1982, pg. 181).

For audio applications this theory is realised with the help of the Kirchhoff integral (Berkhout, 1988) and a signal processing and transducer system typically comprising of approximately one hundred full-range loudspeakers

(at the same height) in a polygon around the listener(s) (Verheÿen, 1998). In a relatively damped room, accurate reconstruction of the intended wave field for multiple sound sources is possible for frequencies in the range of about 100 Hz to 1.5 kHz (the lower limit determined by the bass-woofer driver response, woofer-spacing and room interaction effects, and the upper limit is inversely proportional to the tweeter-unit spacing; Verheÿen, 1998). It is therefore more apt to call the method a wave *front* reconstruction process, due to reflections from the room boundary and the loudspeaker array. However, recent active control methods can compensate for these room effects (to a degree) using the WFS system to reduce room modes and strong reflections by absorbing unwanted acoustic energy (Spors et al., 2003; Gautier et al., 2005). With a fully surrounding WFS system, there is no directional bias and the location of the perceived auditory images is independent of listener position and orientation, over a large fraction of the listening area. Therefore, in the horizontal plane the auditory imagery can be both *homogenous*; "one in which no direction is preferentially treated", and *coherent*; "one in which the image remains stable, subject to no significant discontinuities if the listener changes position within it" (Malham, 1999).

As implied, WFS strives at wave field synthesis over an *area*. Furthermore, real-time manipulation of the position of recorded or live sound sources (i.e. recorded and mixed to a signal electronic channel) is possible, so dynamic movement of sources can be simulated (including convincing doppler effects) typically using four computers. Therefore the sound source can be positioned in terms of both direction and distance at any location in the horizontal plane, including locations *within* the loudspeaker array (Usher et al., 2004b).

An interesting feature of WFS is its compatibility with existing conven-

tional discrete multichannel audio mixes (e.g. reproduced from a commercial 5.1 channel film or musical recordings on a DVD-A). As described by Boone et al. (1999), these recordings can be reproduced by simulating each of the five discrete channels as a virtual plane wave incident at the same angle as discrete loudspeakers would be (e.g. arranged according to the conventional 3/2 ITU-R BS 775-1 standard- discussed shortly). Boone et al. claims this has a number of advantages compared with using five separate loudspeakers; the virtual location of each loudspeaker is not restricted by the size of the listening room; and because each loudspeaker source is simulated as a plane wave there are no frequency-dependant beaming effects, increasing the area of the "sweet-spot".

•Cross-talk canceler.

Also called a crossfeed canceler (Atal and Schroeder, 1962; Schroeder, 1970) and the stereo-dipole (for rather obscure and illogical reasons; Nelson et al., 1997), this system uses a pair of conventional loudspeakers (ideally in an anechoic room) to create a sound pressure at the two ears as if the listener is wearing headphones fed with the same signals fed to the loudspeakers. The advantage of this is that the auditory images are *externalized*; perceived outside of the head rather than *within* the head as is often the case with headphone audition (Kendall, 1995; Begault and McClain, 2001). The reproduced signals are (ideally) *binaural recordings* made with a dummy head or by convolving a single audio channel with a pair of filters; each filter representing the impulse response from a single source to each of a listeners ear, in a particular room.

In the simplest form of a cross-talk canceling system, removal of the cross-talk between the loudspeakers to each ear (i.e. the left loudspeaker going to both the left and right ear instead of just the left ear) is accomplished

by putting a physical barrier between the loudspeakers. This large physical barrier is probably too intrusive for most people and the reproduction sounds over-damped due to excess high-frequency sound absorption by the barrier. A signal processing technique to accomplish acoustic cross-talk cancelation for two-loudspeaker reproduction was originally described in a patent by Atal and Schroeder (1962) and the basic idea is summarized in figure 2.4.

Cross-talk canceling systems must also take into account the head-filtering affect of the cross-talk signal, which is dependant on the angle of incidence of the loudspeaker (i.e. the head related transfer function). Ideally these filters should be calculated individually for each listener (i.e. empirically measured), otherwise the cross-talk performance (e.g. the difference between intended and reported image direction) for different listeners will vary (Takeuchia et al., 2000). Furthermore, head-movement will affect virtual source localization (head rotation should be kept below $\pm 10°$; Schroeder, 1970), though this affect is reduced with closely-spaced loudspeakers (e.g. at 10°; Takeuchia et al., 2000).

•Ambisonics.

The Ambisonics recording, signal processing, and reproduction principle is that the wave field at a point in the recording room (both travelling and standing wave components) can be closely approximated at a point in the listening (i.e. reproduction) room using a special microphone and a number of surrounding loudspeakers (Fellgett, 1974; Gerzon, 1977). In the recording, the output of a *Soundfield* microphone, consisting of four orthogonally arranged yet nearly coincident microphone capsules, can be used to represent (i.e. approximate) the particle velocity for the x, y and z axes and the pressure at the centre of the capsules- a representation in terms of the wave fields' *spherical harmonics*. From these microphone signals (stored according to the

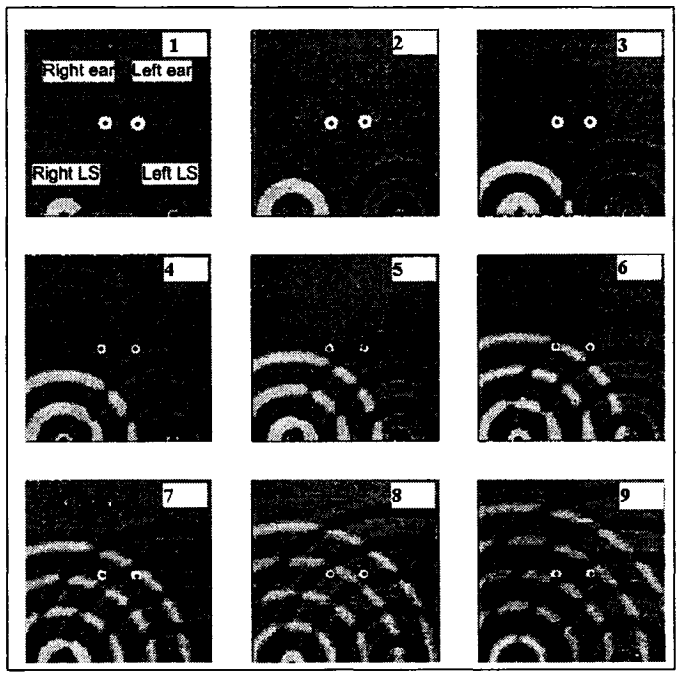


Figure 2.4: Illustration of the principle for cross-talk canceling systems (adapted from Nelson et al., 1997). The example shows a dirac signal that is intended to be heard by the left ear only. The effect is accomplished by radiating an out-of-phase signal from the right loudspeaker (white wave-front) so that it cancels (i.e. in terms of pressure) the cross-talked signal from the left LS when it arrives at the right ear (shown in the 6th frame). The right LS must also radiate a signal to cancel the new cross-talked signal from right LS to left ear (7th frame)- and so on.

UHJ format Rumsey, 2001), the signals which need to be fed to a number of loudspeakers around the listening position can be derived, typically with between six and eight loudspeakers (Farina and Ugolotti, 1999).

A particularly interesting (and topical) application of Ambisonics is for *Spatial Impulse Response Rendering* (Merimaa and Pulkki, 2005): measuring an acoustic impulse response in three orthogonal dimensions using a Soundfield microphone. This allows a convincing reproduction of the recorded space using multiple surround loudspeakers; an anechoicly recorded source is convolved with a different impulse response for each loudspeaker it is reproduced with, so as to maintain the spatially dependant acoustic transfer function of the recorded space.

A number of studies have compared imaging properties of these wavefield/ wavefront reconstruction systems with each other- e.g. WFS and Ambisonics (Nicol and Emerit, 1998), stereo-dipole and Ambisonics (Farina and Ugolotti, 1999), conventional two-loudspeaker audio and Ambisonics (Pulkki and Hirvonen, 2005) or all three (Guastavino and Katz, 2004). Objective measures of auditory spatial imagery are generally related to *source* image localization as this is easier to directly compare (e.g. in terms of the response variation for perceived image direction). However, because recordings for these systems use different recording and mixing techniques, comparing these systems in a general way can be considered a case of "apples and oranges".

Linear upmix systems

What is meant by linear is that the signal data flow structure in the system is unchanged by the condition of the input signals. A linear system obeys the

principle of superposition (PoS), which can be summarized like this; the PoS requires that the response of a system to a weighted sum of signals be equal to the corresponding weighted sum of the outputs to each of the input signals (Proakis and Manolakis, 1996, pg. 65). Linear audio upmix systems can be split into those that use *encoded* input signals and those that don't. An encoded signal means one which has been specially mixed from a combination of n_1 original audio channels and is converted to a smaller number of channels n_2, which is later decoded back to n_1 signals for reproduction with (at least) n_1 loudspeakers. Both the encoding and decoding processes are accomplished by scaling and adding and/or subtracting the input signals (i.e. n_1 input signals are used for the encoding and n_2 for the decoding), and as the scaling parameters are fixed these systems are also called *linear matrix converters* (Miles, 1996; Avendano and Jot, 2004).

•Madsen system.

The system presented by Madsen (1970) is not really a signal processing approach to spatial sound enhancement, but it is the simplest example of a linear upmixer and one which is referred to later on. An overview of the system is given in figure 2.5: A pair of unencoded two-channel signals (for example, from a CD player) are radiated from a front pair of loudspeakers in addition to two delayed copies of these signals radiated from another loudspeaker pair to the side or behind the listener; there are no gains applied to the rear loudspeaker signals. That's it! In accordance with the precedence effect (Blauert, 1997, pg. 225) the auditory spatial imagery of the source image (e.g. of the musical instrument) should be unaffected if the relative time of arrival of a wavefront from the rear loudspeakers to the listener is between about 2 and 30 ms *after* the wavefront from the front loudspeakers. If the delay is greater than this, then the signal fed to the rear speakers must also be attenuated so that only one source image is heard (i.e. the signal radiated

from the rear loudspeakers must be below the echo-threshold, which follows a trend shown in figure 2.6).

Begault and McClain (2001) found that the threshold for detection of a single reflection was dependant on the angle of incidence: thresholds reduced (i.e. the delayed sound could be heard easier) as the delayed sound was reproduced from an increasingly different direction from the earlier sound. This has consequences for the Madsen system because the threshold curve in figure 2.6 would be lowered as the side loudspeakers are moved farther away from the front speakers.

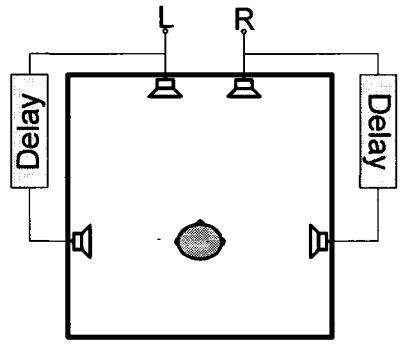


Figure 2.5: Overview of Madsen "ambiance extraction" system (Madsen, 1970). The side/ rear channels are simply a delayed copy of the front channels. The delay ensures that source image appears between the front loudspeakers according to the precedence effect.

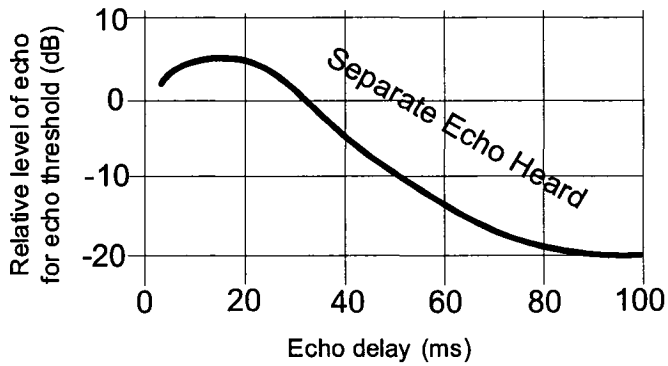


Figure 2.6: Echo threshold for continuous speech signal with single echo reproduced from loudspeaker at 40° and original sound from loudspeaker at -40°. Threshold is defined as the level of delayed sound at which two sounds are heard (from work of Haas discussed by Blauert, 1997, pg. 225). In the Madsen system (Madsen, 1970), suggested delays for the side loudspeaker channel are 2.5-10 ms. In the "K-surround ambiance extractor system" (Katz, 2002), suggested delays are at 30 ms intervals and each delay is 15 dB less than the previous. A theory to explain the rise in echo threshold at approximately 10 ms is given by the author in a past paper (Usher and Woszczyk, 2003), which, put simply, proposes that if the delay is consistent with the inter-speaker time delay (which for typical loudspeaker pairs at $\pm 30°$ is about 10 ms), then the suppression effect is strongest because the delay can be interpreted by the auditory system as a real sound reflection off a surface at the delayed loudspeaker.

•Hafler Stereo/Hookup.

The Hafler system is another converter which works with a pair of unencoded audio channels to create two new signals (Woram, 1970; Hafler, 1972). The idea of bridging the input signals to create the centre channel has been used since at least 1934 by researchers at Bell laboratories (Klipsch, 1958). This system is particularly interesting as it is very simple in essence yet all modern upmixers rely on the same fundamental principle: creating a centre-loudspeaker channel from the sum of the two input signals and a rear ("surround") channel from a difference signal.

The summation (or subtraction) of two signals of unit amplitude but random phase gives an RMS value of $\sqrt{2}$ (Klipsch, 1958). Therefore, if the input signals were uncorrelated the sum and difference signal levels would be about 3 dB more than the two input signals,[8] which is why the bridged centre-channel signal and rear difference signals are attenuated by 3 dB. There are many ways of implementing the idea; one elegant passive solution is suggested in a patent by Harrison (1995), as summarized in figure 2.7.

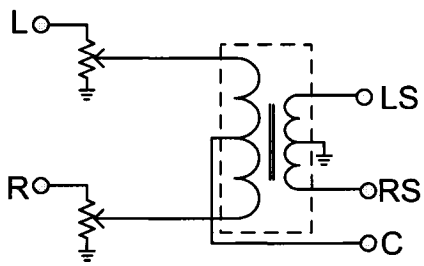


Figure 2.7: Passive signal processing mechanism for Hafler-Hookup using two input signals L and R and a transformer to create a pair of rear loudspeaker signals LS and RS and a single centre loudspeaker signal C (Harrison, 1995).

[8] $20 \times \log_{10}\sqrt{2}=3$

•Gerzon optimum reproduction matrices for multispeaker stereo.

A modification of the centre-channel derivation method using the "bridged centre" is described by Gerzon (1992a) for reproducing n_1 original channels with a greater number of n_2 frontal loudspeakers. This system is an inherent part of the "Trifield" sound processing system which is used in some commercial upmixers.[9] There are two principles which govern the design of the linear upmixing matrix: firstly, that the total energy of the input signal is preserved during the upmix process (the "energy preservation criterion"). And secondly; that the localization properties of the source images in the upmixed scene with n_2 loudspeakers be similar to that in the original scene with n_1 loudspeakers. As alluded to in the introduction, the new spatial audio system introduced in this thesis shares the latter design criterion.

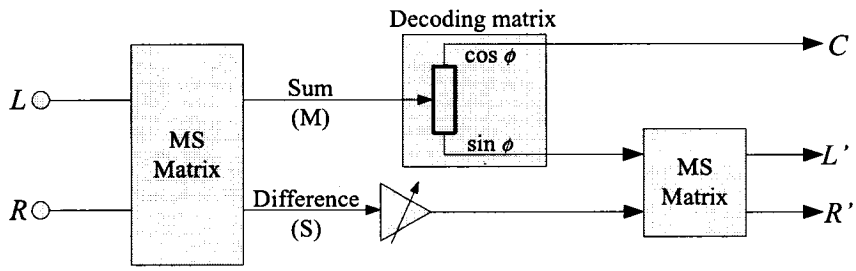


Figure 2.8: Overview of Gerzon matrix decoder (Gerzon, 1992a) to produce a centre channel C and two new channels L' and R' from two input channels L and R. The sum (M) and difference (S) signals are calculated according to (2.3), and are decoded to create L' and R' according to (2.4). The box with the arrow in the decoding matrix can be considered a signal divider, with a position relative to the top (i.e. to centre speaker) of $\cos\phi$. The variable signal gain on the S channel is an image width controller.

To understand the basic idea of the Gerzon matrix upmix approach, we shall consider the case when $n_1 = 2$ and $n_2 = 3$ as summarized in figure 2.8.

[9]Such as some "surround decoders" manufactured by Meridian.

The first stage of the upmix process is to create a sum (M) and difference (S) signal from the input signals L and R according to (2.3):

$$\begin{aligned} M &= 2^{-1/2}(L+R) \\ S &= 2^{-1/2}(L-R). \end{aligned} \tag{2.3}$$

The -3 dB gain is applied as discussed in the previous section. It is trivial to show that recreating the original signals (i.e. the second MS matrix in figure 2.8) according to (2.4) preserves the total energy in the original signal pair.[10]

$$\begin{aligned} L &= 2^{-1/2}(M+S) \\ R &= 2^{-1/2}(M-S). \end{aligned} \tag{2.4}$$

The proportional of summed signal fed to the centre speaker relative to that sent to the second MS matrix is controlled using a weighting scaler $\cos\phi$, as shown in figure 2.8. If $\phi = 90°$, there is no centre channel and the left and right loudspeakers simply radiate the original signals, whereas if $\phi = 0°$, no M signal is sent to the LS and RS loudspeakers, which are fed S and $-S$ (respectively). For a three loudspeaker setup with the base angle between the left and right loudspeaker equal to 90°, a compromise of approximately $\phi = 35°$ is recommended. This value was chosen by an analysis of the predicted image width for a range of time and level panned signals. The "trick" is twofold: To have the time-panned image and level-panned image coincide as a function of panned image direction; and to have the image direction change smoothly as a function of interchannel time and level differences. This analysis was undertaken using data regarding time panning ("velocity vector theory") and amplitude panning ("energy vector

[10] As $M^2 + S^2 = L^2 + R^2$.

theory") to predict image location when three loudspeakers are used (i.e. the same idea behind vector based amplitude panning; Pulkki, 1997). The psychophysical mapping functions for these two theories change with frequency; for example, an image created using time-delay panning with a loudspeaker pair may appear at 14° for low frequency stimuli (< 1.7 kHz), but when higher frequency stimuli are used the same interchannel time delay would cause the image to be located much closer to the 0° azimuth (Pulkki and Karjalainen, 2001). To overcome this, Gerzon suggests using a number of parallel frequency-dependant decoding matrices, each with different values of ϕ. A lower value of ϕ is recommended for the low-frequency M signals, which has the effect the there is more summed high-frequency signal sent to the side (L and R) loudspeakers (in other words; the high-frequency parts of the auditory images are "forced" to be wider).

•Dolby Surround.

Developed by Dolby Laboratories for 35 mm motion pictures in 1976 and used in many films in the 1980's, this system encodes three front and a single surround channel onto two tracks (which are stored either optically alongside the film negative with Dolby Stereo or magnetically for Dolby Surround; Dolby, 2005). The encoding process is described in a classic patent by Scheiber (1972), as summarized in figure 2.9. The decoding process is simply the reverse of the encoding; the surround (S) channel is created from the difference of the encoded signals L_t and R_t and the centre channel from the sum of these signals.

The surround channel can be from the output of a single microphone in a reverberant field (or the output of an artificial reverberator), or as suggested by Woszczyk (1990), created using a separate pair of microphones. This latter method is appropriate for concert hall recordings, as the natural

temporal properties of a diffuse field can be captured using coincident microphones (only a single surround channel can be created with the Dolby Stereo system, so spaced microphones would create strange phase problems when summed). Woszczyk (1990) found that a coincident pair of cardioid microphones facing in opposite directions (i.e. a figure of eight response, but with both lobes in-phase) gave the best spatial imagery, compared with a variety of other microphone techniques for both conventional two-loudspeaker (2/0) reproduction and reproduction using a Dolby Stereo decoder.

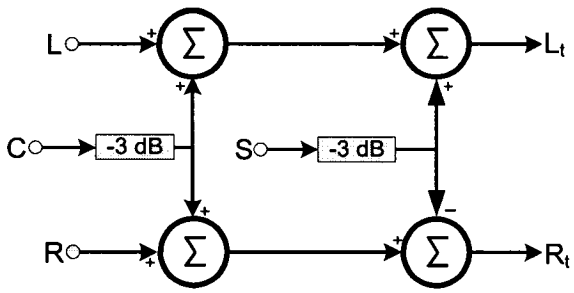


Figure 2.9: Dolby Surround/Stereo encoder (based on the Scheiber (1972) patent). Using the four input signals (the front pair L and R, centre channel C, and the surround channel S), two *encoded* signals are created (L_t and R_t). The surround channel is often band-limited to 7 kHz before it is mixed.

Due to calibration problems with the analog tape machines used to replay the encoded two-channel Dolby Stereo signals, left and right audio tracks could apparently be more than 50 μs and 1 dB misaligned (Griesinger, 1989). The time misalignment means that frequencies with short periods could not be cancelled and summed in the correct way during decoding, which accounts for the 7 kHz low-pass filtering of the surround channel (Rumsey, 2001, pg. 98). To overcome this, a patent by Griesinger (1989) (which is used in the MC-1 Digital Controller decoder manufactured by Lexicon since 1999) describes a method to actively time-align and re-balance the input sig-

nals from the tape machine by looking for constant biases in the sum and difference from the magnitude of the input signals (it is assumed that if there is a strong dialog channel intended for the centre channel, the time-averaged difference in the magnitude of the two encoded input signals should be zero).

Discrete loudspeaker systems

Unlike the other classes of spatial audio systems described here, discrete audio systems differ in that they do not process the signals in any way except amplification before reproduction with loudspeakers (that is, there is no up-mixing: each of the n audio channels feeds each of the n loudspeakers with a one-to-one mapping). However, it is a particularly relevant category as it is the most common class of audio system; in nearly every car and home throughout the world. The most common example of such a system is what is referred in to in this thesis as the conventional two-loudspeaker reproduction method (2/0)- whereby the left and right electronic output of the musical recording from the storage media playback system (e.g. CD player) is simply fed to the left and right loudspeaker (via amplification). As mentioned, a *spatial audio system* is defined as a sound reproduction system where the spatial sound quality is improved compared with the conventional two-speaker (discrete) reproduction- so by this definition, a conventional 2/0 loudspeaker system is a discrete audio system but not a spatial audio system (nor is it conventionally called a multichannel audio system).

Of course, any number of discrete signal channels and loudspeakers could be used in a discrete loudspeaker system. An example of a large multichannel loudspeaker system is described by Woszczyk et al. (2005): A hemi-spherical arrangement of 24 drivers and 6 subwoofers reproduce sound from around

the top hemisphere of the listener to create an enveloping sound field. The independant channels reproducing artificial reverberation ensure that reverberance imagery is appreciated at a similar sound quality over a large listening area. The artificial reverberation is generated from the original mix (consisting of 12 channels, which could be either from a stored or live source) using electronic reverberation techniques such as convolution with a stored impulse response. However, such multichannel audio systems would benefit from a device which extracts reverberance-image sound components which are embedded in the original mix and reproduces these with the surrounding loudspeakers. Such *blind upmixing* devices will be described shortly.

The two commercial media formats for six channel storage of music (i.e. no video) are DVD-A (introduced in 2001, which uses the same PCM data encoding method as CD at a resolution of up to 24 bits sampled at up to 192 kHz) and SACD (introduced in 2003, which uses the DSD modulation technique with 1 bit samples at a sample rate of 2.8 MHz) (Pohlmann, 2000). The sixth channel is often intended to be reproduced with a sub-woofer,[11] but sometimes is a full-bandwidth channel which is to be reproduced with a conventional loudspeaker elevated above the other five. In fact, the German classical music label MDG has a system called "2+2+2 multichannel sound" which recomends that in addition to the front and rear loudspeaker pairs, the fifth and sixth channel from the DVDA are reproduced with another loudspeaker pair directly above the front. In terms of sound quality, DVD-A and SACD systems can be considered identical[12]– however, these two formats

[11]According to ITU-R BS 775-1 (1994), a subwoofer is a dedicated loudspeaker for reproducing frequencies in the range 20–120 Hz, which is also called the Low Frequency Effects channel.

[12]In a double-blind study with 110 people comparing 50 kHz bandwidth recordings processed with 24-bit, 176.4 kHz fs PCM or DSD converters and reproduced with headphones and loudspeakers, there were only four instances out of 145 test presentations that showed people could tell the difference between the PCM (DVDA) or DSD (SACD) systems Blech and Yang (2004).

may not be sold for much longer due to falling sales; according to a RIAA report (RIAA, 2004), in 2004 350,000 DVD-A discs were sold, down by 21% from the previous year, and in the same year 790,000 SACD discs were sold, down by 40% from 2003, compared with over 750 million CD's, up 3% from 2003.[13] However, DVD video and music videos together created about 60 million individual sales in 2004 (RIAA, 2004) and most of these discs contained five descrete audio channels (plus a low-frequency effects channel). This is called the home-theatre market and is responsible for the increase in surround-sound loudspeaker systems in our homes and cars, which is why the upmixing system introduced in this thesis is so relevant today.

For commercially released recordings on any of these media formats (including formats with video) it is assumed that five of these audio channels will be reproduced by five loudspeakers arranged according to (or at least very similar to) the ITU-R BS 775-1 (1994) recommendation. The ITU document describes how three front and two rear (or "surround") loudspeakers should be arranged for reproduction of sound with or without accompanying picture for a film (i.e. the loudspeaker arrangement is the same whether the screen is there or not). Notation relating to the number of front nF and rear nR loudspeakers is given as nF/nR or $nF - nR$. For example, the Hamasaki et al. (2004) system had a 12/10 and 5/10 loudspeaker configuration. The most common arrangements for home-theatre are 3/2 and 2/2. 3/2 means there are three front and two rear loudspeakers, as shown in figure 2.10, which is also called the "5.1" system (the point-one is a sub-woofer). 2/2 means there are only the front left and front right loudspeakers and two rear (i.e. there is no centre loudspeaker)- as the new spatial audio system introduced later in this thesis is. The abbreviations for the Left-Surround and Right-Surround loudspeakers are LS and RS, whilst the front Left, Centre and Right are sim-

[13]These figures do not include on-line sales.

ply L, C and R (respectively). Even though suggestions by ITU-R BS 775-1 for loudspeaker placement with a 3/4 configuration are given, we will just look at the 3/2 configuration with no sub-woofer as this is the most common format for musical recordings. No subwoofer is used in any experiments in this thesis as high-quality, "reference monitor" loudspeakers are assumed in accordance with ITU-R BS 1116 (1994).[14]

Imagery in discrete spatial audio systems

Phantom images created by radiation of a signal between front loudspeakers are much more *defined* and stable than if the same signal is reproduced between side loudspeakers (e.g. listeners report the image direction more consistently, with a smaller width and report that the image does not move as much; Ratliff, 1974; Theile and Plenge, 1977; Segar and Rumsey, 2001; Usher and Woszczyk, 2005). Therefore, the ITU loudspeaker arrangement is well suited for music reproduction which has an inherent "front stage", such as classical or jazz concert-hall recordings where the audience is expected to be in front of the musicians (these are the kinds of recordings that the new audio system in this thesis is investigated with). This is not to say that phantom

[14]Quoting from ITU-R BS 1116 (1994): *"Reference monitor" loudspeaker means high-quality studio listening equipment, comprising an integrated unit of loudspeaker systems in specifically dimensioned housing, combined with special equalization, high-quality power amplifiers and appropriate crossover networks. The electro-acoustic characteristics should fulfil the following minimum requirements, measured under free field conditions. Absolute sound level values are referenced to a measurement distance of 1 m to the acoustic centre, unless otherwise specified. For the pre-selection of loudspeakers, the frequency response curve over the range 40 Hz-16 kHz, measured in one-third octave bands using pink noise on the main axis (directional angle = 0°), should preferably fall within a tolerance band of 4 dB. Frequency response curves measured at directional angles ±10° should not differ from the main axis frequency response by more than 3 dB, and at directional angles ±30° (in the horizontal plane only) by more than 4 dB. The frequency response of different loudspeakers should be matched. The differences should preferably not exceed the value of 1.0 dB in the frequency range of at least 250 Hz to 2 kHz.* An example of the frequency response of loudspeakers used in experiments in this thesis can be found in figure B.2 in the appendix.

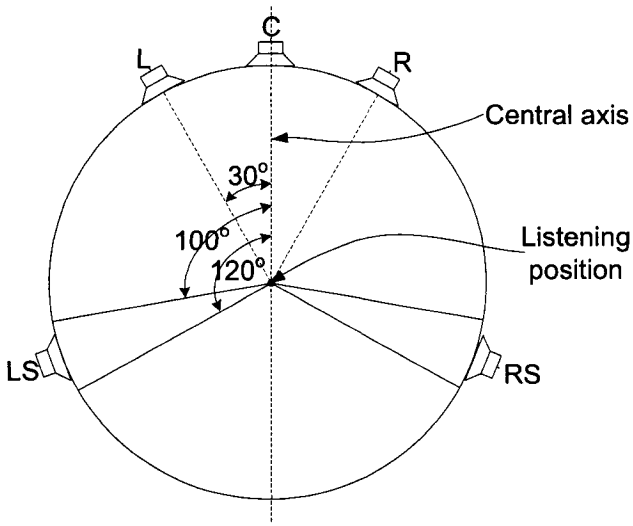


Figure 2.10: ITU-R BS 775-1 (1994) recommended 3/2 loudspeaker configuration for three front (L, C and R) and two rear loudspeakers (LS and RS). All loudspeakers should be at a similar height (approximately 1.2 m) and symmetrical about the central listening axis (with the rear loudspeakers between $\pm 100°$ and $\pm 120°$ relative to the central axis). For the 2/2 configuration- like the new upmix system introduced in the thesis- there is no centre loudspeaker. ITU-R BS 1116 (1994) recommends that the distance between the listening position (also called the "sweet-spot") and the loudspeakers be 2-3 metres.

imaging is impossible between side loudspeakers arranged in this manner- an experiment reported in chapter 3 shows that this is indeed possible for both source and revererberance images. Therefore most multichannel recordings for classical music are mixed so that the *source* images apear between the front loudspeakers and the rear loudspeaker channels are used for radiating sound which affects *reverberance* imagery. This is generally achieved using one or a combination of the following ways:

- By placing the microphones mixed to the rear loudspeakers far away from the musical instrument (Mason and Rumsey, 1999; Theile, 2000).
- Facing the rear microphones away from the instruments (Fukada et al., 1997).
- Decorrelating the original signals using an artificial reverberator (and adding a "pre-delay"; Corey, 2002).

As discussed in section 2.1.8, a high interaural cross-correlation is needed for the formation of a source image and a low interaural cross-correlation would be measured in a reverberant field (Damaske, 1967; Tohyama and Suzuki, 1989; Morimoto, 1993). The suggested locations of the rear loudspeakers in the ITU standard help to achieve this (i.e. better than if the rear loudspeakers were at different positions). Hiyama et al. (2002) found that using only five loudspeakers arranged at the 3/2 locations shown in figure 2.10 with rear loudspeakers at $\pm 120°$, the same sense of perceived subjective diffuseness could be achieved as if there were 24 loudspeakers at equally spaced angles around the listener. In a similar study, Ohgushi et al. (1987) found that for various reproduced music excerpts the listeners' judgement of "sense of reality" were identical whether the rear loudspeakers were located at $\pm 90°$

or ±120°. Also, (as mentioned) Sonke (2000) found that simulating only four or five plane waves at equally spaced angles of incidence with a WFS system was generally indistinguishable from twelve plane waves. So it seems that the five loudspeakers arranged according to the 3/2 ITU 775 format should be adequate to give a convincing *homogeneous* reverberance image; "*one in which no direction is preferentially treated*" (Malham, 1999). An experiment reported in chapter 3 shows that this is not the case for musical recordings reproduced with the 3/2 setup and that even when the listener can't see any loudspeakers, the reverberance images are generally biased in the direction of the rear loudspeakers.

To further reduce the interaural correlation of sound components radiated by the rear loudspeakers, Holman (1991a) suggests using *dipole* radiators (with the null pointing at the listeners' head) rather than *direct* radiating rear loudspeakers.[15] A system suggested by Fosgate (1993) has two separate signals feeding each of the drivers in the dipole surround unit, which are derived from a seven-channel upmixer (a derivation of Dolby Pro Logic II-described later).

The THX theatre standard recommends using rear dipole loudspeakers (Holman, 1991b). This standard is for multichannel cinema and home-theatre loudspeaker systems to ensure a similar listening experience of a film soundtrack in different rooms. With rear dipole speakers, the sound arriving at the listeners' ears from the rear loudspeakers will arrive more by reflected paths than via a direct line-of-sight path. Zacharov (1998) investigated the effect of rear loudspeaker directivity responses on three sound quality judgements; degree of envelopment, detail of directional effects and naturalness. For a

[15]A direct radiator here means one with a cardioid-like directivity pattern, such as with conventional moving-coil drivers in box cabinets and a dipole radiator a transducer with a figure-of-eight pattern, such as with a pair of direct radiators mounted back-to-back.

variety of musical and non-musical stimuli, rear-loudspeakers with higher directivity (i.e. direct radiators rather than dipole radiators) were rated more favourably for all three judgements. However, there seemed to be an effect of the listening room in this investigation, as off-sweet-spot listening positions were preferred over sweet-spot positions, which is suggested as a result of a null in a low-frequency room-mode at the sweet-spot. So maybe the test results for comparing the dipole and direct radiators would be different in other rooms due to the different loudspeaker-room interaction (the generalizability of the conclusions of the study were questioned by Holman, 2000b). Rumsey (2001)(pg. 141) comments that conventional (rather than dipole) rear loudspeakers are probably preferable (i.e. give higher sound quality ratings) if the spatial audio system reproduces multichannel sound mixes designed for panning to the side of the listener.

A modification to the dipole radiation method for rear loudspeaker signals is to radiate only that part of the sound which contains information about the spatial properties of the source (S) image with one loudspeaker, and those parts of the sound which contribute to the reverberance (R) image with another. We will call these two sound components the S and R image components. The so-called "Perceptual Sound Field Reconstruction" (PSFR) system by Rosen and Johnston (2001) radiates the R image component with a conventional loudspeaker facing away from the listener but facing a sound diffusion surface like a quadratic-residue diffusor (an "indirect radiator"). The source-image components are reproduced using a conventional "direct" loudspeaker facing the listener. Five units surround the listener, each with a direct and indirect radiator and a five-channel recording is processed to create ten signals (i.e. five signals for the direct radiators and five for the indirect radiators). The input signal feeds the direct radiator whenever a transient is detected (it is suggested that the switch was triggered by a local

maxima about three times larger than the energy magnitude averaged in the last 10 ms) and is fed to the indirect radiator the rest of the time.

2.2.3 Blind adaptive audio upmixers

This is really another class of spatial audio system, but as it is the one which best applies to the central topic of the thesis, it is described here in a separate section. "Blind" (or "unsupervised") upmixers differ from linear matrix converters in two important ways:

- Blind audio upmixing systems are *active*: the data processing is dependant on the particular input signal properties and the input-to-output relationship will be different for different input signals. The scaling parameters for the input signals to derive the new signals (via addition and subtraction) are adaptive and dependant on the particular audio input signals.
- With blind upmixers, the input signals do not have to be specially encoded.

The systems discussed here are intended to be used with conventional "off-the-shelf" two channel recordings and it is assumed the sound engineer who mixed the CD intended it to be played back with a conventional two-loudspeaker arrangement (i.e. a discrete 2/0 loudspeaker system). The word "blind" is often used with signal processing techniques where nothing is assumed by the system *a priori* about the signals themselves, in contrast to linear upmix systems which often assume encoded signals. Although there are many commercial audio upmixing systems (e.g. Circle Surround II, NEO6

and ARKYMES), they often use similar signal processing structures and five systems are described here to represent a variety of upmixing approaches in enough detail to show a contrasting feature.

Dolby Pro Logic II

Dolby Pro Logic II (DPLII) was brought to the Canadian market just before the author started his PhD program, also in Montreal (at the Festival du Son et de l'Image), in March 2001.[16] In a press release in April 2004[17], Dolby announced that DPLII was used in 15 million products for a variety of applications; car audio, video games, television (including live sports broadcasts), and the most common application; for home theatre. According to Gundry (2001), DPLII was designed to deal with unencoded two-channel signals, replacing Pro Logic (I)[18] which was designed principally for use with encoded signals (according to Waller (1994), Pro Logic (I) was a Dolby Stereo system adapted for home use) and had many limitations, such as a single surround channel limited to 7 kHz (Dressler, 2000).

The idea is summarized in figure 2.11, which is based on details described in the Fosgate (2005) patent (filed in March 2000) and the conference paper by Gundry (2001).[19] A key feature of DPLII (which is also shared by the new upmix system is introduced later) is the use of a feedback-derived active control system. The active control system (called a *servo*) attempts to match

[16]Also introduced in 2003 was Dolby Pro Logic IIx, which upmixes 2 to 7 audio channels (reproduced as a 3/4 loudspeaker system)- with a flanking pair of "centre-side" loudspeakers, though this will not be discussed as it uses the same basic idea of DPLII.

[17]Online article: `http://investor.dolby.com/ReleaseDetail.cfm?ReleaseID=155740`.

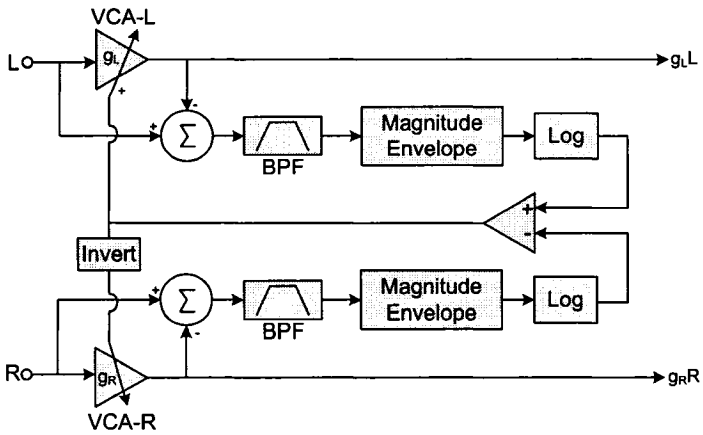
[18]Which is generally just called Dolby Pro Logic.

[19]Not all details of the DPLII system are discussed in these documents. For example, it is unclear how the surround channel is decorrelated to create a pair of surround loudspeaker channels. In fact, Dolby do never explicitly say which patents are in DPLII units.

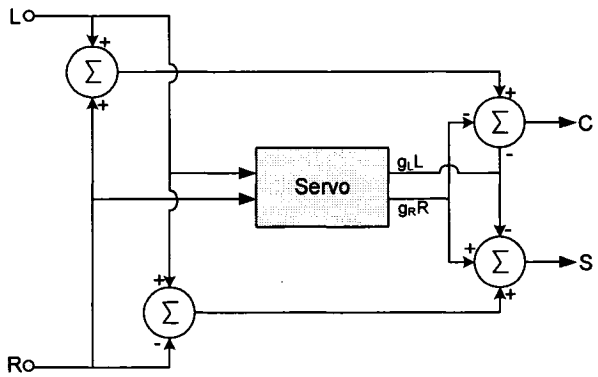
the input signal magnitudes so as to cancel the effect of an amplitude-panned source which is more dominant (i.e. with a higher energy contribution) in one channel than the other. The signal level is determined in a very similar way to other upmix systems which also use the left-to-right channel level ratio as a means for determining the dominant signal "direction" (e.g. the Logic7 system (Griesinger, 1998) and the system proposed by Choi et al., 1995). The level estimate is based on a band-limited weighting of each servo input signal (to remove the possible confusion with strong low-frequency sound effects being treated as the dominant signal rather than speech) and the magnitude is estimated using an envelope smoothing function with a fast response time and slower decay time, with time constants in the order of a few milliseconds and close to one second, respectively (Gundry, 2001).

The difference in level between these two signals is calculated using a simple comparator (e.g. an op-amp) which gives the difference between the two input voltages (i.e. left-channel minus right-channel). If the output of the comparator is positive, then the left channel is considered to be dominant and the left input channel is *attenuated* using the VCA on the left channel (VCA-L in figure 2.11(a)) whilst the right input channel is *boosted*. Likewise, if the comparator output is negative then the left channel is boosted whilst the right input channel is attenuated. An important point is that the magnitude estimate be converted to a logarithmic value to take into account the relative absolute voltage between the two signals (i.e. converted to a *level*) before the difference is calculated with the comparator. If the magnitude was calculated using the absolute magnitude value- in Volts- then the gain of the VCA would be dependent on the voltage level. For instance, if the left and right input signal magnitudes were 1.0 V and 1.01 V, then VCA-L would multiply the left input channel by the same amount as if the two input magnitudes were 0.01 V and 0.02 V, even though in the first case the level difference between

the signals was 0.1 dB but in the second case it was 6 dB.



(a) Servo ("steering") system. The aim is to "cancel out" the effect of amplitude panning of the dominant channel. This is accomplished using the Band Pass Filtered (BPF) signals from each channel and the comparator tries to balance the (smoothed) level of each input signals using the variable-gain amplifiers (i.e. VCA's).



(b) Derivation of Centre and Surround channels using output of servo. To reduce perceptual fusion of the front and rear images a 0-30 ms delay of the surround channels is implemented.

Figure 2.11: Functional overview of Dolby Pro Logic II (Gundry, 2001; Fosgate, 2005). Front left and right loudspeaker channels are created in the same way as the C and S channels except the servo inputs are the C and S channels. Although the gain values calculated by the servo (i.e. the gain of each VCA) are calculated in real-time, the input signals are delayed by 5 ms as a look-ahead digital implementation for better tracking performance. The surround channel is further processed to create an approximately uncorrelated pair of surround signals.

Maher "Stereophonic image enhancement" system

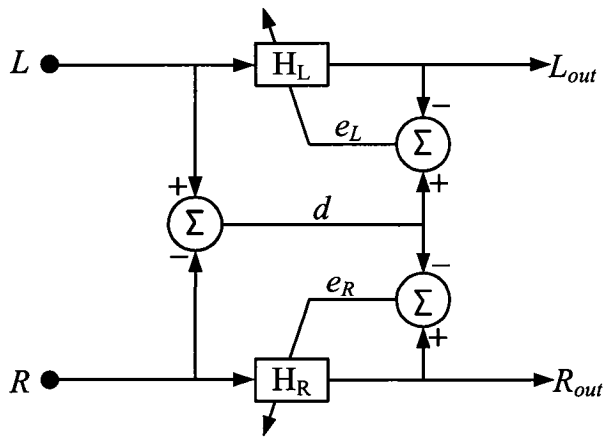


Figure 2.12: Maher "spatial enhancement" system (Maher et al., 1996). The two unencoded input signals are filtered by the two adaptive (FIR) filters H_L and H_R, which approximates the difference signal d, to give two new output signals L_{out} and R_{out}. The output signals are then radiated from loudspeakers behind the listener.

The blind upmix system discussed by Maher et al. (1996) is ostensibly a mixture of the Dolby Pro Logic II system and the new upmixer introduced in this thesis. However, the motivation for the signal processing configuration shown in figure 2.12 is inherently different. If the adaptive filter H_L was a simple gain function, with a gain controlled by the magnitude of the signal e_L, then it would be similar to the servo of the DPLII system (though the Maher system does not suggest any further signal processing, contrary to DPLII as shown in figure 2.11(b)). The two adaptive filters are a 12-tap FIR filter, which is updated so as to minimize the level of the error signals e_L and e_R (updated according to the LMS algorithm, which is described in depth in chapter 4). If the error signals (e_L and e_R) were taken as the output signals, then the basic idea of this system would be much more in-line with that of the new upmixer in the thesis; which is to reduce the level of the overall

difference signal (there are some other major differences; for instance the new upmixer has mechanisms on the input signals to deal with time-delay and "hard" amplitude panning; which the Maher system does not have provisions for dealing with).

As there is no delay on the difference (d) signal, then the filtering must be undertaken with minimal latency to ensure that the correlated sound components can be canceled (i.e. by the second pair of differencing units). Therefore, the benefits of an increased frequency resolution and computational efficiency gained by using a large filter with frequency-domain convolution are traded against the increased IO latency and reduced efficacy of the system. Also, because the output signals are entirely from the filtered input signals, any filtering artifacts (e.g. for fast changing coefficients) would be more audible than if the output signals were calculated from the difference of the filtered signal and a non-filtered signal (as with the new system and the final stage of DPLII).

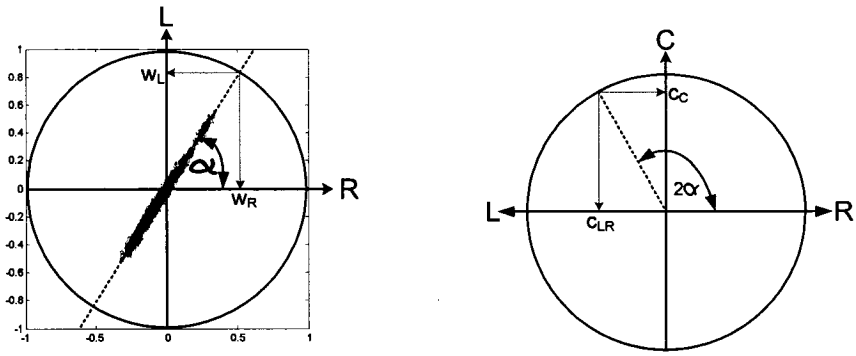
Aarts/ Irwan upmixer

This two-to-five channel upmixer (Aarts, 1995; Irwan and Aarts, 2002a,b) has three functional parts:

1. Find the principal image direction: how far off-centre the image is panned in the horizontal plane and determine whether this direction is in the front or rear.

2. Create a pair of signals for the centre-channel and rear (or surround) channels with a weighted sum and difference of the two input signals.

3. Decorrelate the rear-channel signal to create two new rear-channel signals.



(a) Lissajous phase plot of highly correlated noise showing relationship of image direction α and weights w_L and w_R.

(b) Three-channel representation showing how gains for summed speaker (c_c) and left and right speaker signals (c_{LR}) are derived.

Figure 2.13: Calculation of image direction for Irwan and Aarts system.

The output of the first part is a two-valued vector comprising of the average energy of the left channel, w_L, and the right, w_R. This is summarized in figure 2.13. As can be seen from the Lissajous phase plot, the average slope corresponds to the angle of this vector- i.e. at time k,

$$\alpha(k) = \text{atan}\frac{w_L(k)}{w_R(k)}. \tag{2.5}$$

The relative strength of the left and right channels is calculated using a single-tap LMS filter (though a this can also be done using a Principal Component Analysis technique described in Irwan and Aarts, 2002b- but the PCA approach is more computationally demanding). The LMS filter takes a weighted sum of the input signals (with the left channel weighted by w_L and the right by w_R) and tries to maximise the energy of this sum with respect

to these weights whilst keeping the norm of these weights equal to one, to preserve energy. The rear-loudspeaker (or surround) channel is created by a weighted difference of the input signal, but the weights are reversed (i.e. the left channel is weighted by w_R and *vice versa*). This is intuitively obvious: for example, if the left channel energy is 6 dB larger than the right channel, then we can only cancel the correlated components by first boosting the weaker right channel by 6 dB and then subtracting the two signals. The calculation of the front-loudspeaker signal $y(k)$ and surround signal $q(k)$ are summarized in (2.6).

$$\begin{aligned} y(k) &= w_L(k)L(k) + w_R(k)R(k), \\ q(k) &= w_R(k)L(k) - w_L(k)R(k). \end{aligned} \tag{2.6}$$

When the two input signals are unrelated, $\alpha(k)$ will be undefined. This can be checked against by ensuring that $\alpha(k)$ does not go negative. Also, when the difference signal $q(k)$ exceeds the sum signal $y(k)$, the front channel weights c_{LR} and c_C shown in figure 2.13(b) are weighted less using a weighting coefficient β. The new weighted values can be visualized by adding a third dimension orthogonal to the circular plane shown in figure 2.13(b). The image direction vector in this 3D representation would have a vertical angle (i.e. elevation) of β relative to the plane and the surround image weight c_s equal to the vertical height of this vector, as can be seen in figure 2.14. In practice, β is calculated using an estimate of the cross-correlation coefficient, (ρ at lag $\tau = 0$), which can be approximated using an accurate and computationally simple recursive technique described by Aarts et al. (2002). β is inversely

proportional to ρ and is bound between zero and $\pi/2$, as described by (2.7).

$$\beta(k) = \begin{cases} \text{asin}\left[1 - \rho(k)\right] & \text{if } 0 < \rho < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

The front channel weights c_{LR} and c_C are both multiplied by $\cos\beta$ to give c'_{LR} and c'_C, and the rear-channel weight c_S is simply $\sin\beta$- in other words, when the cross-correlation is zero β is equal to $\frac{\pi}{2}$, so the front channels are both weighted by zero ($\cos\frac{\pi}{2}$=0) and the weighted difference signal steered fully to the rear channels.

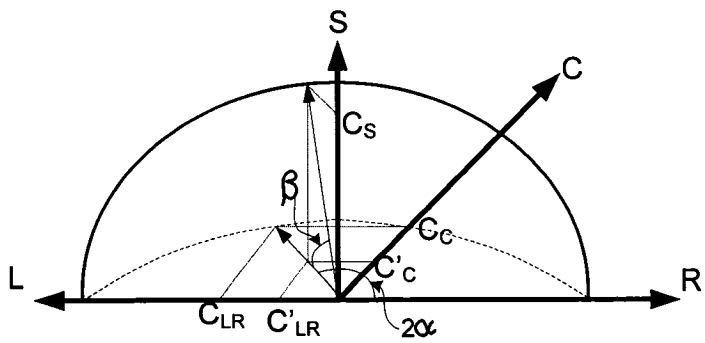


Figure 2.14: 3D representation of output signal derivation for Aarts/ Irwan upmixer. The principal direction (2α) is calculated using an iterative approximation of the gradient of the lissajous phase plot (figure 2.13(a)), and the scaler β is related to the degree of correlation between the input signals; when the input signals are uncorrelated it is high and the difference signal is then steered to the rear (surround) loudspeakers.

In summary, this entire process to generate the front loudspeaker signals u_L, u_R and u_C and the rear channel u_S can be represented with the following

adaptive matrix operation:

$$\begin{bmatrix} u_L(k) \\ u_R(k) \\ u_C(k) \\ u_S(k) \end{bmatrix} = \begin{bmatrix} c_L(k) & w_L(k) \\ c_R(k) & w_R(k) \\ c'_C(k) & 0 \\ 0 & c_S(k) \end{bmatrix} \begin{bmatrix} y(k) \\ q(k) \end{bmatrix}, \tag{2.8}$$

where:

$$c_L = \begin{cases} -c'_{LR}, & \text{If } c_{LR} < 0, \text{ i.e. image direction is left of centre,} \\ 0 & \text{otherwise.} \end{cases}$$

$$c_R = \begin{cases} c'_{LR}, & \text{If } c_{LR} \geq 0, \text{ i.e. image is right of centre,} \\ 0 & \text{otherwise.} \end{cases}$$

The final stage in this upmix process is to create a pair of independent rear loudspeaker signals from the single surround signal u_S. This is achieved by convolving the surround signal with a pair of orthogonal impulse responses in what is know as a *Lauridsen decorrelation filter* (Lauridsen, 1954)- also called an "all-pass" filter. These two signals are reproduced with a 10 ms delay relative to the front channels to reduce perceptual integration of the source and reverberance images. Each impulse response has just two non-zero coefficients, separated by a 10 ms gap. For both channels, the first coefficient is the same positive value (e.g. 0.5). For the second coefficient, the value is positive in one whilst negative in the other. This gives two orthogonal signals, but with a comb-filter magnitude response: there are dips in one response where there are peaks in the other (the peaks are 100 Hz apart with a "height" of about 25 dB at 100 Hz and about 15 dB at 1 kHz- analysed using a sliding 2048-point, Hanning windowed FFT analysis of white noise).

Therefore, the two outputs will sum to a flat magnitude response but only if the acoustic path from each speaker to the listener is identical. Informal listening off the sweet-spot reveals the harsh metallic timbral colouration of each channel and head movement causes a noticeable timbre change.

Logic7

The basic idea of this 2-to-7 channel upmixer (Griesinger, 1996a, 1998) is to determine a single dominant direction, which is calculated (like Dolby Pro Logic II) as a logarithmic level difference of the magnitude of the two input channels. The front-to-back energy ratio (like β in the Aarts system) is then calculated using the log-magnitude ratio of the sum and difference signals which is used to weight the contribution of the energy to the front and rear loudspeaker channels. The weighted difference channel (weighted by the front:back energy ratio) is then sent to the LS loudspeaker channel with a 180° phase-shift relative to the RS channel. The additional two rear-surround channels are simply time-delayed copies of the other surround channels (Griesinger, 1996a)- i.e. employing the Madsen effect.

According to the above three patent references, the idea of the Logic7 system is basically the same as the upmix system proposed by Choi et al. (1995), except the Choi system only has a single surround channel and the Logic7 system is detailed for an implementation using all analog components whereas the Choi system is for operation on a digital computer. The digital system allows some error-checking to help with computational overflow problems for extreme signal conditions. For instance, if the left channel magnitude is much larger than the right, then the sum:difference ratio calculation is not undertaken and is estimated at unity.

Avendano/ Jot upmixer

This 2-to-5 channel upmix system (Avendano and Jot, 2002, 2004) has two main aspects:

1. Extraction of frequency bands which are deemed uncorrelated, for radiation with a pair of rear loudspeakers - as shown in figure 2.15.

2. Redistribution of the two-channel signal into three font loudspeakers by determining a principal source direction and repanning the sound using vector-based panning for three channels (using the same idea as for the Gerzon "Trifield" system discussed earlier).

Chernyak and Dubrovsky (1968) found that when the interaural correlation of headphone-presented noise was zero, two images were heard yet only a single image was heard when the coherence was 0.4. Similarly, the Avendano/ Jot system considers all frequency components with an inter-channel coherence less than 0.15 to be uncorrelated and passes these frequencies to the rear loudspeakers (Avendano and Jot, 2002), as shown in figure 2.15.

To create the signals for the front three loudspeakers, the principal image direction is calculated on a frequency-by-frequency basis using the tangent panning law for amplitude panning. If the direction is to the left of the centre, then only the left and centre loudspeaker is used (and *vice versa* if the principal direction is on the right).

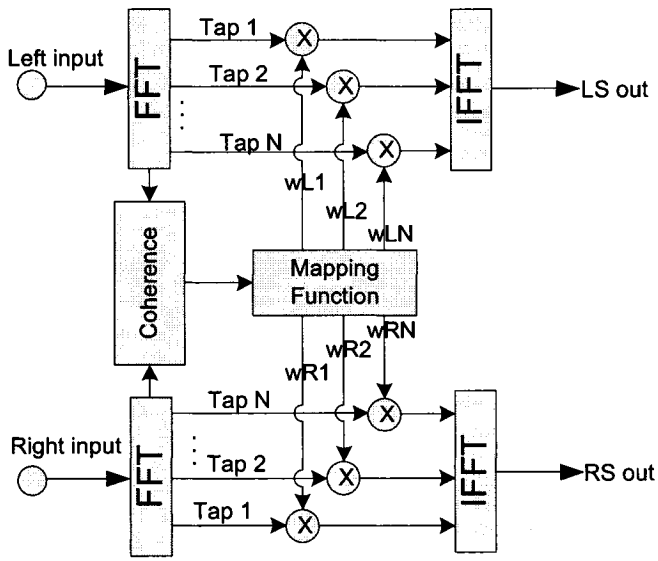


Figure 2.15: Functional schematic for generation of rear-loudspeaker signals LS and RS from two-channel input audio signals for Avendano/ Jot upmixer (adapted from; Avendano and Jot, 2002). 1024-point (I)FFT's are suggested with input data sectioned according to the overlap-add windowing technique. The mapping function returns a weight (wLk for the k^{th} weight on the left input channel) of close to zero if the interchannel coherence is above approximately 0.15 and unity if the below 0.15. So if a particular frequency band was deemed incoherent, it would be fed entirely to the rear loudspeakers (i.e. channels LS and RS).

Chapter 3

Subjective Design Criteria for the New Upmixer

In the introduction to the thesis, two general goals of the new upmixer were outlined:

1. To create a source image with a spatial quality similar to the original 2/0 mix (i.e. a high spatial *fidelity*).
2. To create a natural-sounding reverberance (ambiance) image.
3. To create a listening experience that would not be dispreferred to the original 2/0 listening experience.

It was also noted that the third goal is subservient to the first two, which highlights the general design principle that the new audio processor is *not* so much a special effect- which modifies the 2/0 listening experience with sound artifacts not necessarily present in the original recording- but rather

an audio *enhancer*, which by implication intensifies aspects already present in the original recording. In this chapter, these first two general goals are translated into specific subjective design criteria, and a method (i.e. listening test) for evaluating these criteria is also be detailed.

The "music intelligibility" or *readability* (Guastavino and Katz, 2004) of a reproduced sound scene can be explained as a source image being masked by a reverberance image, and an exploratory investigation is reported which looked at how spatial properties of a source image are affected by spatial properties of a reverberance image in a multichannel loudspeaker audio scene.

3.1 Design of a Graphical Mapping Tool for describing Auditory Spatial Imagery

3.1.1 Introduction

In this section a method to visualize the geometric properties of auditory imagery perceived by a listener in an audio scene is presented. This mapping tool is designed specifically for evaluating imagery in a loudspeaker audio system arranged according to the 2/0, 2/2 or 3/2 ITU-R BS 775-1 (1994) configurations. The new tool is a graphical mapping system, as described in the discussion on methods for investigating auditory spatial imagery (page 34).This tool is called a GUI (Graphical User Interface) as it is a word commonly used for software which has a strong graphical component to the user interface.

In order to allow a comparison of auditory spatial imagery in the original

Attribute	Dimensions
Image centre	Direction (azimuth) Distance (ego-centric range)
Image size	Width Depth

Table 3.1: Summary of image properties the GUI should measure, as graphically shown in figure 2.1.

2/0 and upmixed scenes, the GUI must be able to describe the perceived spatial extent and location of a source and reverberance image and also the *definition* of the source image.[1] Secondly, the GUI must provide data which can be used to consistently reveal differences in the perception of auditory images in different sound scenes. Data for these analyses come from the geometrical image properties shown in figure 2.1, as summarized in table 3.1. This must be available in a way which allows a statistical interpretation that gives an indication as to how reliably the GUI can be used to map perceived auditory images.

3.1.2 Design of the GUI

The GUI was programmed using MATLAB (with over 3000 lines of code!) and ran on either PC1 or PC2 (see appendix C). A screenshot of the GUI is shown in figure 3.1. The GUI allowed a listener to describe the Source (S) image in terms its definition (either as being a stable or unstable image) and the Reverberance (R) image on a 2D plan-view map showing the listening room. All 2D image attributes could be described simultaneously using any number of ellipses. For each scene drawing, ellipses of the same type (e.g.

[1]Definition here means the spatial image definition as described in section 2.1.5, such as its temporal stability.

stable S image) which overlapped were "flattened" and treated as the same image, so multiple ellipse could be drawn to represent a single image.

Images were drawn in a similar way as with the graphics programs Adobe Photoshop or Microsoft Paint; the image select button was clicked (one of the three buttons in the top-left of the screen) and the subject left-clicked the mouse on part of the screen to start drawing the ellipse from the edge, dragging the mouse to increase the ellipse size. When the left-mouse button was unclicked, the ellipse appeared as a solid object the same colour as the button (red for stable S image; pink for unstable S image; blue for R image).

Once drawn, the ellipse could then be repositioned by clicking and dragging, and could be resized or deleted using the GUI buttons below the three ellipse select buttons shown in figure 3.1. The top-down view shows the listening position at the centre and the two curtains in the listening room (at a distance of 0.9 m and 1.8 m from the listening position). The numbered markers correspond to numbered markers on the inner curtain and the radiating lines (at 10° intervals) are used to help the listener map the image distance (i.e. ego-centric range; the perceived distance of the auditory image from the listener). The view could be zoomed-out to an infinite distance and concentric markers at 10 metre intervals are shown. To wipe the screen, the re-start button is hit and when the listener has finished they hit the "next scene" button (a window then appears asking the user to confirm their action).

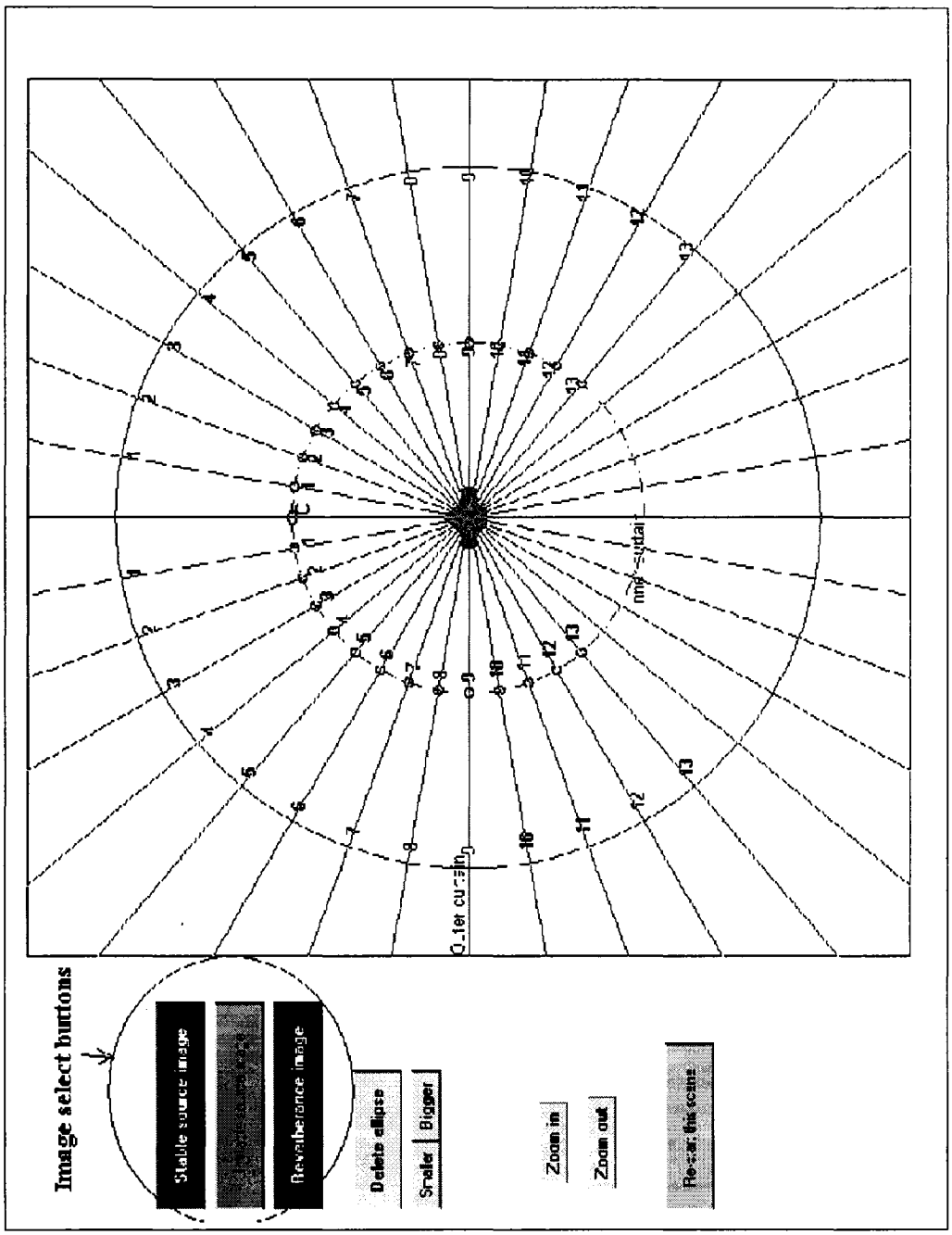


Figure 3.1: Screenshot of GUI.

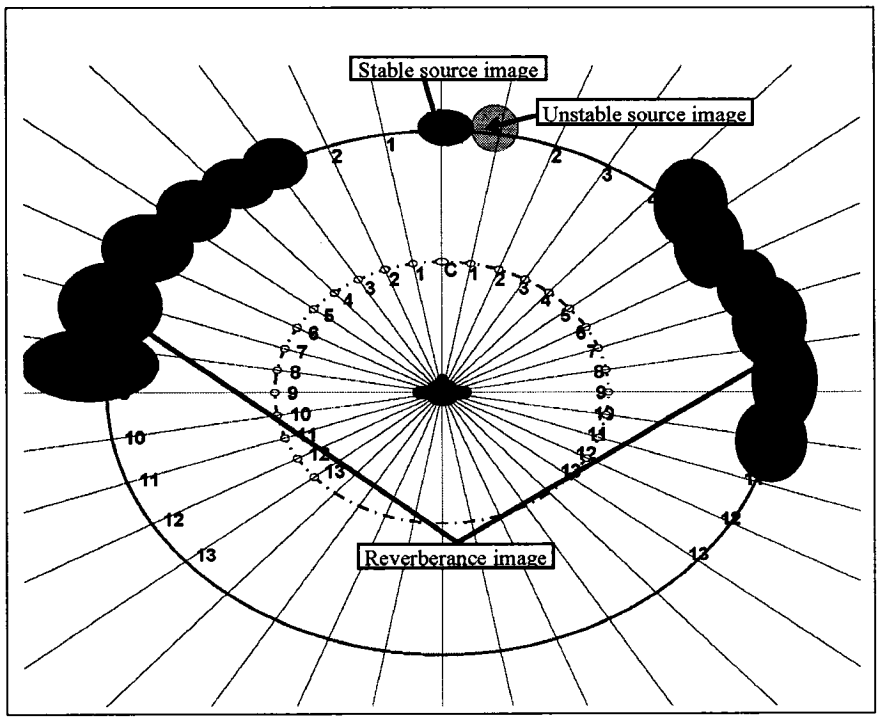


Figure 3.2: Example response for the GUI showing elicited source images (stable and unstable) and reverberance images. An interpretation of the graphical response shown here is: A stable source image was heard between -3° and +5°. The right-hand side of the source image (i.e. from 5° to 13°) was spatially unstable; it may be that certain transients in the sound pulled the source image to the right, or that when the listener rotated their head the source image "jumped" to the right. An enveloping reverberance image was heard, but not in the direction of the source image. The multiple ellipses used to create the left and right-hand side R images are treated as just two separate reverberance images. How this graphical representation is analysed (e.g. to calculate image width, direction and distance) is explained in figure 3.5.

Saving the data.

For each scene description, a text file was saved which contained a numerical matrix similar to this example:

```
1 103.26 139.49 1.71 1.42 3
2 102.23 139.43 4.8 2.42 2
3 102.29 131.65 5.42 9.65 1
```

Each row corresponds to a geometry description for each ellipse drawn. Taking the first row as an example (the value is shown in brackets), the general format follows:

First column: Ellipse number (1).
Second column: x-axis ellipse centre (103.26).
Third column: y-axis ellipse centre (139.49).
Forth column: ellipse width (1.71).
Fifth column: ellipse height (1.42).
Sixth column: ellipse definition code (3).
For the ellipse definition code, 1 is for a stable source image; 2 is an unstable source image; and 3 is a reverberance image. The units for the ellipse centre, width and height are as a fraction of the overall screen width.

3.1.3 Analysis of data from the GUI

The data file can be used to recreate the elicited ellipses for each image type (e.g. for only the stable S image). To extract the image properties in table 3.1, the recreated scene is analysed using the method summarized in figure 3.5; this gives image width, azimuth and distance values for each sound scene which can then be analysed using conventional statistical methods such

as ANOVA. A method to visualize all the responses for a particular sound scene is using density plots and image directional strength. It should be noted that the density plots (which contain the results of multiple scene representations) are not used to calculate image properties; they are just used as an exploratory method to visualize imagery in the sound scenes.

Density plots

Density plots enable all scene descriptions of a unique audio stimulus, as drawn by different subjects and by the same subject for different runs, to be visualized on a single pair of axes. Wagener (1971) accomplished this by layering the scene drawings, which were drawn on see-through plastic, on top of each other and photographing the net result. If the original drawing is done using a computer, this layering can be done computationally (Mason, 2002; Usher and Woszczyk, 2003; Ford, 2005) allowing for a separate analysis of the elicited source and reverberance images. The result is a *density plot*, which shows regions where an image was reported more frequently as being darker.

The process of creating a density plot is now described. For n different scene configurations (e.g. loudspeaker arrangements) and F fragments (e.g. recordings of different instruments) we have $n * F$ unique audio scenes. This is described by S subjects with R runs, giving $S * R$ unique representation of each audio scene. The density plot for each audio scene is created with this three-step procedure:

1. For each image category (e.g. stable S image), each of the $S \times R$ unique

scene descriptions are recreated using the saved image geometry text file.

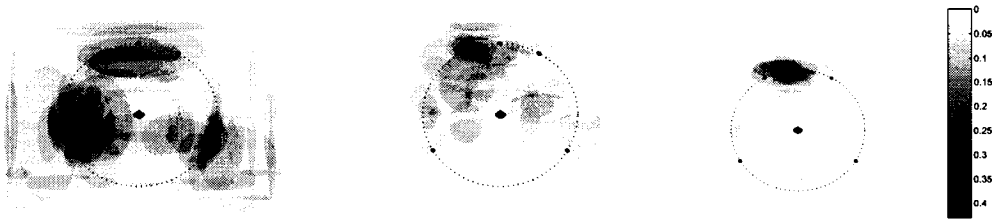
2. Each description is then quantized onto a grid of approximately 1-by-1 "real-world" centimetre pixels. This resolution is better than the auditory system; which has a localization accuracy for broadband sources reproduced with loudspeakers around the listener of 3°-10° (Blauert, 1997, pg. 41), which is 5-17 cm at a 1 metre range. The ellipses for each scene are quantized so that the maximum value (i.e. density) of each pixel in the scene is one (i.e. if an image was reported at a particular location for every stimulus presentation, then that grid unit would have a density of one).

3. Each scene is then summed with other responses of the same image category for the same audio scene. The maximum value for each unit of the density plot would be $S \times R$; which would occur if an image was elicited at that particular location of the GUI-map for every stimulus presentation.

Ideally, the edge of the ellipses would be smoothed (tapered) so as to reflect a more convincing representation of perceptual imagery, where auditory images do not categorically stop but rather "fade-out" over a region of space. This spatial smoothing could be accomplished by convolving each elicited map with a two-dimensional Gabor function (i.e. a 2D sinc function) (Marr and Hildreth, 1980).

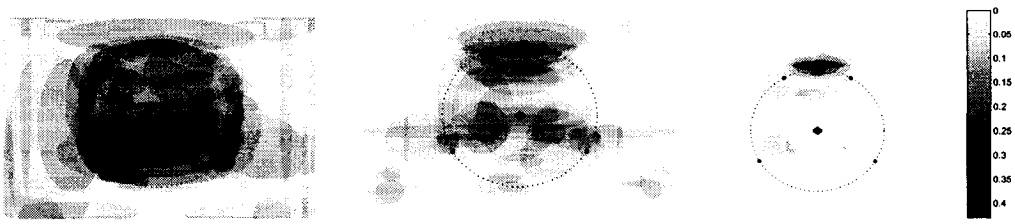
An example of a density plot from an earlier experiment (Usher and Woszczyk, 2004) is shown in figures 3.3 and 3.4. The two sound scenes were created using two five-channel commercially available recordings (released on

SACD) of a piano[2] and church organ[3] performance in a (different) concert hall and reproduced with a conventional ITU-R BS 775-1 3/2 loudspeaker configuration, as shown in figure 2.10. The density plots are shown for both elicited stable and unstable source images and for elicited reverberance images.



(a) Reverberance image (b) Unstable S image (c) Stable S image

Figure 3.3: Density plots arranged by image-type: Reverberance (R) image; stable and unstable source (S) image. The stimulus was a commercially released five-channel recording of a solo piano, reproduced with an SACD player and five loudspeakers arranged to the ITU-R BS 775-1 3/2 configuration; the location of the loudspeakers are shown with black dots. In the listening tests, the loudspeakers were hidden from the listener with a visually opaque yet acoustically transparent curtain. The density scale corresponds to the number of times an image was reported at a particular location using the GUI (which is expressed as a fraction with 1=100%). Each density plot is created from the data of six listeners and three presentations; i.e. 18 unique scene responses.



(a) Reverberance image (b) Unstable S image (c) Stable S image

Figure 3.4: As figure 3.3 but with a church organ recording.

Density plots can be analysed to describe how consistently the same person represents the same audio scene using the GUI or how consistently differ-

[2]Franz Schubert. *Ave Maria* D. 839. SACD. PentaTone classics 5186 043, 2003.
[3]Henri Tomasi. *Semaine Sainte à Cuzco*. BIS-SACD-1109, 2000.

ent people represent the same audio scene. This consistency can be calculated with a measure devised by Mason et al. (2001) called the "similarity statistic" (S). This is a single value between 1 and 0 which is proportional to the total percentage of over-lapped responses for that sound scene calculated according to (3.1):

$$S = \frac{\sum_{d=1}^{N}(d-1)A_d}{(N-1)\sum_{d=1}^{N}A_d}, \quad \text{where:} \tag{3.1}$$

A_d = area of response with a density level of d.

N = number of summed response sheets.

In an earlier experiment (the time-panning experiment summarized in figure 2.3), we found the within-subject consistency to be remarkably high; for one subject it was S=0.834 from four repeated presentations of an anechoic recording of a solo bass (Usher and Woszczyk, 2003). The between-subject similarity was typically much higher than that found in a similar spatial audio scene mapping experiment using a loudspeaker pair and a phase-modulated noise source (Mason, 2002, chapter 5) for which S ranged from 0.018 to 0.060 (mean value 0.041). However, this statistic was not used in any further experiments as it is easier to interpret established auditory image descriptions (such as width, distance and azimuth) using conventional statistical methods such as analyses of variance tests. Furthermore, the analysis technique we will use by extracting those image attributes summarized in table 3.1 will enable a comparison of the data from the GUI with data from previous studies.

Image directional strength

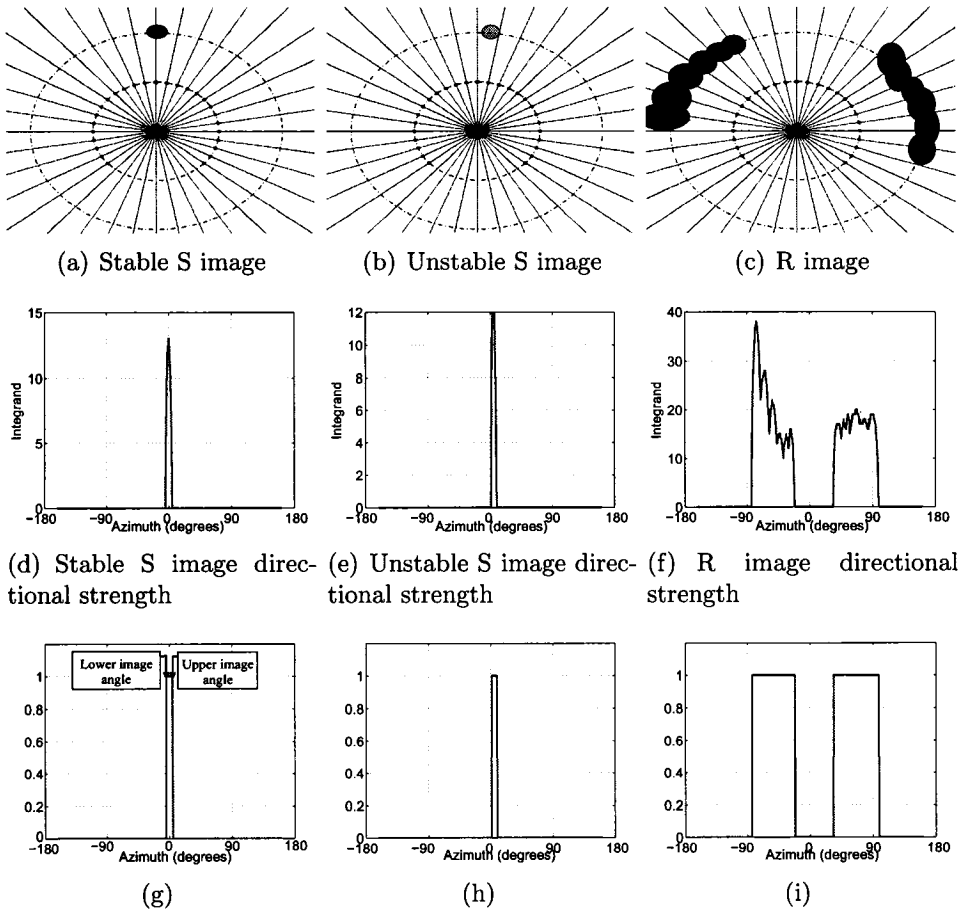
The image directional strength is a measure of the frequency of occurrence which listeners report auditory images from a particular *direction* (originally introduced by Usher and Woszczyk, 2003). It can immediately visualize the distribution of imagery in the horizontal plane, ignoring the distance (ego-centric range) dimension of the density plot. This is advantageous as it has been shown that listeners report perceived image distance (i.e. ego-centric image range) inconsistently with conventional loudspeaker audio scenes (Usher and Woszczyk, 2003, 2004; though in an experiment with a wave field synthesis system, the reported image distance/ range using the GUI was as consistent as reported image distance using a verbal description; Usher et al., 2004b). Also, Neher (2004) found that image depth was reported in a consistent manner only with specially tailored stimuli- as discussed in his experiment called "validation of ensemble depth".

In previous experiments with the GUI, we found that reported image width and depth were strongly positively correlated and informal questioning of the listeners revealed that they were generally very unsure about reporting image depth. This is expected, as in real world listening situations the auditory system is constructed in a way to determine spatial image properties of the sound source in the horizontal dimension, using interaural cues, much better than in the distance dimension (where the use of different cues depends strongly on the sound source and direction; Zahorik, 2002). This is supported by sound localization studies which show a much greater accuracy for reported source azimuth (as mentioned; 3°-10° for broadband sources around a listener, or 26-88 cm for a source 5 m away; Blauert, 1997, pg. 41) than for distance: In blind localization studies in echoic environment, we have a consistent bias to overestimate the distance of close source (closer

than about 1.2 m) and underestimate the distance of far sources (Nielsen, 1993; Zahorik, 2002)- for instance, a real sound source 5 m away was generally reported between 3 and 4 m (Zahorik, 2002). The method for calculation of image directional strength is summarized in figure 3.5.

Calculation of image attributes

To calculate the image attributes of table 3.1 (image width, azimuth and distance), the procedure described in figure 3.5 is undertaken.



(a) Stable S image (b) Unstable S image (c) R image

(d) Stable S image directional strength (e) Unstable S image directional strength (f) R image directional strength

(g) (h) (i)

Figure 3.5: Analysis of scene descriptions using the GUI for calculating image width and azimuth for a single scene description by one listener. Using the example scene response in figure 3.2, the image properties are calculated in the procedure summarized above for each of the image categories: Stable Source image (left column); unstable source image (centre column) and reverberance images (right column). The "raw" image description for each category is shown in the top row. The image directional strength plots (i.e. middle row) are then truncated so the maximum is one- as shown in plots (bottom row). The image width is calculated by the absolute difference between the maximum and minimum angles covered by each image. This accounts for "wrap around"- i.e. a polar coordinate system is used. The image direction is calculated as the centre of each unique image; that is, the half-way point between the upper angle and lower angle in subplot (g). The distance is calculated as an average of the ego-centric distance (i.e. range) to the centre of each ellipse which makes up a single image. In the lower plots (g-i), the directional strength is quantized to 1 if it is greater than 0. If the lower three plots are summed with other responses for the same stimuli and normalized by the total number of responses, then we have the image directional strength (i.e. a measure of the frequency which an image of a particular type was reported as a function of azimuth).

3.2 Interaction of Source and Reverberance Imagery in a Multichannel Loudspeaker Audio System

3.2.1 Purpose of investigation

An exploratory study is reported here, where two areas of auditory spatial imagery in loudspeaker audio scenes were investigated: The generalizability of, and the interaction between, source (S) and reverberance (R) images. This study was undertaken to gain an insight into the consequences of radiating the reverberance image components from loudspeakers around the listener whilst maintaining a frontal source image, as the new upmixer is designed to achieve (in-keeping with the goal of minimizing distortion of source imagery in the upmixed scene compared with the original 2/0 scene).

Nearly all of the literature on the geometry of auditory spatial imagery in loudspeaker audio scenes has concentrated on *source* imagery; two exceptions are the work of Hanyu and Sekiguchi (2004) and Ford (2005). Likewise, the control of source image direction for loudspeaker audio is well researched (as discussed in the previous section on "panning").

To summarize; in this study there were two aspects of the degree of generalizability between source and reverberance images which were investigated: (1) It was wondered whether the spatial geometry of a reverberance image could be described with a similar degree of reliability (i.e. consistency) as with a source image.

(2) The degree to which S and R images can be independently controlled using pair-wise amplitude panning with loudspeakers around the listener.

These research topics were investigated for both *real* and *virtual* images. The difference between a real and a virtual (or phantom) image is that a real image is perceived to exist at the same location as the sound-creating loudspeaker(s) whereas a virtual image is perceived to exist at a position generally between them (Pulkki and Hirvonen, 2005; Usher and Woszczyk, 2005); both real and virtual images commonly occur in loudspeaker music reproduction. The large body of data on virtual-image localization around a listener for source images can then be compared with the data obtained from the GUI for R-image spatial properties to quantitatively provide data for the second question regarding panning reverberance images.

In order to generate both a source image and a reverberance image, two different signals were reproduced. For the formation of an S image, an anechoic recording of a solo musical performance was reproduced. Since reverberance is, by definition, the perceptual correlate of reverberation, then a reverberance image will be perceived if we reproduce a recording of the anechoic signal processed by a reverberant path impulse response; that is, one where the coefficients have an exponentially decaying time envelope of random noise (this definition is explored in more detail in the next chapter); how this was achieved is described in the *stimuli* section.

There were therefore two experimental paradigms:

- Experiment 1: All images were *real images* created by reproducing either artificial reverberation or an anechoicly recorded solo musical performance with a single loudspeaker.

- Experiment 2: Images were *virtual (phantom) images* created by radiating these two signal with a pair of loudspeakers.

The perceptual interaction of S and R images was also investigated; what happens to the auditory spatial imagery of an S image when an R image is also perceived, and how is the spatial imagery of the S image affected as the R perceived image direction is changed? The motivation for this is summarized with three hypotheses, which are also rephrased as a testable research question:

H 1. *An auditory source and reverberance image can perceptually interact: In what ways is the auditory spatial imagery of a source image affected if a reverberance image is also perceived in the same audio scene?*

H 2. *Increasing the perceived spatial separation of the source (S) and reverberance (R) image reduces the interaction between the images: Does the perceived spatial separation of S and R images affect the auditory spatial imagery of an S image?*

H 3. *Spatial homogeneity of reverberance is positively correlated with perceived naturalness ("fidelity") and preference: If reverberance images are reported as being more evenly distributed around the listener, is this scene preferred over scenes with a less homogenous distribution of reverberance imagery?*

Regarding the interaction of the source and reverberance images, hypotheses 1 and 2 were investigated by affecting the spatial separation of the loudspeakers radiating the anechoic and reverberation channel (for the real-image experiment), and by pair-wise amplitude panning the R image to the side of the listener for the virtual-image experiment.

In a normal musical listening experience, it is generally the source image which we direct our attention to rather than the reverberance image (i.e. the source image creates the foreground stream and the reverberance is the background stream; Griesinger, 1996b). In this sense, the source image can be considered the *target* of our attention. If there is too much reverberation (i.e. the relative loudness of the reverberance is too high) then the temporal nuances of the musical performance will be smeared. We can therefore consider the reverberation image to be *masking* the source image.[4]

This target/ masker analogy is related to the Cocktail Party Problem (or Cocktail Party Effect), summarized by Cherry (1953) as: *"How do we recognize what one person is saying when others are speaking?"* In this problem, the person we want to listen to is considered the *target* and the other people (at the cocktail party) are *maskers*, because these other people acoustically mask the sound of the target and make it more difficult to understand what the target is saying. Cherry conducted an experiment with headphones using one persons voice as the target and another person as the masker. The subjects' task was to report what the target voice said (a "shadowing" experiment). In summary, there are five main factors which affect the listeners' ability to follow the target voice (Yost, 1997):

1. Location of the target and masker(s).
2. Visual information (e.g. lip movement).
3. Tonal properties of the speech (e.g. pitch).
4. Accents ("prosody").

[4] *Masking* here means the process by which the intelligibility or readability of a target sound is affected by the presence of another sound. The term *readability* has been used in other studies (e.g. Guastavino and Katz, 2004) to mean how clear-sounding the recorded source seems when reproduced with loudspeakers.

5. Transition probabilities (from the listeners knowledge of how certain speech sounds go together).

Most studies which investigate the first factor- i.e. the role of sound-source locations in the cocktail party problem- have generally used speech as a target and speech or speech-like noise as a masker, using speech intelligibility as a metric for the degree of unmasking (e.g. Ebata, 2003; Hawley et al., 2004). In these studies using speech, spatially separating the target from the masker improves the listeners' understanding of the target speech; an effect called "spatial release from masking" (e.g. Plomp and Mimpen, 1981; Shinn-Cunningham et al., 2001; Hawley et al., 2004). This improvement in semantic understanding is generally measured in two ways: In one method, an intelligibility rating is calculated by presenting either meaningless one-syllable words (called logatomes) or short sentences to the subject, and the percentage of syllables or words correctly understood is noted (Blauert, 1997, pg. 265). Another method uses the speech reception threshold (SRT), defined as: "*The level at which the target must be presented in order for speech intelligibility to reach some predetermined threshold level. The amount of spatial unmasking can be summarized as the difference (in dB) between the SRT for the target/masker configuration of interest and the SRT when target and masker are located at the same position*" (Shinn-Cunningham et al., 2001).

Begault and Erbe (1994) investigated how the SRT was affected by spatial unmasking using four-letter call-signs as the target and continuous speech "babble" (a mixture of two spoken voices) as the masker. The stimuli were presented with headphones, with the target speaker source simulated at 0° (i.e. straight-ahead) and the masking direction (azimuth) changed at 30° increments using HRTF filtering. A maximum intelligibility improvement of about 6 dB was obtained when the masker was panned to the 60° or 90°

position; in other words, when the masker was at 90° it could be 6 dB louder than when it was at 0°, yet the speech intelligibility was the same.

In a similar study by Plomp (1976) (reported by Blauert, 1997, pg. 337), noise with a spectrum similar to speech was used as a masking signal with a speech target at 0° (i.e. straight in front of the listener). The sounds were presented using loudspeakers (i.e. a single loudspeaker for the target and masker) and the SRT was measured for different masker azimuths. This experiment was conducted in rooms with different reverberation times (RT) and the results are summarized in figure 3.6 (from Blauert, 1997, pg. 337). As can be seen, with a reference SRT of 0 dB for the target and masker at the same direction in an anechoic room, the SRT decreases steadily as RT increases (i.e. the target had to be louder in order to have the same level of intelligibility). For all RT's, spatial unmasking occurred as the masker was moved away from the target, with an improvement of about 6 dB in the anechoic room and 1.8 dB with a RT of 2.3 s.

The above discussion explains the reasoning for hypothesis 2: *Increasing the perceived spatial separation of the S and R image will spatially unmask the S image.* A further, related hypothesis is that *increasing the homogeneity of the R image will spatially unmask the S image*, and we look at the results of an experiment by Hawley et al. (2004) to support this. Hawley et al. (2004) used a recording of a spoken sentence as the target sound and a variety of maskers such as speech-shaped noise and other voices (including time-reverse speech). What is interesting in this study is that the number of masking sources around the listener was varied; it could be a single source (with an azimuth of 0°, -30°, 60° and 90°); or two or three masker sources (at a combination of these bearings). The stimuli were presented using headphones, with HRTF filtering to affect the virtual image direction, and the

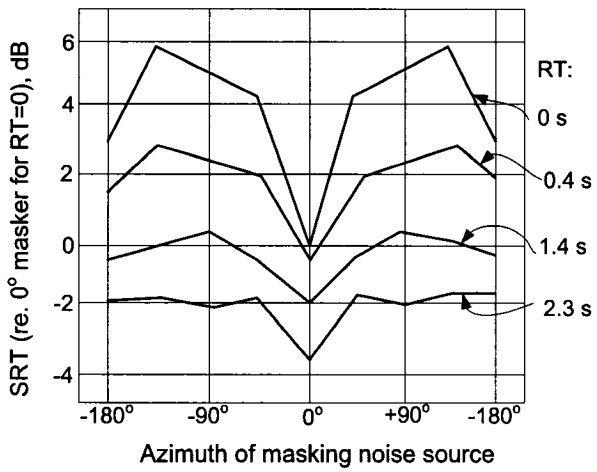


Figure 3.6: Equal intelligibility curves showing spatial unmasking of a target speech source reproduced with a loudspeaker at 0° azimuth, 0° elevation, with a noise masker reproduced from another loudspeaker around the listener. The experiment was repeated in rooms with different T60 reverberation times (RT) (adapted from Plomp (1976), reported in Blauert, 1997, pg. 337).

target was always panned in front at 0°. SRT was measured by subjects adjusting the level of the target voice until a 50% intelligibility rating was achieved. For the speech masker, the spatial unmasking effect was much stronger when there were three simultaneous masking sources distributed around the listener than when there was one or two; with an SRT 2-5 dB higher than the "best" single-masker case (a statistically significant difference). The unmasking effect was strongest when the maskers were located at 30°, 60° and 90° (the target was always at 0°). What is particularly interesting is that the level of each additional masking source was always the same- so when there were three maskers the overall level contribution from the maskers was nearly 10 dB louder than the single masker case- which is perceptually over twice as loud (Moore, 1997, pg. 133).

So if we consider a reverberance image to mask a source image, then an

increased spatial homogeneity of the reverberance image should help unmask the source image; even if the overall level of reverberation is higher. This last point explains the caveat; *Increased levels of reverberant sound can be achieved by this spatial unmasking without necessarily reducing the intelligibility (or readability) of the source image.*

In terms of masking effects of the source image by the reverberance image, we expected to see some spatial fusion of the perceived source and reverberance images when they are perceived to originate from a similar direction; leading to a source image-spread in the direction of the reported reverberance image. This fusion was expected to decrease as the reverberance image is perceived to originate from an increasingly disparate direction. As we will be dealing with relatively large reverberation times in typical concert halls (2-4 seconds), going by the results of the Plomp (1976) study we expected the maximum effect of spatial unmasking to occur when the source image is at 0° and the reverberance image is perceived from about between 45° and 90°. This is objectively investigated by looking at how the geometric properties of the source image were affected by the perceived location of the reverberance image.

Regarding H3; homogeneity of reverberance imagery in audio is noted by Malham (1999) as a key requirement of a spatial audio system to give a *natural* listening experience and many studies have found that perceived naturalness is positively correlated with *preference* (e.g. Berg and Rumsey, 2000; Zacharov and Koivuniemi, 2001a; Guastavino and Katz, 2004). This hypothesis is left as an intuition, though a formal preference experiment comparing the upmixed and original 2/0 audio scenes is reported on in chapter 6.

3.2.2 Method

Stimuli

The original audio signal used was a single channel, anechoic recording of a flute performance (Debussy's "Syrinx", 20 seconds excerpt, *legato*); the temporal and spectral envelopes of the signal are shown in figure 3.7. In order to investigate the ability of listeners to describe reverberance imagery, we wanted to create a reverberation channel with the temporal properties of reverberation. A commercially-available artificial digital reverberator[5] was used to create the reverberation channel with a 3.0 second T60 reverberation time. As this experiment concerns pair-wise amplitude panning, we only needed a single ("mono") channel of artificial reverberation ("reverb") which was fed in parallel to two loudspeakers. In accordance with the common idiom in sound recording practice, we call this reverb channel the **Wet** channel and the anechoic music channel the **Dry** channel.

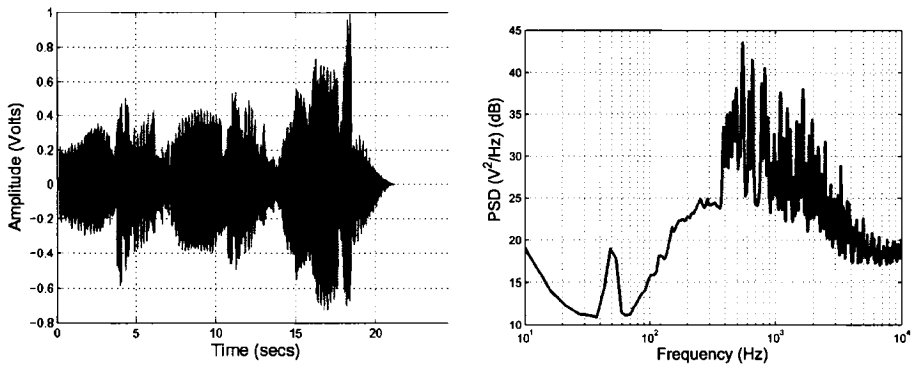
Listening room set-up

The experiment was conducted in the MARLAB[6] with a 2/2 ITU-R BS 775-1 loudspeaker set-up (rear loudspeakers at $\pm 120°$). We chose this setup as this is the loudspeaker arrangement for the new spatial audio system (as outlined in the thesis introduction). For a description of the acoustic characteristic of the MARLAB and the loudspeakers used, see appendix B.1.

In an earlier experiment investigating perceived source distance of source

[5]Manufactured by T.C. Electronics, model M3000; configuration details are given in appendix C.4.

[6]The Multichannel Audio Research Laboratory, at McGill University.



(a) Time-domain response. (b) Power spectrum of anechoic flute.

Figure 3.7: Temporal and spectral details of anechoic flute recording used in exploratory study of source and reverberance image interaction in loudspeaker audio scenes.

images in a wave field synthesis system (Usher et al., 2004b), it was found that if the listener could *see* that there was no sound source at a particular location then they would not report localizing an auditory image there. This visual biasing effect (as discussed in the section of graphical mapping techniques on page 34) was frustrating, as we wished to create a listening environment where subtle changes in auditory spatial imagery, such as image distance (i.e. egocentric range) could be perceived and reported. In experiments with the GUI, a closely surrounding visually opaque yet acoustically transparent curtain was therefore used (as shown in figure B.1 in the appendix; the curtain was 0.9 m in front of the listening position). Of course, this is somewhat ecologically invalid as the new upmix system is intended for use in domestic environments, or rooms conforming to ITU-R BS 1116 where such curtains are not used. However, it was taken as justified for controlling the visual dominance effect for auditory source localization.

To help with spatial correspondence between the apparent "real-world"

location of the auditory image and the 2D GUI map, numbered markers at 10° intervals were placed on the curtain with suspended string. Subjects were given a laser-pointer to mark the perceived image locations projected on to the curtain, and would then read-off the azimuth to help representing the image on the GUI (Choisel and Zimmer (2003) found that using a laser pointer increased subject consistency in reported image directions for both real and virtual source images).

In experiment 1 (i.e. the real-image experiment) there were additional loudspeakers at 0°, 60° and 90°. All loudspeakers were at the same distance to the listening position (1.85 m) and the same height (tweeter at 1.20 m). The angles were measured using a laser pointer on a tripod at the listening position and the distance fine-tuned by time-aligning impulses sent to each loudspeaker. Using pink-noise, the level of each loudspeaker was equalized at 70 dBA, ±0.5 dB, slow-time-weighted, measured about the listening position. For the same electronic input signal, the loudness was judged to be approximately equal for all loudspeakers (loudness is discussed latter in section 3.2.4).

Scene configurations

There were two separate experiments to investigate source and reverberance spatial imagery in loudspeaker audio. To recap:

- Experiment 1: All images were *real images* created with either artificial reverberation or anechoicly recorded sound reproduced from a single loudspeaker.

- Experiment 2: Images were *virtual (phantom) images* created by radiating the same electronic signal with a pair of loudspeakers.

There were three scene configurations for both the real and the virtual image experiment:

1. Only the wet channel (artificial reverberation) was active (scene **W**).
2. Only the dry channel (anechoic recording) was active (scene **D**).
3. Both the wet and dry channel were active together (scene **D+W**).

In the first two cases, the wet and dry channel were panned at various locations around the listener, whereas in the case when both wet and dry channel were active together scene (**D+W**) the dry channel was reproduced by either only the centre loudspeaker (in experiment 1) or with equal gain by the front left and right loudspeaker pair (in other words, the intended source image location was at 0°). This is summarized in table 3.2.

Scene Config.	Stimuli	Intended image direction
D	100% Dry	0°; 30°; 60°; 90°; or 120°.
W	100% Wet	0°; 30°; 60°; 90°; or 120°.
D+W	50% Dry 50% Wet	Source image: 0° R image: 0°; 30°; 60°; 90°; or 120°.

Table 3.2: Scene configurations for experiment 1 and 2. Experiment 1 had no virtual image panning- all images were real; and experiment 2 used pairwise amplitude panning. In scene configuration **D+W**, the dry signal was reproduced from the centre channel only (experiment 1) or equally from the front loudspeakers (experiment 2). In other words, for scene **D+W** the *intended* (or panned) direction of the source image was at 0°. The 50% wet or dry level means that in this scene, the contribution of the reproduced wet and dry channel to the sound pressure level at the listening position was the same.

In the real-image experiment, the image direction was affected by simply radiating the dry or wet channel from a particular loudspeaker. For the virtual-image experiment, the only active loudspeakers were in the conventional 2/2 ITU-R BS 775-1 configuration (no centre-speaker) as shown in figure B.1. For the virtual-image experiment, the image direction was controlled by sending the dry or wet audio signal to a loudspeaker pair (i.e. the two loudspeakers closest to the intended image direction) with different gains. This is a form of pair-wise amplitude panning (PWAP). The five intended image angles we investigated were: 0°, 30°, 60°, 90° and 120°. The images at 30° and 120° were always real images (i.e. only a single channel was active).

The loudspeaker gain coefficients were calculated using the tangent-panning law (see equation (2.2) on page 44), as it has been found that this predicts perceived image direction better than Blumlein's classic stereophonic law of sines for mobile-head listeners (Bernfeld, 1973). The side angles were chosen because of the importance of lateral-incident sound reflections on subjective "spatial impression" (Barron, 1971; Barron and Marshall, 1981). Furthermore, we wanted to investigate the spatial-unmasking effect of separating the source and reverberance image, and in the Plomp (1976) study on spatial (un)masking with reverberation the unmasking effect for a target source at 0° was strongest when the reverberation was reproduced from 90°. With the target-masker analogy, it was wondered if there would be any change in the reported spatial imagery of the target image (i.e. source image) if the masking image (i.e. reverberance image) was perceived to originate from a different spatial location.

Trial#	Scene config.	Image direction: S image	R image
1	**D**	0°	X
2	**D**	30°	X
3	**D**	60°	X
4	**D**	90°	X
5	**D**	120°	X
6	**W**	X	0°
7	**W**	X	30°
8	**W**	X	60°
9	**W**	X	90°
10	**W**	X	120°
11	**D+W**	0°	0°
12	**D+W**	0°	30°
13	**D+W**	0°	60°
14	**D+W**	0°	90°
15	**D+W**	0°	120°

Table 3.3: Stimuli permutations. In experiment 1 the image direction was a real image direction, from a single loudspeaker source at the intended direction. In experiment 2 a virtual image was created by amplitude panning using the loudspeakers closest to the intended direction. In both experiments for the 30° and 120° image, only a single loudspeaker was active and it is therefore a real image not a virtual image. The symbol X indicates that either the source or reverberation channel was not active for this trial. The trial order was randomized for each test.

Subject training and instructions

All subjects who took part in the experiments were students in a Tonmeister sound recording program with at least three years critical listening experience. Six subjects took part in the real image experiment and five in the virtual image experiment (i.e. experiments 1 and 2). Subjects were paid $15 for each experiment. A training experiment was undertaken where the listeners were explicitly told whether there is a dry, wet, or a mixture of channels being reproduced, though they were not told the intended (panned) image direction. This enabled the subjects to familiarize themselves with the instructions (such as what is meant by a source or reverberance image) and the GUI. The subjects were free to rotate their heads but told to keep it beneath the mark on the ceiling corresponding to the listening position, as described in the instructions in appendix D. The instructions were identical for the real and virtual-image experiments.

Summary of experiments

- Two experiments: the first with single loudspeakers (i.e. *real* images), the second with pair-wise amplitude panning (i.e. with *virtual* source and reverberance images) with a 2/2 loudspeaker configuration.
- Intended (panned) image directions: 0°, 30°, 60°, 90°, 120°. All on the right-hand side of the listening position.
- Three scene configurations for each experiment:
  - Scene **D**: A mono anechoic (dry) channel of a flute recording is panned around the right-hand side of the listener.

  - Scene **W**: A mono channel of artificial reverberation (the wet channel) created from the flute recording is panned around the right-hand side of the listener.
  - Scene **D+W**: The anechoic channel is panned at 0° and the reverberation channel is panned around the right-hand side of the listener.

- Listeners drew perceived source and reverberance image location and extent (in the horizontal plane) using the GUI.
- Six paid, experienced subjects took part in the experiments.
- 15 unique trials for each experiment (see table 3.3), presented in 3 sessions (i.e. 2 repeats) with a 5-15 minute break between sessions. In each session the trial-order was randomized.
- Each stimulus played continuously until the subject proceeded to the next trial.
- No feedback was given to the subject at any time during the test.

3.2.3 Results

For experiment 1 (i.e. all images were real images), there were 6 (subjects) * 3 (runs)=18 graphical responses for each unique audio scene, and there were 15 responses for experiment 2 (as there were only 5 subjects). Density plots were created by overlaying these responses and are shown in figures 3.8 and 3.9.

From each individual scene drawing, image width; distance and direction were calculated for all three image categories (i.e. stable and unstable source

Real-images (experiment 1)

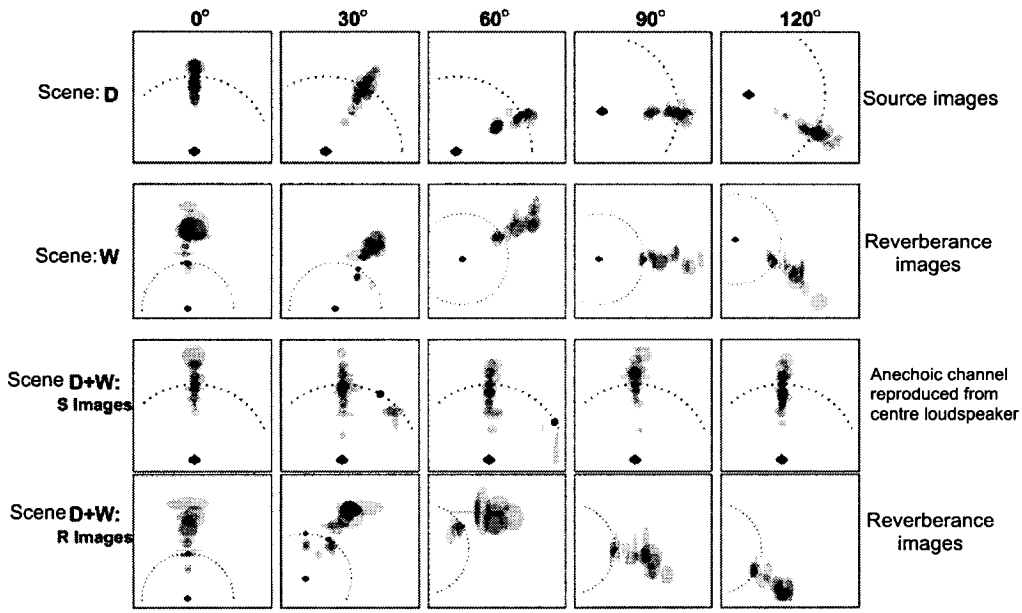


Figure 3.8: Density plots for real image experiment (i.e. each audio channel reproduced with a separate loudspeaker). Responses summed from 6 subjects, 3 runs. Arranged by panned image direction (column) and scene configuration (row):
First row: Scene **D**: Only Dry channel presented; reported stable and unstable S images are shown.
Second: Scene **W**: Only Wet channel presented; reported R images shown.
Third and forth row: Scene **D+W**: Dry channel from centre speaker, wet channel panned to the side. Third row shows reported stable S images, bottom row shows reported R images.
Black dot indicates active loudspeaker location (only one loudspeaker was active at a time in scene **D** and scene **W**). The spatial scale is the same for each scene (i.e. row) and the density scale is the same for all plots.

Virtual-images (experiment 2)

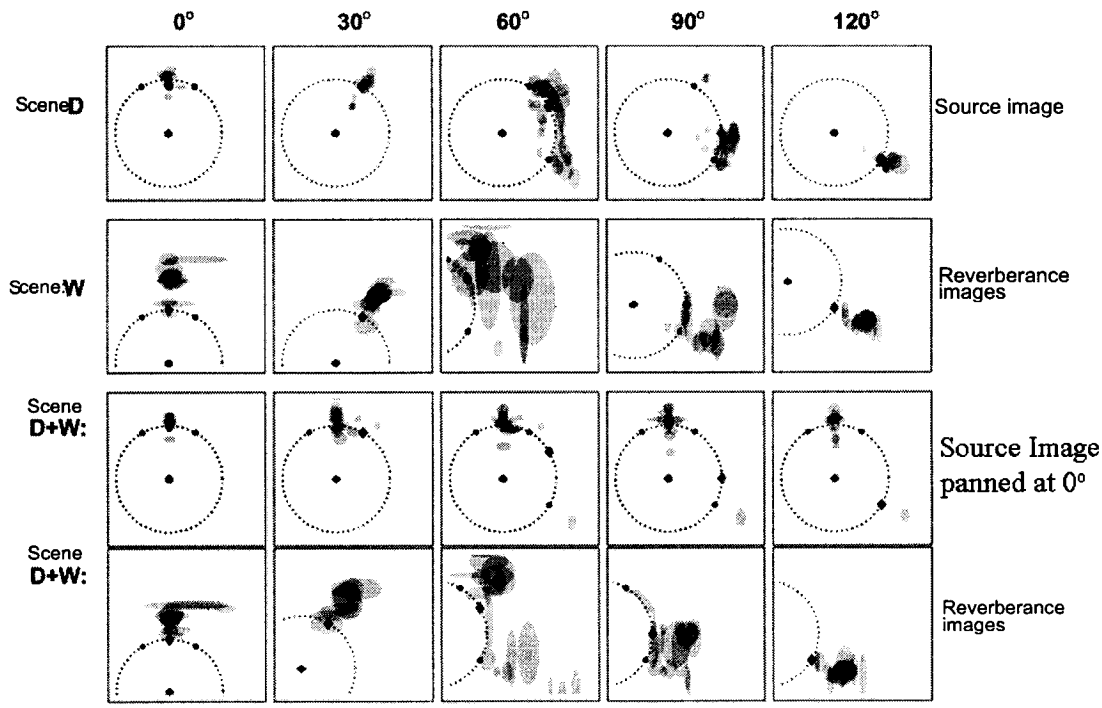


Figure 3.9: Density plots for virtual (phantom) image experiment. Responses are summed from 5 subjects, with 3 runs for each stimulus. Arranged by panned image direction (column) and configuration (row):
First row: Scene **D**: Only Dry channel presented; elicited stable and unstable source images shown.
Second row: Scene **W**: Only Wet channel presented; elicited reverberance images shown.
Third and forth rows: Scene **D+W**: Dry channel panned between front loudspeakers, wet channel panned to the side. Third row shows reported stable S images, bottom row shows reported R images.
The black dot indicates the location of active loudspeakers and the green diamond shows the intended image direction. When the intended image direction is 30° and 120°, the image was a real image not a virtual image. The spatial scale is the same for each scene (i.e. row) and the density-scale is the same for all plots.

(S) images and reverberance (R) images): These were calculated using the method shown in figure 3.5. These image attributes for the **D** and **W** scenes are summarized in figure 3.10, and for the **D+W** scene (i.e. when the S image was panned at 0° and the R image was panned to the right of the listener) the image properties are summarized in figure 3.11. In all these plots, mean image values (i.e. mean width, distance and azimuth) are shown, ± two standard deviations (SD's). The mean and SD were calculated for all descriptions of a particular scene; which were from 18 responses in the real-image experiment and 15 in the virtual-image experiment. If more than one unique source or reverberance image was drawn for a given scene (though many ellipses could have been drawn to describe a single image) then the image direction was not be calculated (it would not make sense to average the azimuths of two separate images).

3.2.4 Discussion

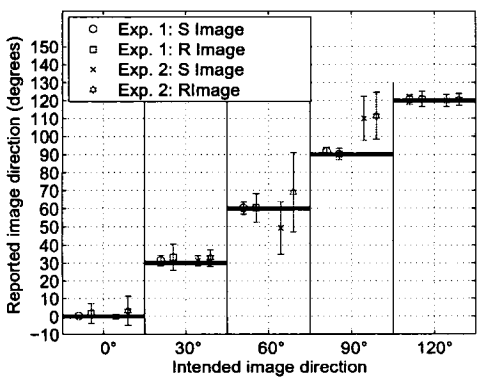
Configurations D and W

The discussion shall first consider the scenes when either just the dry channel (scene **D**) or just the wet channel (scene **W**) was reproduced for a given stimulus.

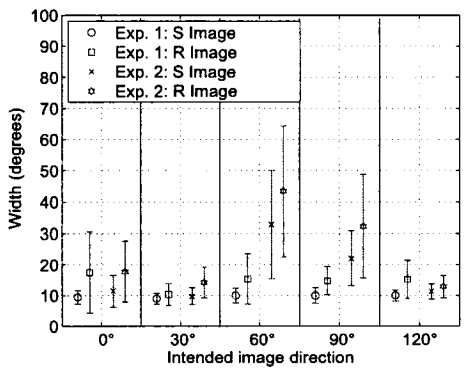
- Localization "error":

For the virtual image experiment, the term *error* is not an absolute error- as true veridicality doesn't exist- the error here was calculated as the difference between the intended (i.e. panned according to the tangent panning law) and reported image direction.

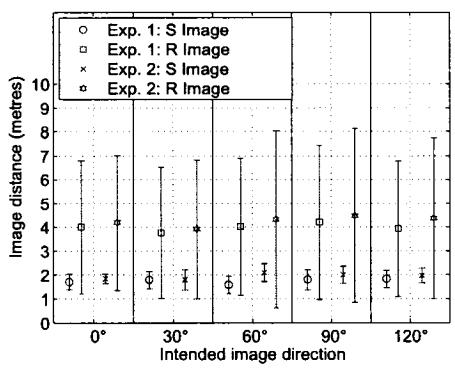
Elicited image properties: configurations D and W (i.e. only anechoic or reverberation channel reproduced at one time).



(a) Reported image direction (intended direction shown with heavy line).



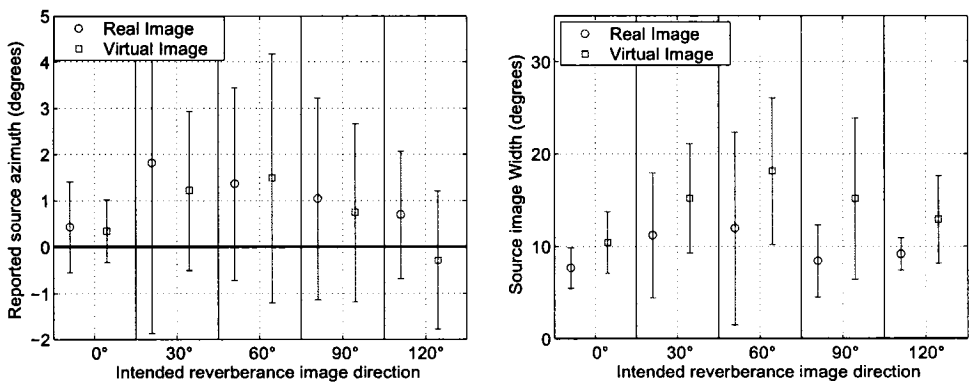
(b) S and R image width.



(c) Absolute image distance.

Figure 3.10: Stable source (S) and reverberance (R) image properties for real (exp. 1) and virtual (exp. 2) image experiments.

Configuration D+W (source image always panned at 0° and R image panned on the right-hand side).



(a) Stable source image direction (intended at 0°).

(b) Stable source image width (ASW).

Figure 3.11: Reported stable source image direction and width for real (exp. 1) and virtual (exp. 2) image experiments. Intended source image direction is zero degrees for all trials in scene configuration **D+W**. Anechoic channel is always reproduced from the centre speaker in the real-image experiment and is panned at 0° in the virtual-image experiment. The reverberance image is always a real image when panned at 30° or 120°. Due to the non-normal distribution of the data, using the error bars to determine statistical significance of the trends in the mean are misleading; both the direction and width of the stable S image were, in fact, significantly affected by the panned direction of the reverberance image (please see discussion section for details).

As can be seen from a quick eye-ball inspection of the density plots, for **real images** subjects reported the direction of both source and reverberance images with a very high degree of accuracy. Looking at figure 3.10(a), the maximum mean-error was +3.1° for the reverberance image (at 30°) and +1.8° for the source image (90°), with 95% confidence intervals of 8° and 3° for the reverberant and source images. The mean localization errors and standard-deviations were similar or smaller than found with other experiments which used single loudspeaker sources around the listener (which is typically 3° to 10°; Blauert, 1997, pg. 41).

Side **virtual images** showed a consistent "error" bias. We find that for the virtual images panned at 60°, the source image is reported about 10° (mean) closer to the front-right loudspeaker, and that the reverberance image is pulled by a similar amount in the direction of the rear-right loudspeaker. This is interesting, but we can not draw firm conclusions as the between-subject standard deviation for reported source and reverberance image azimuth was 15° and 20°. This response variation (i.e. localization blur) was similar to that found in other studies with amplitude-panning of a single audio channel using two side loudspeakers (Ratliff, 1974; Martin et al., 1999; Pulkki and Karjalainen, 2001). For the virtual images panned at 90°, both the source and reverberance images were reported at about 112°; in agreement with Corey (2002) who found that using a similar loudspeaker set-up and PWAP, images panned at 80° and 100° were also pulled towards the rear loudspeaker. However, this is in contrast to the Theile and Plenge (1977) study, discussed previously (summarized in table 2.1) where images panned to the side of the listener were pulled *forward*; for example, images panned at 90° were reported at 77°.[7]

[7]The Theile and Plenge (1977) study used amplitude-panned noise and anechoic speech. The loudspeaker base-angle was 60° and the central loudspeaker axis was shifted around

• Image width:

Still for scene **D** and scene **W** (i.e. when just one channel was reproduced at a time) it was found that for both real and virtual images the width of the reverberance images were consistently larger than for the source images (see figure 3.10(b)). To investigate the effect of intended image direction on image width, a type III sum of squares GLM procedure was used with fixed factors SUBJECT and (intended) DIRECTION. This method was chosen because it is robust to non-normal distributions of data (SPSS, 1995).

For the **real images**; the width of the *source* image was not significantly affected by the intended image direction, but was affected by the SUBJECT factor ($F = 19.386, p < 0.00$). The *reverberance* image width was smallest when reproduced from 30° (mean width of 12°, 95% CI of 6°), and was largest at 0° (mean width of 17.6°, 95 % CI of 12°). The relationship between reproduction direction and width for reverberance images was also investigated and found to be "statistically interesting" ($F = 2.207, p = 0.079$). Again, the SUBJECT factor was significant ($F = 3.311, p = 0.010$), as was the SUBJECT*DIRECTION interaction ($F = 2.607, p = 0.002$)- suggesting that subjects had different methods for reporting reverberance image width as a function of image direction.

For the **virtual image** experiment; the width of the side images (i.e. when the image was panned at 60° or 90°) was larger than for front (0°) image. For the source image, the GLM analysis revealed that SUBJECT, intended DIRECTION, and interactions were significant ($p < 0.001$). The same was found for reverberance images. The width was greater than the directional spread predicted by the Pulkki (1999b) model- which looks at the

the listener. When the axis was at 90° to the side of the listener and the inter-speaker level difference was 0 dB, the image was reported at 80°, but with an interquartile range of about 80°.

discrepancy between azimuth predicted by interaural time differences and the azimuth predicted by interaural level differences (his model predicts a maximum spread of 6° for a source panned at 60°, but the reported image width in this experiment was closer to 15°).

Although the SPL of the reproduced anechoic and reverberant audio channels was the same at the listening position (i.e. for all real and virtual locations), the S image was consistently heard to be louder than the R image. This could be explained by a forward-masking suppression-of-reverberation effect, similar to the precedence effect with discrete echoes (Blauert, 1997; Moore, 1997). As already discussed; a simple example of the forward masking effects of reverberation is to record speech in an ordinary room and to replay the recording backwards (Houtsma et al., 1987): the perceived reverberance is louder when the speech is played in reverse. Source image width is positively correlated with loudness (Keet, 1968; Usher and Woszczyk, 2003), but as reverberance images were generally larger than source images, yet less loud, it is difficult to interpret the loudness-width relationship as it applies to reverberance images.

- Image distance:

Various investigations into the effect of sound source azimuth on perceived loudness have found that although the direction-effect varied for different rooms, it was generally <1 dB. For example, Ratliff (1974) used octave-band filtered noise reproduced with a single loudspeaker around the listener and concluded that in a "normal" listening room (70 m^3, RT=0.35 s) "*the average auditory response is almost equally sensitive around the full azimuthal circle, although there is a tendency for the back to be less sensitive than the front by about 1 dB*". Due to the approximately inverse-square relationship of real-world source distance with proximal level (which holds for close sources),

loudness is a major cue for determining image distance for close sources. For instance, using speech stimuli Zahorik (2002) found the loudness cue was used by listeners to determine source distance in preference to a conflicting direct sound-to-reverberation level cue. We would therefore expect the perceptually softer reverberance images to be heard further away than the source images. A quick visual inspection of the density plots shows this is the case for both real and virtual images. However, as shown in figure 3.10(c) the response variation for reported R image distance is very large.

The direction-dependance of image distance for **real images** is minimal for source images for all directions except at 60°, where it is closest (mean 1.5 m). A type III sum of squares GLM procedure was used, with fixed-factors SUBJECT and DIRECTION, to investigate the direction-distance relationship, and this difference was found to be statistically significant ($p = 0.042$). The reverberance image distance was not significantly affected by panned direction.

For **virtual images**, contrary to the previous finding the source image is heard farthest at 60° (about 2.1 m) and closest at 0° (1.84 m). It was found that the panned source image direction significantly affected reported image distance ($F = 6.723, p = 0.000$), as did the subject factor and the interactions. As with real sources, the reverberance image distance is not affected by (intended) direction in a statistically significant way ($F = 1.568, p = 0.198$), though the SUBJECT factor does affect reported reverberance image distance ($p = 0.000$).

Configuration D+W

This section will only consider the **D+W** scene configuration, where both the dry channel and wet channel are reproduced in the same audio scene. In this scene configuration, the intended source image direction was 0° for both the real and virtual image experiment. The reverberation (wet) channel was reproduced from either real speakers at 0°, 30°, 60°, 90° or 120°, or was pair-wise amplitude panned so that the intended reverberance image direction (according to the tangent panning law) corresponded to these directions (though when the panned direction was 30° or 120°, it was always a real image with just a single loudspeaker active).

In order to investigate the interaction of S and R images as a function of spatial separation (i.e. hypotheses 1 and 2), we will look at how the elicited spatial properties of the (stable) source image change as the (intended) reverberance image direction is changed. A type III sum of squares GLM procedure was used with two factors: subject and intended reverberance image direction. Three analyses with dependant variables corresponding to source image spatial properties were investigated to see if they were affected by the change in intended reverberance image direction. Results are summarized in table 3.4.

	Significance:	
Dependant variable:	Real Source	Virtual Source
S. Image Azimuth	$p = 0.213$	$p = 0.021$
S. Image Width	$p = 0.011$	$p = 0.001$
S. Image Distance	$p = 0.203$	$p = 0.336$

Table 3.4: Statistical significance of changes in stable source image spatial properties as the reverberance image is panned around the listener.

For the real-source experiment, the number of cases was 90- i.e. 6 subjects * 3 runs * 5 intended reverberance image directions. For the virtual image experiment the number of cases was 75 for 5 subjects. As can be seen from figure 3.11(a), the source image was pulled in the direction of the reverberance image, but this detent affect was only significant for the virtual image experiment. This detent was maximal when the R image was panned at 60° (mean detent of +1.5°). This confirms hypotheses 1 and 2 about the spatial fusion of the source and reverberance images when perceived spatial separation is small. From figure 3.11(b), a similar trend is observed with stable source image width, which is also largest when the reverberance image is panned at 60° for both the real image (14° source width) and the virtual image (18° source width). As found in the discussion for scene configurations **D** and **W**, the large inter-subject variability means there are no consistent trends in perceived source image distance. Even though inspection of the plots in figure 3.11 do not make this trend obvious (due to the non-normal distribution of the data, the error bars overlap), the data in table 3.4 from the type III GLM analysis show that both these trends in source image distortion artifacts are in fact significant for the independent variable of intended (i.e. panned) R image direction.

Summary of findings

When a single channel of artificial reverberation or an anechoic recording is pair-wise amplitude panned around a listener using loudspeakers arranged in the conventional 2/2 manner:

- The response variation for reporting the perceived direction of virtual

images panned at 60° and 90° is similar for both source and reverberance images.

- Reverberance images are consistently reported wider and farther than source images for both real and virtual images.
- Using pairwise amplitude panning (PWAP), both source and reverberance virtual images panned at 60° and 90° are wider than virtual images at 0°.
- S and R images panned at 90° (using PWAP with two loudspeakers at 30° and 120°) are pulled towards the rear loudspeaker and reported at 110° (mean).

When a single channel of anechoic sound is panned at 0° with a loudspeaker pair at ±30° and a reverberation channel is panned around the listener at 0°, 30°, 60°, 90° and 120°:

- Spatial distortion of the source image is greatest when the reverberance image is panned at 30° and 60°, as shown by a pulling of the source image in the direction of the reverberance image and an increase in the perceived source image width.
- Spatial distortion of a source image panned at 0° reduces as the reverberance image is panned away from 0° (i.e. panned at 60°, 90° and 120°).

3.3 Subjective Design Criteria of the New Audio System

The design criteria of the new spatial audio system which can be evaluated with listening tests are outlined here. As mentioned in the thesis introduction, the new audio system should enhance the perceived spatial imagery of a conventional two-channel recording- specifically a two-microphone recording of a solo musical performance in a concert hall- reproduced with a 2/0 loudspeaker pair. It should be borne in mind that the new audio system is a signal processing-based solution, not a new transduction mechanism. This ensures compatibility of the new upmix system with existing domestic home-theatre systems such as a system arranged according to the 2/2 ITU-R BS 775-1 recommendation described on page 66. The 2/0 reproduction is therefore the reference case which applies to the three criteria outlined below. How these goals can be realised using signal processing is discussed in the next chapter.

1. *Spatial distortion of the source image in the upmixed scene should be minimized (compared with 2/0 loudspeaker audition)* .

   The auditory spatial imagery of the source image should be the same in the upmixed and original 2/0 scenes. This can be evaluated using the GUI as was done in the experiment just described, where the elicited image width, distance and azimuth was compared in 2/0 and 2/2 scenes. Another way of interpreting this goal is that the new system should be designed in such a way that it *respects the mixing intentions of the sound engineers involved in the production of the original two-channel recording*, at least in terms of the source ("front-stage") imagery. The degree to which this criterion is satisfied can be measured in the same

way as with the previous experiment; using data from the GUI to compare source image geometry in the 2/0 and upmixed scenes.

2. *Reverberance imagery should have a homogenous distribution in the horizontal plane; in particular, reverberance image directional strength should be high from lateral (±90°) directions.*

   It was found in the experiment that when the reverberance image was panned at 90°, from the density plots it can be seen that although the R images were sometimes located in the direction of the rear loudspeakers, the imagery was spread out between the side speakers; the mean image direction was about 112°, with error bars just touching the direction of the rear loudspeaker at 120°. Phantom imaging to the side would increase the perceived lateral reverberance image content of the sound scene, and lateral reflections have been shown to contribute to increased spatial impression (Barron and Marshall, 1981). Furthermore, it was found in the study just reported that source image distortion was less when the reverberance image was panned to the back and side of the listener (i.e. at 90° or 120°) than when it was panned to the front (i.e. 30° or 60°).

3. *In terms of overall sound, the new system should not be dispreferred to a conventional 2/0 system.*

   In the context of a listening for *pleasure*, the upmixed audio scene should ideally be preferred over a reference 2/0 reproduction created using the same recording. For the specific subjective design criteria for the upmixer introduced in this thesis, this ideal goal is relaxed to saying "not dispreferred" to reflect the reduced importance of this goal compared with the first two.

Chapter 4

Theory and design on the New System

In this chapter the signal processing mechanism for the new spatial audio system is introduced. It can be classed as a blind active upmixer (see section 2.2.3) for creating two new signals from two original audio signals. For the sake of brevity, the acronym ASUS is used to stand for the new system; **Adaptive Sound Upmix System**. An implementation for five input signals, for example from a film sound-track or musical recording reproduced with a DVDA player, is shown in appendix A. The implementation of the adaptive filtering method as it relates to audio upmixing presented here is entirely the work of the author, and from an extensive review of the published academic and patent literature, no previous work ("prior art") which suggests such a method could be found. However, the algorithm and its implementation used to accomplish this *is* well established in the literature.

Although the electroacoustic performance criteria which can be used to

evaluate the system in this chapter are introduced here, the actual analysis of performance is reported in chapter 5 and a subjective evaluation with listening tests is discussed in chapter 6.

4.1 Electronic Design Criteria

Here, the subjective design criteria for the new upmix system outlined in section 3.3 are translated into a set of criteria which can be evaluated using electronic measurements. The criteria can be divided into two categories; those which concern source imagery and those which concern reverberance imagery. To describe how to realize these goals in signal processing terms the input signals to the upmixer are modeled as two parts; a part which affects spatial aspects of the *source* imagery and a part that affects spatial aspects of *reverberance* imagery. How these two parts are distinguished in electronic terms is discussed shortly with the *signal model*, but for now these two electronic components of the input signals are simply called the Source (image) component and the Reverberance (image) component. In the left channel, these components are abbreviated to S_L and R_L, and in the right channel S_R and R_R. This is just an abstract representation to make the foregoing translation of the subjective performance criteria to the electronic criteria easier. Other sound components which do not contribute to S or R imagery, i.e. noise in the recording environment from a source other than the musical instrument, are assumed to be absent or at least very low in level. Therefore the two input signals (e.g. the left and right channels from a CD player) can simply be modeled as the sum of these two sound components- as summarized in figure 4.1.

According to the principles of pair-wise panning discussed in section 2.1.8, if the source components S_L and S_R are coherent (i.e. with a high absolute cross-correlation peak at a lag less than about 1 ms) then radiation of these signals with two loudspeakers either in front (as with a conventional 2/0 loudspeaker system) or to the side (as was investigated in the **D** scene configuration in the phantom-image experiment in section 3.2) of the listener will create a phantom source image between the loudspeakers. As was shown with the **W** scene configuration in the phantom-image experiment, the same applies to the radiation of the reverberance components; so if R_R could be extracted from the right channel and radiated from the rear-right loudspeaker, a listener would perceive a reverberance image on the right-hand side, as shown in figure 4.1. As we are dealing with a noise free (or at least, very low noise) recording environment, the reverberance image components can simply be defined by exclusion: they are those sound components of the two input signals which are not correlated.[1] We shall later see how this general definition is limited with a frequency-time model.

The two subjective design criteria regarding source and reverberance imagery are now translated into a method which can be undertaken empirically on the output signals of the new upmixer:

1. *Spatial distortion of the source image in the upmixed scene should be minimized.*

   To maximize the source image fidelity in the upmixed audio scene, source image components L_S and R_S should not be radiated from the rear loudspeakers in the upmixed sound scene. If they were, then they

[1]This definition is valid for the context of this thesis- i.e. recordings in concert halls with no artificial reverberation- but is not entirely robust if we consider recordings where a single channel of reverberation is mixed to two channels. However, mono-reverberation is rarely used in recordings today.

could perceptually interact with the source image components radiated from the front loudspeakers and cause the source image to be distorted (it was shown in the previous chapter that side phantom images can be created if correlated audio signals are radiated from front and rear loudspeakers). Therefore, all those sound components which contribute to the formation of a source image should be removed from the rear loudspeaker signals, yet those source image components radiated from the front loudspeakers should be maintained. A way of measuring this in electronic terms is to ensure that the signal RS is uncorrelated with signal L, and that LS is uncorrelated with R. For a signal sampled at time n, this is mathematically expressed in (4.1):

$$\begin{aligned} 0 &\approx \sum_{n=-\infty}^{\infty} RS(n)L(n-k) \\ &\text{and} \\ 0 &\approx \sum_{n=-\infty}^{\infty} LS(n)R(n-k). \\ k &= \pm 0, \pm 1, \pm 2, \ldots, \pm N. \end{aligned} \tag{4.1}$$

The lag range N should be equal to 10-20 ms (500-1000 samples for a 44.1 kHz sample-rate digital system), as it is the early sound after the direct-path sound which primarily contributes to spatial aspects of source imagery (such as source width) and the latter part to reverberance imagery (Barron, 1971; Morimoto, 2002; Soulodre et al., 2003). For lag times (k) greater than 20 ms or so, the two signals may be somewhat correlated at low frequencies- as explained later.

2. *Reverberance imagery should have a homogenous distribution in the horizontal plane; in particular, reverberance image directional strength should be high from lateral ($\pm 90°$) directions.*

The implication of this statement is that in order to create new reverberance images to the side of the listener, the side loudspeaker channels (e.g. R and RS) should have some degree of correlation. Under such circumstances, pair-wise amplitude panning could occur between the two loudspeakers; with the perceptual consequence that the reverberance image would be pulled *away* from the side loudspeakers and to a region *between* them, as was found in the experiment reported in chapter 3. This is summarized in (4.2):

$$\begin{aligned} 0 &\neq \sum_{n=-\infty}^{\infty} LS(n)L(n-k) \\ &\text{and} \\ 0 &\neq \sum_{n=-\infty}^{\infty} RS(n)R(n-k), \\ k &= \pm 0, \pm 1, \pm 2, \ldots, \pm N. \end{aligned} \tag{4.2}$$

Again, N would be equal to 10-20 ms.

Regarding the degree of correlation between the two rear channels (i.e. the "extracted ambiance" signals), the optimal relationship is not as straightforward as with the above two electronic criteria. Although a low interaural coherence is conducive for enveloping wide auditory imagery (Kurozumi and Ohgushi, 1983; Martens, 1999), this does not necessarily mean the rear loudspeaker channels should be uncorrelated *de facto*. The correlation between two locations in a reverberant field is dependant on the distance between them and is frequency dependant (Jacobsen and Roisin, 2000). For instance, at 100 Hz the measuring points in a reverberant field must by approximately 1.7 m apart to have a coherence of zero (assuming the Schroeder frequency of the hall is less than 100 Hz). Microphone-pair recordings in concert halls

therefore rarely have total decorrelation at low-frequencies. Furthermore, for sound reproduced with a loudspeaker pair in normal echoic rooms, due to loudspeaker cross-talk, head diffraction and room reflections, the interaural coherence (especially at low frequencies) is close to unity regardless of the interchannel coherence of the loudspeaker signals (Kim et al., 2005) (as discussed in the section on panning in chapter 2.1.8).

4.2 System Overview

The ASUS can be adapted for any number of input channels (> 2)- an example for five input channels (e.g. from a DVDA or SACD recording of a solo instrument performance) is shown in appendix A. To limit the size of the thesis and to ensure a suitably in-depth investigation, only the two input signal configuration is considered.

In the description of the ASUS in this chapter, we will assume the two input signals are directly from the microphone pair; therefore the recording media can be eliminated from the discussion. These two signals from each microphone at sample time n are $m_1(n)$ and $m_2(n)$. As discussed in the electronic design criteria, the goal of the ASUS is to remove those sound-image components in the two mike signals which are correlated (i.e. the source image components) leaving the reverberance-image components to be radiated from the rear loudspeakers. Therefore, if a function can be found which can be applied to one mike signal to make it electronically the same as the other (generally; in the frequency-domain this is called the transfer function and in the time-domain the impulse response; Oppenheim and Shafer, 1999), then the correlated sound components which contribute to source imagery can be

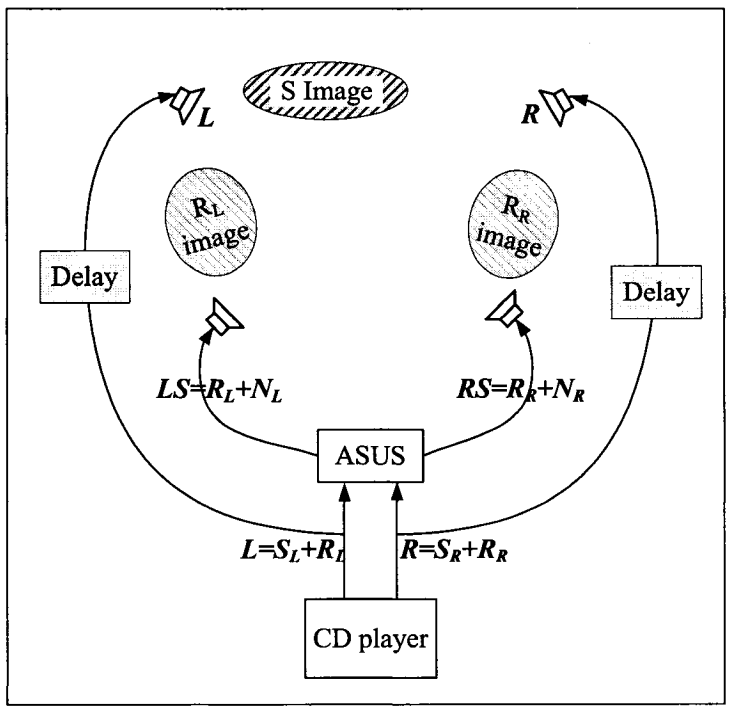


Figure 4.1: Overview of source and reverberance imagery created by the new upmixer (i.e. the ASUS). In this 2-to-4 channel embodiment which is studied in the thesis, the system processes two input signals (L and R) of a musical recording (such as the output of a CD player) and creates two new signals LS and RS which are radiated by the rear loudspeakers. The front loudspeakers radiate the original signals with a time delay to account for the IO latency of the ASUS. It is assumed that those components in the original signals which are correlated contribute to the formation of the source image; these components are S_L and S_R in the left and right channel from the CD player. The remaining components in each channel contribute to the perceived reverberance image (these sound components are R_L and R_R). The ASUS tries to extract these reverberance image components and radiate them from the rear channels, creating reverberance virtual (phantom) images to the side of the listener. Other "noise" artifacts created by the ASUS, N_L and N_R, are also radiated. To avoid distortion of the source image there should be no source image components in the rear loudspeaker channel; that is, signal RS should be uncorrelated with signal L, and LS uncorrelated with R.

removed by subtracting the two signals after one of these signals has been processed by this function. An overview of the signal processing structure of the proposed system is given in figures 4.2- 4.4, which can be summarized as four important elements:

1. Filtering a first input audio signal with respect to a set of filtering coefficients (typically, with a 1024-tap FIR filter).

2. Time-shifting a second audio signals with respect to the first signal (typically with a delay of about 5 ms).

3. Determining a first difference between the filtered and the time-shifted signals. This difference signal is then radiated with a separate loudspeaker.

4. Adjusting the set of filtering coefficients based on the first difference so that the difference signal is essentially orthogonal to the first input signal.

4.2.1 Signal Model

The impulse response (IR) between two locations in a concert hall can simply be measured by creating a large acoustic impulse- such as with popping a balloon- and measuring the pressure change at the other location using a microphone, an electronic amplifier and signal recorder.[2] The instantaneous time-domain transfer function can only be measured with this "impulsive excitation" method if the onset of the impulse is instantaneous and a single

[2]When the IR is undertaken in a reverberant environment it is sometimes called the Room Impulse Response (RIR; Neely and Allen, 1979; Elko et al., 2002).

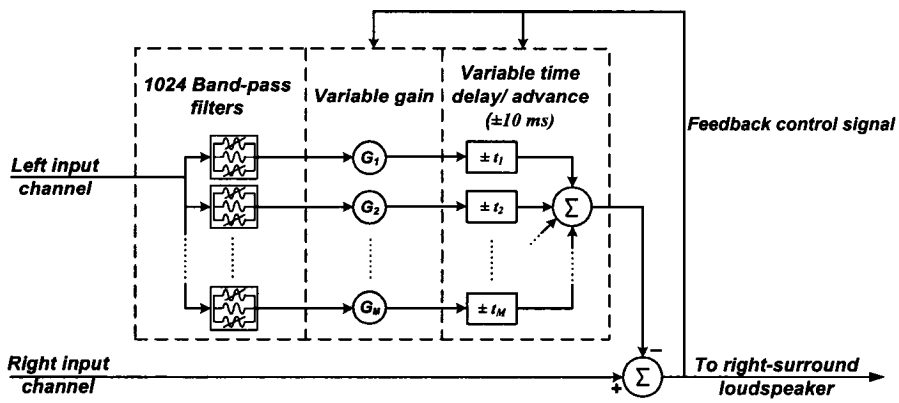


Figure 4.2: Conceptual overview of system, for generation of one surround channel. M equal band-width band-pass filters (typically, M=1024) are weighted by a positive or negative gain and the filtered channel can then be time advanced or delayed to minimize the difference signal (the differencing is actually done on a frequency-by-frequency basis). In the actual implementation, the right-input signal is delayed by about 10 ms, and the variable delay operates between a delay of 0 and about 20 ms. The feedback control signal is used to update the filter gains and delays so as to reduce the output signal level in the mean-square sense. The left-surround signal is generated by swapping the left and right input channels.

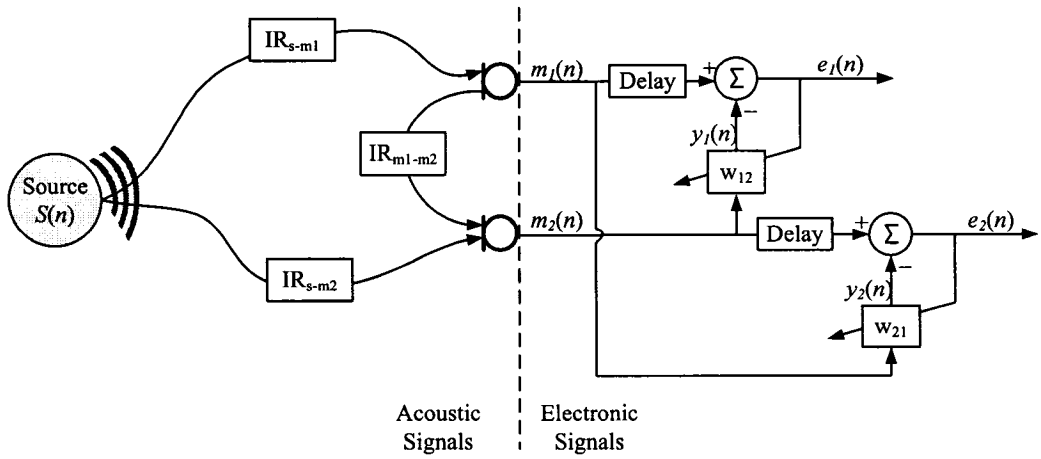


Figure 4.3: Electroacoustic signal chain of two input channel ASUS embodiment. A single sound source (i.e. musical instrument) is recorded using two microphones. The output signal of each mike at sample time n is $m_1(n)$ and $m_2(n)$. These signals are modeled as a convolution of the radiated source signal, $S(n)$, and the two source-microphone impulse responses IR_{S-m1} and IR_{S-m2}. The two mike signals are then filtered with the adaptive FIR filters w_{12} and w_{21}. The filters are adapted over time so that the level of the difference (error) signals is minimized. Ideally, the difference signals $e_1(n)$ and $e_2(n)$ do not contain any information which would contribute to the source image (i.e. if these two signals were auditioned alone). These two difference signals are radiated by the rear loudspeakers as shown in figure 4.4.

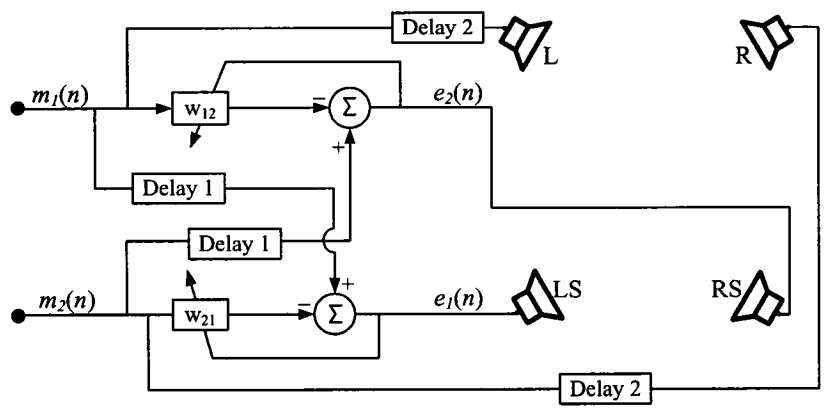


Figure 4.4: Another view of the system showing the signals feeding the four loudspeakers. Delay 1 and delay 2 are different: the former is to allow for non-minimum phase impulse responses for when the direct sound arrives at one microphone before the other, and the latter is to account for IO delay caused by the DSP system. The four loudspeakers are intended to be arranged according to the ITU-R BS 775-1 2/2 recommendation.

sample in duration, shaped like a (scaled) Kronecker delta function. The IR obtained by measuring the voltage of the microphone output signal actually includes three separate IR's: the mechanical IR of the sound producing device; the acoustic transfer function- affected by both the air between the two locations and by sound reflecting objects in the room; and the electro-mechanical transfer function of the microphone, electronic signal processing and recording system; which is equivalent to a convolution of these three IR's (Vogel and de Vries, 1994).

The IR is affected by the level of the excitation signal due to non-linearities in the mechanical, electronic or acoustic parts involved in the IR measurement (for instance, as shown by Dunn and Hawksford, 1993, an IR measured using loudspeakers is affected in a non-linear way by the signal level). An impulse response can also apply to the time-domain output of a (digital) electronic system when excited with a signal shaped liked a Kronecker delta function. Therefore, to avoid confusion the term *acoustic impulse response* will be used to refer to any impulse response which involves the transmission of the excitation signal through air, as distinguished from a purely electronic IR.

In a recording of a solo musical performance using two microphones, there are three acoustic impulse responses, as summarized in figure 4.3: the inter-microphone impulse response IR_{m1-m2}[3]; and the two impulse responses between the sound source and the two microphones (IR_{S-m1} and IR_{S-m2}). All three IR's can change due to various factors, and these factors can be distinguished as being related to either the sound source or to its surrounding

[3]The inter-microphone impulse response means the time-domain transfer function between the two mike signals. For recordings in this thesis, we will use high-quality, large-bandwidth microphones so the acoustic impulse response between the two diaphragms of the microphones are very similar to the electronic impulse response between signals m_1 and m_2.

environment:

- Movement of the sound source or microphones (Pelorson et al., 1992; de Vries et al., 2001).
- The instrument is not a point-source so there will generally be a different impulse response for different notes which are played (especially for large instruments such as a grand piano or church organ) due to the direction-dependant acoustic radiation pattern of the instrument. If a loudspeaker is used to create the excitation signal, the radiation pattern of the loudspeaker will affect the measured IR (Pelorson et al., 1992).
- Air turbulence (Ueda and Ando, 1997; Blesser, 2001) and temperature variations (Morse and Ingard, 1968; Elko et al., 2002) within the recording environment will affect all three impulse responses.
- Physical changes in room boundary surfaces and moving objects (rotating fans, audience etc).

Clearly, the first two factors which affect the acoustic IR's in the above list are source-related and the second two are environment related.[4] These factors will be investigated later with a real-time system, however, the algorithm for the ASUS will be described for time-invariant IR's and stationary source signals. The word *stationary* here means that the statistical properties of the microphone signals (such as mean and autocorrelation) are invariant over time- i.e. they are both *strictly stationary* and *wide sense stationary* (Papoulis and Pillai, 2002). Of course, when dealing with live musical instru-

[4]Movement of the musician is really a source and environment factor, as this will affect all IR's due to a change in the source spatial radiation and a change in the sound absorbing surfaces in the concert hall.

ments the signals at the microphones are non-stationary; it will be shown later how time-varying signals such as recorded music affect the performance of the algorithm. Finally, for the time-being any sound in the room which is caused by sources other than our single source S is ignored; that is, a noise-free (or at least, very low noise) acoustic and electronic environment is assumed. For the foregoing analysis in this section, these three major assumptions are summarized:

- Time invariant IR.
- Stationary source statistics.
- Noise-free operating environment.

The time-domain acoustic transfer function between two locations in an enclosed space- such as between a sound source and a microphone diaphragm- can be modeled as a two-part IR (Polack, 1993; Blesser, 2001; Avendano and Jot, 2004). The two-component IR model is summarized in figure 4.5.

In this model, the L-length acoustic IR is represented as two decaying time sequences; one of which is defined between sample times $n = 0$ and $n = L_r - 1$, the other between $n = L_r$ and $n = L$. The first of these sequences represents the IR from the direct sound and early-reflections (ER's), and the other sequence represents the reverberation; accordingly called the "direct-path" and "reverberant-path" components of the IR. In acoustical terms, reflected sound can be thought of as consisting of two parts: early reflections (ER's) and reverberation (reverb). ER's are defined as "those reflections which arrive at the ear via a predictable, non-stochastic directional path, generally within 80 ms of the direct sound" (Beranek, 1996) whereas reverberation is generally considered to be sound reflections impinging on a point (e.g. microphone)

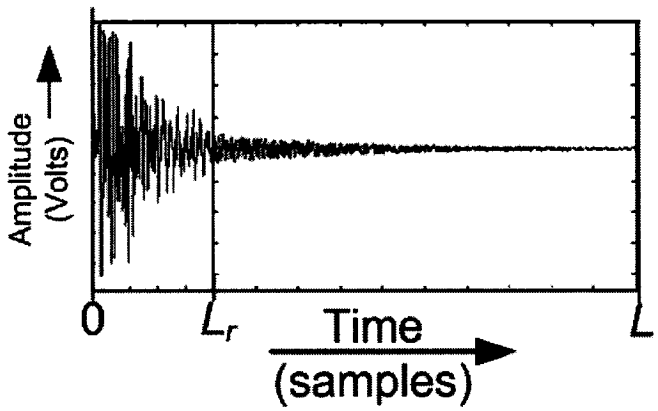


Figure 4.5: Two-component acoustic impulse response model for the time-domain transfer function between two locations in a room. The first part up to the *mixing time* L_r (about 80 ms for a typical concert hall) is called the early or direct component and consists of the direct sound and early reflections, which primarily affects source imagery. The last part is called the reverberant component, which primarily affects reverberance imagery and is modeled as a decaying noise sequence.

which can be modeled as a *stochastic process*, with a Gaussian distribution and a mean of zero (Schroeder, 1987; Blesser, 2001).

The source signals involved in the described filtering processes are also modeled as discrete-time *stochastic processes* (Papoulis and Pillai, 2002). This means a random process whose time evolution can (only) be described using probabilistic laws; it is not possible to define exactly how the process will evolve once it has started, but it can be modeled according to a number of statistical criteria.

As discussed; it is the direct-component of the IR which primarily affects source imagery, such as perceived source direction, width and distance, and the reverberant-component which affects reverberance imagery, such as envelopment and feeling for the size of the room (Barron and Marshall, 1981;

Morimoto, 2001; Soulodre et al., 2003). The time boundary between these two components is called the *mixing time*: "*The mixing time defines how long it takes for there to be no memory of the initial state of the system. There is statistically equal energy in all regions of the space [in the concert hall] after the mixing time [creating a diffuse sound field]*" (Blesser, 2001). The mixing time is approximated by (4.3) (Reichardt and Lehmann, 1978; Polack, 1993):

$$L_r \approx \sqrt{V} \qquad \text{(ms)}, \tag{4.3}$$

where V is the volume of the room (in m^3).

The mixing time can also be defined in terms of the local statistics of the impulse response. Individual, late-arriving sound reflections in a room impinging upon a point (say, a microphone capsule) will give a pressure which can be modeled as being statistically independent from each other; that is, they are independent identically distributed (IID). According to the central limit theorem, the summation of many IID signals gives a Gaussian distribution (Nelson, 1995; Papoulis and Pillai, 2002). The distribution can therefore be used as a basis for determining the mixing time, and two methods for empirically investigating this are presented in section 4.2.3.

After establishing the two-component acoustic IR model, the input signals $m_1(n)$ and $m_2(n)$ can be described by the acoustic convolution between the sound source $s(n)$ and the L_r-length direct-path coefficients summed with the convolution of $s(n)$ with the $(L - L_r)$-length reverberant-path coefficients. Of course, this convolution is undertaken acoustically but to simplify the mathematics we will consider that all signals are electronic as if there is a direct mapping of pressure to voltage, sampled at time (n). Furthermore, for simplicity the two microphone signals m_1 and m_2 are not referred to

explicitly, instead each system is generalized using the subscripts i and j, where i or $j = 1$ or 2 and $i \neq j$. This convolution can therefore be written as:

$$m_i(n) = \sum_{k=0}^{L_r-1} s(n-k)d_{i,k} + \sum_{l=L_r}^{L} s(n-l)r_{i,l}, \quad i = 1 \text{ or } 2. \tag{4.4}$$

A vector formulation of the convolution in (4.4) is now developed, as vector representations of discrete summations are visually more simple to understand and will be used throughout this chapter to describe the ASUS. In-keeping with convention, vectors will always be represented as bold text, contrasted with the italic text style used to represent discrete signal samples in the time-domain.

As mentioned, the direct-path IR coefficients are the first L_r samples of the L-length IR between the source and two microphones, and the reverberant-path IR coefficients are the remaining $(L - L_r)$ samples of these IR's. The time-varying source samples and time-invariant IR's are now defined as the vectors:

- $\mathbf{s}_d(n) = [s(n), s(n-1), ..., s(n-L_r+1)]^T$.
- $\mathbf{s}_r(n) = [s(n-L_r), s(n-L_r-1), ..., s(n-L)]^T$.
- $\mathbf{d}_i = [d_{i,0}, d_{i,1}, ..., d_{i,L_r-1}]^T$.
- $\mathbf{r}_i = [r_{i,0}, r_{i,1}, ..., r_{i,L-L_r-1}]^T$.

And the acoustic convolutions between the radiated acoustic source and

the early and reverberant-path IR's in (4.4) can now be written as:

$$m_i(n) = \mathbf{s}_d^T(n)\mathbf{d}_i + \mathbf{s}_r^T(n)\mathbf{r}_i. \tag{4.5}$$

For convenience, the early and reverberant path convolutions are replaced with:

$$\begin{aligned} s_{di}(n) &= \mathbf{s}_d^T(n)\mathbf{d}_i \\ &\text{and} \\ s_{ri}(n) &= \mathbf{s}_r^T(n)\mathbf{r}_i. \end{aligned} \tag{4.6}$$

So (4.5) becomes:

$$m_i(n) = s_{di}(n) + s_{ri}(n). \tag{4.7}$$

4.2.2 Signal assumptions

With the following definitions for the last L samples of the early and reverberant path sound arriving at time n:

- $\mathbf{s}_{di}(n) = [s_{di}(n), s_{di}(n-1), ..., s_{di}(n-L+1)]^T$
- $\mathbf{s}_{ri}(n) = [s_{ri}(n), s_{ri}(n-1), ..., s_{ri}(n-L+1)]^T$,

the following assumptions about these early and reverberant path sounds are expressed using the statistical expectation operator $E\{\cdot\}$.[5]

[5]Rather than a discrete-time summation to represent properties of the signals discussed in this chapter, *expectations* of stochastic signals are generally used as a visual simplifica-

- The early part of both IR's ("direct-path") are at least partially correlated:
  $E\left\{\mathbf{d}_i^T(n)\mathbf{d}_j(n)\right\} \neq 0,$
  $E\left\{\mathbf{s}_{di}^T(n)\mathbf{s}_{dj}(n)\right\} \neq 0.$

- The late part of each IR (the "reverberant path") are uncorrelated with each other:
  $E\left\{\mathbf{r}_i^T(n)\mathbf{r}_j(n)\right\} = 0,$
  $E\left\{\mathbf{s}_{ri}^T(n)\mathbf{s}_{rj}(n)\right\} = 0.$

- The two reverberant path IR's are uncorrelated with both early parts:
  $E\left\{\mathbf{r}_i^T(n)\mathbf{d}_i(n)\right\} = 0,$
  $E\left\{\mathbf{s}_{ri}^T(n)\mathbf{s}_{di}(n)\right\} = 0.$

- The reverberant path IR is decaying random noise with a normal distribution and a mean of zero:
  $E\left\{\mathbf{r}_i(n)\right\} = 0,$
  $E\left\{\mathbf{s}_{ri}(n)\right\} = 0.$

4.2.3 Validity of assumptions

So far, the effect of room modes or resonances has been ignored. These occur due to reflections normal to the room boundaries which constructively interfere with itself. At low frequencies, the distance between the maxima of the modes will be large (i.e. the wavelength λ). If the level of these modes (i.e. sound pressure level) is large, and there are no other modes of the same frequency overlapping the measurement point, then the response at

tion. This is in accordance with the mean ergodic theorem (Haykin, 2001, pg. 37), which allows the generalization of the local time averaged mean to an ensemble average across the entire process.

that particular frequency will be sinusoidal (i.e. not a Gaussian distribution) and therefore the stochastic model does not apply. This was first discussed by Schroeder (1987) (originally published in 1954); *"the mean spacing of the normal modes must be small compared to the half-power [-3 dB point] width of an individual resonance"*. Schroeder calculated (and later empirically verified; Schroeder and Kuttruff, 1962) that the frequency $f_{Schroeder}$ where we can assume a high modal overlap, and therefore a stochastic model of reverberation, is related to the -60 dB reverberation time RT_{60} as shown in (4.8):

$$f_{Schroeder} > 2000\sqrt{\frac{RT_{60}}{V}} \quad \text{(Hz)}. \tag{4.8}$$

Therefore, the stochastic model for the reverberant component of the IR is only valid after the mixing time and for frequencies above the Schroeder frequency; as summarized in figure 4.6.

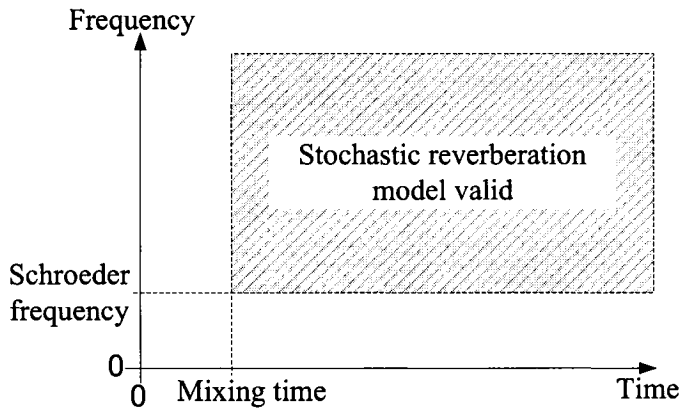


Figure 4.6: Validity domain of the stochastic IR model in the time-frequency plane (hatched). The time axis corresponds to the IR sample time. For Pollack Hall, the Schroeder Frequency is about 80 Hz and the mixing time according to (4.3) is about 40 ms. The figure is adapted from the model presented by Jot and Chaigne (1997).

The pressure-correlation between two locations in a concert hall, e.g. measured using a microphone pair, is unpredictable below the Schroeder frequency: the measurement locations may both be located on a modal maximum (increasing the correlation), or on a maxima and minima (giving a negative correlation), or may be uncorrelated due to modal interference. Above the Schroeder frequency, however, the spatial correlation function in a reverberant (diffuse) sound field can be predicted very accurately according to (4.9) (Cremer and Müller, 1982; Jacobsen and Roisin, 2000):

$$\kappa = \frac{\sin(kx)}{kx}, \tag{4.9}$$

where the correlation between the two locations (κ) is dependant on the wavenumber[6] k and the distance between them x. The resulting *slit* function is summarized in figure 4.7, where it can be seen that the assumption that the reverberant-path components are uncorrelated; i.e.

$E\left\{\mathbf{s}_{ri}^{T}(n)\mathbf{s}_{di}(n)\right\} = 0$

for a typical recording with a 20 cm spaced microphone pair, is only valid for frequencies greater than approximately 1 kHz.

As mentioned, according to the central limit theorem (also called the de Moivre-Laplace integral theorem; Papoulis and Pillai, 2002) the distribution of the sum of a large number of IID random variables is Gaussian shaped: a so-called *normal* distribution.[7] We define reverberation as a physical phenomenon where the density of sound reflections impinging on a point is such that the pressure can be modeled as a stochastic function with a normal distribution and a mean of zero, and that part of an impulse response which has a normal distribution is called the reverberant part. In this section, we will

[6]For a frequency f, the wavenumber $k = 2\pi f/c$, where c is the speed of sound.

[7]Gaussian and normal distributions are not necessarily equivalent, but the distinction is generally ignored in the literature (Tukey, 1977, pg. 623).

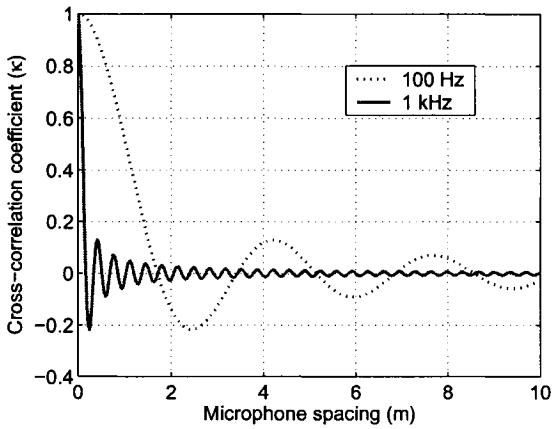


Figure 4.7: Correlation between two locations in a reverberant field, for 100 Hz and 1 kHz sound sources, according to (4.9).

look at acoustic impulse responses and show how the reverberant component can be identified. This is important, as we shall see that the length of the adaptive filter determines the degree to which the source-image components can be removed from the ASUS output channels.

The degree to which a part of an IR has a normal distribution (the "degree of normality") has been used to determine the onset of the reverberation. Two such measures of normality are "average RMS response fluctuation" (Schroeder, 1987) and *kurtosis*. Abel and Berners (2004) calls this first measure *"echo density"*, although this term is not a direct measure of echo density (i.e. the average number of reflections per second; Kutruff, 1991), it still gives a value which is related to this. Schroeder (1987) referred to this as "a response fluctuation" of a frequency response curve measured in a hall using a variable sine-wave reproduced from a loudspeaker. Schroeders' model predicts that 68.26% of the response curve will lie within about 10 dB of the average level. Similarly, the Abel and Berners (2004) criterion for the reverberant part of an IR is based on the principle that 68.26% of the obser-

vations of a normal population will be found within 1 standard deviation of the mean (Schroeder, 1987; Papoulis and Pillai, 2002). So "echo density" is another measure for normality in the IR, defined as the number of samples in a local IR sample which are greater than a standard deviation from the sample mean (i.e. $\pm$ 1 SD). An example of the time-domain response fluctuation (i.e. what Abel and Berners (2004) calls "echo-density") for a concert hall is shown in figure 4.10.

Another measure of the degree of normality is kurtosis. Kurtosis is a measure of the degree of peakedness of a distribution, or the degree of bimodality (Darlington, 1970). Kurtosis is calculated according to (4.10), a normalized form of the fourth central moment of a distribution:

$$kurtosis = \frac{E\{x-\mu\}^4}{\sigma^4}, \tag{4.10}$$

where $E\{\cdot\}$ is the statistical expectation operator, μ is the mean and σ the standard deviation of x. If there are a few very large samples in the series of x then the kurtosis will be high (*leptokurtic*); a flat distribution- a low peakedness- will have a kurtosis of less than 1 (*platykurtic*); and a normal distribution has a kurtosis of 3 (Darlington, 1970). Although kurtosis has not been used before in such a context as for determining the reverberant part of an IR, the reverberant path IR is defined by the kurtosis being approximately equal to 3. The running-kurtosis for the measured impulse responses in figures 4.8 are shown in figure 4.11.

Impulse responses used to investigate how the early ("direct") and reverberant components of an IR can be identified were measured in Pollack Hall at McGill University[8] to test the validity of assumptions regarding the early

[8]Physical details about this 600 seater, 2000 m^3 hall are shown in appendix B.2.

and reverberant components; specifically regarding the statistical properties used to discriminate the two parts. Furthermore, the measurements provide useful data which will be used for electroacoustic and subjective evaluation of the ASUS in later chapters.

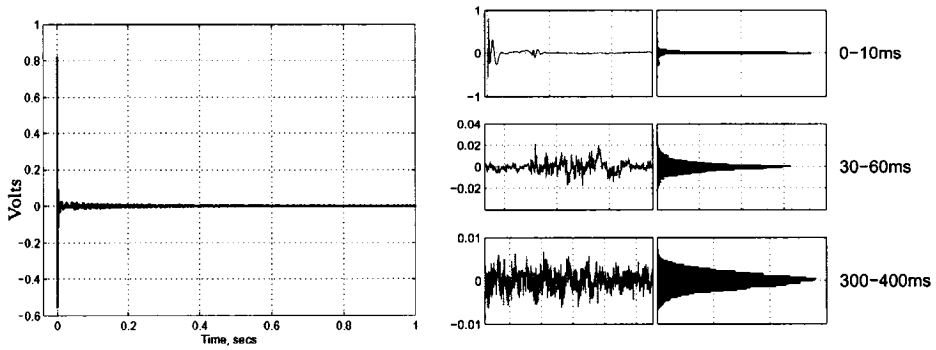
Two IR measurements techniques were investigated: using maximum-length-sequence (MLS) and dirac excitation signals. The signal was reproduced using a single loudspeaker (manufactured by Genelec; model 1032) on the stage and recorded using a single B&K type 4011 (cardioid) microphone facing the loudspeaker 3.5 m away on a 3 m high mike stand (as shown in figure 5.2) with a Grace amplifier on to a Pro-tools computer recording system; D/A and A/D conversion was effected with 16-bit precision at a sampling frequency of 44.1 kHz (as with all measurements in this thesis).

For the MLS measurements, a 17^{th} order (approximately 3 second) excitation signal was reproduced from a single loudspeaker and gave 90 dB SPL (slow weighted) measured at the base of the microphone stand. 45 periods of the MLS signal were recorded, with the final 40 used for the cross-correlation analysis. For the dirac analysis, a 0.2 μs pulse (one sample at 44.1 kHz fs) was reproduced with a loudspeaker that was deliberately overloaded (a signal overload light showed) to get a high SPL (95 dB peak). This was primarily to get a high signal to noise ratio, but as mentioned by Vanderkooy (1994); "The distortion during such a short pulse is somewhat irrelevant, serving only perhaps to renormalize the excitation level, since the total mechanical energy imparted to the driver is manageable, and the resulting impulse response is unaffected by distortion."

Looking at the time-domain plots from the dirac and MLS excitation methods (figure 4.8 and 4.9) we see that the temporal decay and energy

distribution is very similar- especially for the early "direct component". It should be noted that impulse response measurements for the calculation of the reverberation time (RT) in an enclosed space are usually undertaken with an omni-directional microphone (Schroeder, 1965) and calculated as the time taken for the impulse response level to decay by 60 dB (the T60) or 30 dB (T30) from the maximum IR level. Therefore, the decay times for the measurements with the cardioid microphones shown here give a shorter RT than the "true" hall RT as the sound contribution from the rear of the hall was weighted less (using an omni-directional mike, the T30 of Pollack Hall was measured to be 2.3 s at 63 Hz and 1.8 s at 1 kHz).

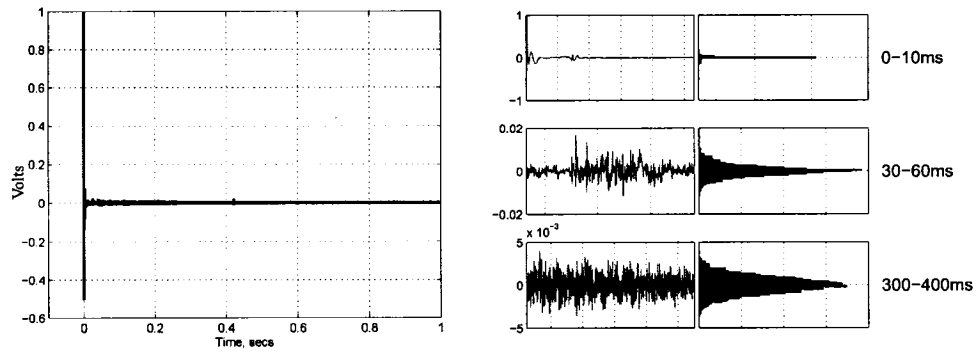
The dirac and MLS excitation methods show some differences between the "echo-density" and kurtosis measures, as shown in figures 4.10 and 4.11. The "echo-density" method predicts a longer mixing time than the kurtosis method; about 100 ms and 80 ms (respectively). Strong individual reflections can be identified easier with the MLS method than the dirac method; maybe because of the low level of low-frequency energy introduced into the hall with the dirac method, and these later echoes would be predominantly low-frequency due to air absorption which is significant for such a long reflection path with an absorption coefficient of about 0.1 dB per metre at 4 kHz but only 0.001 dB/m at 100 Hz (Kinsler et al., 1999, pg. 224).



(a) Time-domain response (voltage) for single IR with dirac excitation.

(b) Zoomed-in view of 1 IR and local PDF from 40 repeated impulses.

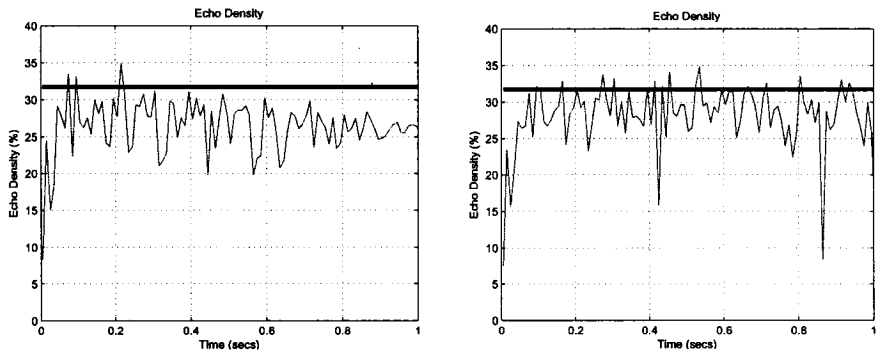
Figure 4.8: IR measurement in Pollack concert hall obtained with single cardioid microphone after excitation using dirac signal reproduced with a loudspeaker.



(a) Time-domain IR obtained by MLS excitation measurement (40 averages).

(b) Zoomed-in view and local PDF.

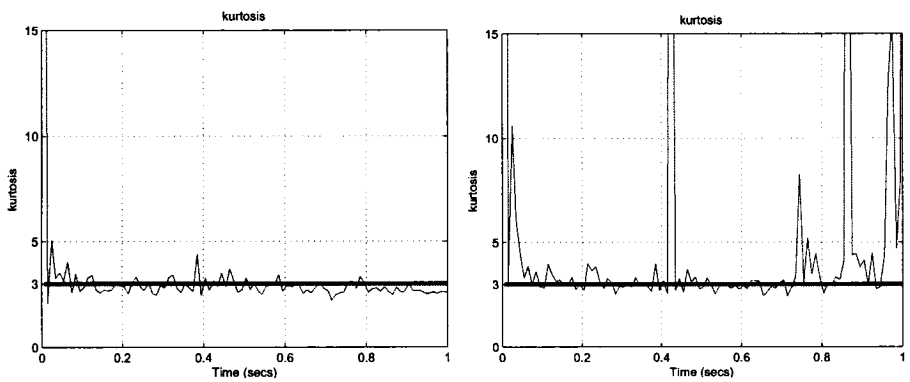
Figure 4.9: IR statistics from 40 repetitions of 17^{th} order (3 seconds) MLS excitation in Pollack hall. Also shown in (b): section of IR and PDF.



(a) IR "echo density" for dirac excitation.

(b) IR Echo density for MLS excitation.

Figure 4.10: "Echo density" of IR's with dirac and MLS measurement techniques. "Echo density" is a measure of the degree of normality for a local region of the IR (here; a 10 ms rectangular window). It is the percentage of samples in the region which are outside a standard deviation from the mean. For a data sequence normal distribution, this value will be approximately 32%.



(a) IR Kurtosis for dirac excitation.

(b) IR Kurtosis for MLS excitation.

Figure 4.11: Kurtosis: a measurement of the degree of peakedness of a distribution. A normal distribution has a kurtosis of 3. The local kurtosis was calculated using a 10 ms non-overlapping rectangular window.

4.3 The Adaptive Filter

The adaptive filter which can accomplish the design criteria described in the beginning of this chapter is now introduced. To reiterate; we want the ASUS to operate in such a way that diagonally opposite loudspeaker signals are uncorrelated; i.e. the rear-right loudspeaker channel is uncorrelated with the front-left and the rear-left with the front right. In other words; the output error signal e_i affected by adaptive filter $\mathbf{w}_{ij}$ must be uncorrelated with the input signal which is *not* processed by this filter, m_j. The procedure for updating the FIR adaptive filter so as to accomplish this is developed according to the principle of orthogonality which shall be explained shortly.

4.3.1 The Least Means Squares algorithm

Each input signal m_1 and m_2 is filtered by an M-sample length filter ($\mathbf{w}_{21}$ and $\mathbf{w}_{12}$, respectively). As mentioned, these filters model the early component ("direct sound component") of the impulse response between the two microphone signals; so ideally the filter length is equal to the mixing time (i.e. $M{=}L_r$). However, for the foregoing analysis we don't make any assumptions about "knowing" L_r a priori, so we will just call the time-domain filter size M. A delay is added to input channel m_i before the filtered signal y_i is subtracted; this is to allow for non-minimum phase impulse responses which can occur if the sound source is closer to one microphone than the other. However, for the foregoing analysis we will not consider this delay as it makes the mathematical description more straight-forward (and it would make no difference to the theory if it were included).

The filtering of signal m_j by the adaptive filter $\mathbf{w}_{ij}$ gives signal $y_i(n)$. This

subscript notation may seem confusing, but helps describing the loudspeaker output signals because signal m_i and e_i are both phase-coherent (have a non-zero correlation) and are reproduced by loudspeakers on the same side (e.g. signals m_1 and e_1 are both reproduced with loudspeakers on the same side). This filtering process is summarized as the discrete time linear convolution in (4.11):

$$y_i(n) = \sum_{k=0}^{M-1} m_j(n-k) w_{ij,k}, \tag{4.11}$$

which with the following definitions:

- $\mathbf{m}_j(n) = [m_j(n), m_j(n-1), ..., m_j(n-M+1)]^T$
- $\mathbf{w}_{ij} = [w_{ij,0}, w_{ij,1}, ..., w_{ij,M-1}]^T$

allow the convolution to be written in vector form as:

$$y_i(n) = \mathbf{m}_j^T(n)\mathbf{w}_{ij}. \tag{4.12}$$

If we look at filter $\mathbf{w}_{12}$ in figure 4.3, it is seen that the filtered m_2 signal, y_1 is subtracted from the unfiltered m_1 signal (sample-by-sample) to give the error signal e_1:

$$e_i(n) = m_i(n) - y_i(n). \tag{4.13}$$

The output signal is conventionally called an error signal as it can be interpreted as being a mismatch between y_i and m_i caused by the filter coefficients $\mathbf{w}_{ij}$ being "not-good enough" to model m_i as a linear transformation

of m_j; these terms are used for the sake of convention and these two error signals are the output signals of the system which are reproduced with separate loudspeakers behind the listener.

If the filter coefficients $\mathbf{w}_{ij}$ can be adapted so as to approximate the early part of the inter-microphone impulse response, then the early-correlated sound component will be removed and the "left-over" signal will be the reverberant (or reverberance-image) component in the m_j channel, plus a filtered version of the reverberant component in the m_i channel. In this case, the error signal level will be smaller than the original level of m_j. The "goal" of the algorithm which changes the adaptive filter coefficients can therefore be interpreted as to minimize the level of the error signals. This level can simply be calculated as a power estimate of the output signal $\mathbf{e}_i$, which is an average of the squares of the individual samples, and it is for this reason that the algorithm is called the Least Mean Square (LMS) *algorithm* (Widrow and Hoff, 1960; Haykin and Widrow, 2003). (In fact, one doesn't really have to *average* the error signal output; we can simply take the sum of squares, which is why the French call this algorithm the "sum of squares" algorithm.) This goal is formally expressed as a "performance index" or "cost" scaler J, where for a given filter vector $\mathbf{w}_{ij}$:

$$J_i(\mathbf{w}_{ij}) = E\left\{e_i^2(n)\right\}, \tag{4.14}$$

and $E\{\cdot\}$ is the statistical expectation operator. The requirement for the algorithm is to determine the operating conditions for which J attains its minimum value: this state of the adaptive filter is called the "optimal state" (Haykin, 2001).

When a filter is in the optimal state, the rate of change in the error signal

level (i.e. J) with respect to the filter coefficients $\mathbf{w}$ will be minimal. This rate of change (or gradient operator) is a M-length vector ∇, and applying it to the cost function J gives:

$$\nabla J_i(\mathbf{w}_{ij}) = \frac{\partial J_i(\mathbf{w}_{ij})}{\partial \mathbf{w}_{ij}(n)}. \tag{4.15}$$

The right-hand-side of (4.15) is expanded using partial derivatives in terms of the error signal $e(n)$:

$$\frac{\partial J_i(\mathbf{w}_{ij})}{\partial \mathbf{w}_{ij}(n)} = 2E\left\{\frac{\partial e_i(n)}{\partial \mathbf{w}_{ij}(n)} e_i(n)\right\}, \tag{4.16}$$

and the general solution to this differential equation, for any filter state, can be obtained by first substituting (4.12) into (4.13):

$$e_i(n) = m_i(n) - \mathbf{m}_j^T(n)\mathbf{w}_{ij}(n) \tag{4.17}$$

and then differentiating with respect to $\mathbf{w}_{ij}(n)$:

$$\frac{\partial e_i(n)}{\partial \mathbf{w}_{ij}(n)} = -\mathbf{m}_j(n). \tag{4.18}$$

This gives a solution for the differential on the right-hand-side of (4.16):

$$\nabla J_i(\mathbf{w}_{ij}) = -2E\left\{\mathbf{m}_j(n) e_i(n)\right\}. \tag{4.19}$$

Updating the filter vector $\mathbf{w}_{ij}(n)$ from time $n-1$ to time n is done by multiplying the negative of the gradient operator by a constant scaler

μ. The expectation operator in equation (4.19) is replaced with a vector multiplication and the filter update is:

$$\mathbf{w}_{ij}(n) = \mathbf{w}_{ij}(n-1) + \mu \mathbf{m}_j(n) e_i(n). \tag{4.20}$$

It should be noted that the adaptive filtering algorithm which is used (i.e. based on the LMS algorithm) is chosen because of its relative mathematical simplicity compared with others (such as the affine projection; Gay, 2000 or RLS; Haykin, 2001 algorithms), yet it will be shown in the next two chapters that it is powerful enough to satisfy both the subjective and electronic design criteria.

4.3.2 The Normalized Least–Mean–Square algorithm

From the filter update equation (4.20) it can be seen that the adjustment from $\mathbf{w}_{ij}(n-1)$ to $\mathbf{w}_{ij}(n)$ is proportional to the filtered input vector $\mathbf{m}_j(n)$. When the filter has converged to the optimal solution, the gradient ∇ in (4.15) should be zero but the actual ∇ will be equal to $\mu \mathbf{m}_j(n) e_i(n)$. This product may be not equal to zero and results in *gradient noise* (Widrow and McCool, 1976) which is proportional to the level of $\mathbf{m}_j(n)$. This undesirable consequence can be mitigated by normalizing the gradient estimation with another scaler which is inversely proportional to the power of $\mathbf{m}_j(n)$, and the algorithm is therefore called the Normalized Least–Mean–Square (NLMS)

algorithm (Slock, 1993). The tap-weight adaptation is then:

$$\mathbf{w}_{ij}(n) = \mathbf{w}_{ij}(n-1) + \frac{\alpha}{\delta + \mathbf{m}_j^T(n)\mathbf{m}_j(n)}\mathbf{m}_j(n)e_i(n), \tag{4.21}$$

with

$$0 < \alpha < 1.$$

When the input signals $\mathbf{m}_1(n)$ and $\mathbf{m}_2(n)$ are very small, inverting the power estimate could become computationally problematic. Therefore a small constant δ is added to the power estimate in the denominator of the gradient estimate- a process called *regularization* (Haykin, 2001, pg. 338). How the regularization parameter affects filter convergence properties is investigated empirically with a variety of input signals in the next chapter.

4.3.3 The Principle of Orthogonality

As mentioned, the "optimal state" is attained when the gradient operator is equal to zero, so under these conditions at sample time n, (4.19) becomes:

$$E\{\mathbf{m}_j(n)e_i(n)\} = \mathbf{0}_{M\times 1}. \tag{4.22}$$

This last statement represents the Principle of Orthogonality (PoO) (Haykin, 2001, pg. 96). The elegant relationship means that when the optimal filter state is attained, referring back to figure 4.4 at the beginning of the chapter, e_1 (the rear-left loudspeaker signal) is uncorrelated with m_2 (the front-right loudspeaker signal). This means that when the adaptive filter is in its opti-

mal solution, diagonally opposite loudspeaker signals are uncorrelated: *Quod Erat Demonstrandum.*

The PoO is summarized graphically in figure 4.12. Under such a condition, distortion of the source image is minimized because signal e_i contains reverberance-image components which are unique to m_i, and as the source image is only affected by correlated components within m_i and m_j (by definition; correlated components within an approximately 20 ms window), then a radiated signal which is uncorrelated with *either* m_i or m_j can not contain a sound component which affects source imagery. This is a very important idea behind the ASUS, and the degree to which the PoO operates will be investigated by measuring both the electronic correlation between signals m_j and e_i and also the subjective differences in auditory spatial imagery of the source image within a conventional 2/0 audio scene and an upmixed audio scene created with the ASUS.

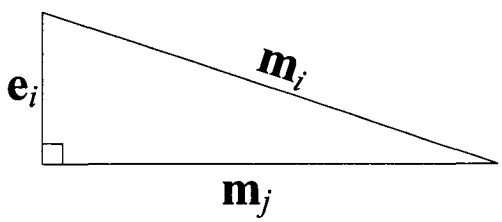


Figure 4.12: Geometrical representation of the principle of orthogonality.

4.3.4 The normal equations

For optimal state conditions, using (4.17) to rewrite (4.22) and then expanding gives:

$$\begin{aligned}\mathbf{0}_{M\times 1} &= E\left\{\mathbf{m}_j(n)e_i(n)\right\} \\ &= E\left\{\mathbf{m}_j(n)\left(m_i(n) - \mathbf{m}_j^T(n)\mathbf{w}_{ij}\right)\right\} \\ &= E\left\{\mathbf{m}_j(n)m_i(n) - \mathbf{m}_j(n)\mathbf{m}_j^T(n)\mathbf{w}_{ij}\right\}.\end{aligned} \tag{4.23}$$

These equations- called the *normal equations* because they are constructed using the equations supporting the corollary to the principle of orthogonality (Haykin, 2001, pg. 393)- can now be written in terms of the correlation between the input signals m_j and m_i, which is called the M-by-1 vector $\mathbf{r}$:

- $\mathbf{r}_{m_j m_i} = E\left\{\mathbf{m}_j(n)m_i(n)\right\}$

and the autocorrelation of each signal is the M-by-M matrix $\mathbf{R}$:

- $\mathbf{R}_{m_i m_i} = E\left\{\mathbf{m}_i(n)\mathbf{m}_i^T(n)\right\}$.

This allows (4.23) to be expressed as:

$$\mathbf{0}_{M\times 1} = \mathbf{r}_{m_j m_i} - \mathbf{R}_{m_j m_j}\mathbf{w}_{ij}. \tag{4.24}$$

The filter in this state is called the *Wiener solution* and the normal equation becomes:

$$\mathbf{w}_{ij} = \mathbf{R}_{m_j m_j}^{-1}\mathbf{r}_{m_j m_i}. \tag{4.25}$$

4.4 Theoretical Performance of the New Upmixer

4.4.1 Performance criteria

In this section various measurements which can be used to evaluate the effectiveness of the new upmixer (i.e. the ASUS) in electroacoustic terms are discussed. From psychoacoustic theory we are able to gain an insight into the subjective implications of these measures.

There are two main reasons why in practical applications there is a difference between the current adaptive filter condition and the filter state which would minimize the error signal level (i.e. the Wiener solution). Firstly, the finite filter update step-size (μ in the LMS algorithm in (4.20), which is affected by α in the NLMS version) ensures that the exact Wiener solution can not be reached and the algorithm executes a random motion (like Brownian motion) around the minimum point on the error performance surface; called *gradient noise* (Widrow and McCool, 1976). Secondly, the Wiener solution itself is nonstationary; due to the factors listed on page 143, such as temperature variations in the recording environment and source movement. There is therefore a *lag* between the optimal (Wiener) solution and actual filter conditions, so it therefore makes more sense to talk about an *optimal* solution rather than a Wiener solution. Also, because of this time variance the filter can be considered *non-linear* - the filter does not obey the principles of *superposition* and *homogeneity* (Haykin, 2001, pg. 258).[9] This is a bit con-

[9]The principle of superposition (POS) can be summarized like this; if the response of a system to an input vector $\mathbf{u}_1(n)$ is $y_1(n)$, and the response to another input $\mathbf{u}_2(n)$ is $y_2(n)$, then the system obeys the POS only if the response to the combined input vector $\mathbf{u}_1(n)+\mathbf{u}_2(n)$ is equal to $y_1(n) + y_2(n)$. The system can be considered linear only if it

fusing, because if we take the filter at a given instance it is a linear FIR-type filter, but it is the time-variation of the optimal solution which makes the filter non-linear.

For stationary or near-stationary inputs, the rate of convergence (RoC) is defined as the number of iterations taken by the update algorithm (4.21) for the filter coefficients $\mathbf{w}_{ij}$ to be "close enough" to the optimal solution in the mean-square sense (Haykin, 2001, pg. 5). Alternatively, the RoC could be taken as the time for the output level to reach a certain level (-20 dB is often used; Elko et al., 2002). The RoC can also be investigated by allowing the filter to converge to (or within a given tolerance of) the optimal solution and then suddenly changing the impulse response (for example, by moving the sound source); a method called "enclosure dislocation" (Mader et al., 2000; Hänsler and Schmidt, 2005). When the optimal solution is constantly changing, the RoC will affect the *tracking* performance. These two important features of the ASUS are summarized as:

- Rate of convergence to optimal solution.
- Tracking of optimal solution.

Considering an optimal solution set of filter coefficients $\mathbf{h}$ and a current set of filter coefficients $\hat{\mathbf{h}}$ then the magnitude of the difference or mismatch between the two can be expressed as a simple dimensionless quantity ξ called the *misalignment* (Benesty et al., 2000a, pg. 106):

$$\xi = \frac{\left|\left|\mathbf{h} - \hat{\mathbf{h}}\right|\right|}{||\mathbf{h}||}, \tag{4.26}$$

obeys the POS and the homogeneity principle. The latter is that the response of the same hypothetical system we just described to a scaled input signal $a\mathbf{u}_1(n)$ must be $ay(n)$.

where $||\cdot||$ denotes the two-norm of a vector.

A difference in the current filter and the required (optimal) filter results in an increased error level. The implication of this is that the filtered input signal and the output error signals are not orthogonal: the PoO (4.22) would not be satisfied. Therefore, there would be components in the error signal which were common to both input signals, which would be manifested by an increased cross-correlation between diagonally-opposite signals (e.g. m_j and e_i). This could affect the perceived spatial character of the source image due to perceptual fusion of correlated sound components between all four loudspeaker signals. The consequence would be that the source image would be spatially *distorted* in the ASUS audio scene compared with a conventional 2/0 loudspeaker pair reproduction of the same two input signals.

Of course, measuring the algorithm performance using misalignment requires that we know the acoustic inter-microphone impulse response a priori. Another method for evaluating algorithm performance is the relative level of the output error signal; conventionally called *misadjustment* (e.g. Widrow and McCool, 1976; Sommen et al., 1987; Elko et al., 2002). The misadjustment is calculated as the relative level of the error signal e_i to the unfiltered input signal m_i. As with misalignment, the misadjustment Ψ_i is a dimensionless quantity (often expressed in dB) as shown in (4.27):

$$\Psi_i = \frac{E\{e_i^2(n)\}}{E\{m_i^2(n)\}}. \tag{4.27}$$

Both the misadjustment and cross-correlation between input and output signals are two very important criteria. Together they give an insight on the formation of side reverberance phantom images created between the un-

filtered input signal and the output error signal. The rate of convergence governs how quickly the adaptive filter responds to movement of the musician or any other artifacts which affects the impulse response between the two microphones. A slow RoC means that there is "leakage" of correlated components to the rear loudspeakers and the source image may be distorted as a consequence.

Other performance criteria relating to practical implementation of the NLMS algorithm are:

- Computational complexity: the number of operations (multiplications and additions) per iteration, and memory requirements.
- Structure of informational flow in the algorithm. This is related to how the input data is sectioned for processing which affects the number of iterations (i.e. algorithm updates) per second and the input-output time latency.

Although these two criteria are always bared in mind throughout this thesis, the bent of the work is towards a thorough understanding of the proposed idea rather than a computationally optimized implementation.

4.4.2 Effect of inter-mike signal correlation on performance

In this section the effect of the inter-microphone correlation on output signal level is investigated with a theoretical mathematical model. This theory is very relevant to predicting how different microphone techniques affect the ASUS performance, and will be empirically verified in chapter 5.

To reiterate: we wish to investigate the misadjustment Ψ_i, which is:

$$\Psi_i = \frac{E\left\{e_i^2(n)\right\}}{E\left\{m_i^2(n)\right\}}, \tag{4.28}$$

in terms of the correlation between the two input signals.

Given:

$$\mathbf{m}_j(n) = [m_j(n), m_j(n-1), \ldots, m_j(n-M+1)]^T, \tag{4.29}$$

$$e_i(n) = m_i(n) - \mathbf{m}_j^T(n)\mathbf{w}_{ij}, \tag{4.30}$$

and the normal equation (see page 165):

$$\mathbf{w}_{ij} = \mathbf{R}_{m_j m_j}^{-1}\mathbf{r}_{m_j m_i}, \tag{4.31}$$

where:

$$\mathbf{r}_{m_j m_i} = E\left\{\mathbf{m}_j(n) m_i(n)\right\}$$

and

$$\mathbf{R}_{m_j m_j} = E\left\{\mathbf{m}_j(n)\mathbf{m}_j^T(n)\right\}.$$

Using (4.30), $e_i^2(n)$ becomes:

$$\begin{aligned} e_i^2(n) &= e_i^T(n)e_i(n) \\ &= m_i^2(n) - 2\mathbf{m}_j^T(n)\mathbf{w}_{ij}m_i^T(n) + \left(\mathbf{m}_j^T(n)\mathbf{w}_{ij}\right)^2, \end{aligned} \tag{4.32}$$

which can be expressed with the expectation operator $E\{\cdot\}$ as:

$$\begin{aligned} E\left\{e_i^2(n)\right\} &= E\left\{e_i^T(n)e_i(n)\right\} \\ &= E\left\{m_i^2(n)\right\} - 2E\left\{\mathbf{m}_j^T(n)m_i^T(n)\right\}\mathbf{w}_{ij} + E\left\{\left(\mathbf{m}_j^T(n)\mathbf{w}_{ij}\right)^2\right\}. \end{aligned} \tag{4.33}$$

Using the previously given definitions for the cross-correlation vector $\mathbf{r}_{m_j m_i}$ and auto-correlation matrix $\mathbf{R}_{m_j m_j}$, (4.33) can be expressed as:

$$E\left\{e_i^2(n)\right\} = \sigma_{m_i}^2 - 2\mathbf{r}_{m_j m_i}\mathbf{w}_{ij} + \mathbf{w}_{ij}^T\mathbf{R}_{m_j m_j}\mathbf{w}_{ij}, \tag{4.34}$$

where σ_{m_i} is the variance of the signal $m_i(n)$:

$$\sigma_{m_i}^2 = E\left\{m_i^2(n)\right\}. \tag{4.35}$$

Substituting the definition of $\mathbf{w}_{ij}$ given in (4.31), we obtain from (4.34):

$$E\left\{e_i^2(n)\right\} = \sigma_{m_i}^2 - 2\mathbf{r}_{m_j m_i}^T\mathbf{R}_{m_j m_j}^{-1}\mathbf{r}_{m_j m_i} + \left(\mathbf{R}_{m_j m_j}^{-1}\mathbf{r}_{m_j m_i}\right)^T\mathbf{r}_{m_j m_i}. \tag{4.36}$$

Considering that

$$\mathbf{r}_{m_j m_i}^T\mathbf{R}_{m_j m_j}^{-1}\mathbf{r}_{m_j m_i} = \left(\mathbf{R}_{m_j m_j}^{-1}\mathbf{r}_{m_j m_i}\right)^T\mathbf{r}_{m_j m_i},$$

(4.36) becomes:

$$E\left\{e_i^2(n)\right\} = \sigma_{m_i}^2 - \mathbf{r}_{m_j m_i}^T\mathbf{R}_{m_j m_j}^{-1}\mathbf{r}_{m_j m_i}. \tag{4.37}$$

Ψ_i, defined in (4.28), can now be obtained by dividing (4.37) by $\sigma^2_{m_i}$:

$$\Psi_i = 1 - \frac{\mathbf{r}^T_{m_j m_i} \mathbf{R}^{-1}_{m_j m_j} \mathbf{r}_{m_j m_i}}{\sigma^2_{m_i}}. \tag{4.38}$$

Now, defining the cross-correlation coefficient between $\mathbf{m}_j$ and m_i as the vector $\mathbf{c}_{m_j m_i}$ (see e.g. Benesty et al., 2000b):

$$\mathbf{c}_{m_j m_i} = \frac{\mathbf{r}_{m_j m_i}}{\sigma_{m_i} \sigma_{m_j}}, \tag{4.39}$$

(4.38) becomes:

$$\begin{aligned} \Psi_i &= 1 - \frac{\sigma_{m_i} \sigma_{m_j} \mathbf{c}^T_{m_j m_i} \mathbf{R}^{-1}_{m_j m_j} \sigma_{m_i} \sigma_{m_j} \mathbf{c}_{m_j m_i}}{\sigma^2_{m_i}} \\ &= 1 - \frac{\sigma^2_{m_i} \sigma^2_{m_j} \mathbf{c}^T_{m_j m_i} \mathbf{R}^{-1}_{m_j m_j} \mathbf{c}_{m_j m_i}}{\sigma^2_{m_i}}. \end{aligned} \tag{4.40}$$

When $m_j(n)$ is white noise, the autocorrelation matrix $\mathbf{R}_{m_j m_j}$ is diagonal such that:

$$\mathbf{R}_{m_j m_j} = \sigma^2_{m_j} \mathbf{I}, \tag{4.41}$$

where $\mathbf{I}$ is the $M \times M$ identity matrix (Haykin, 2001, pg. 358). So for near-white input signals, Ψ_i approximates:

$$\begin{aligned} \Psi_i &= 1 - \frac{\sigma^2_{m_i} \sigma^2_{m_j} \mathbf{c}^T_{m_j m_i} \mathbf{c}_{m_j m_i}}{\sigma^2_{m_i} \sigma^2_{m_j}} \\ &= 1 - \mathbf{c}^T_{m_j m_i} \mathbf{c}_{m_j m_i}. \end{aligned} \tag{4.42}$$

Allowing Ψ_i to be conveniently expressed as a function of the norm of the cross-correlation vector $\mathbf{c}_{m_j m_i}$:

$$\Psi_i = 1 - \left|\left|\mathbf{c}_{m_j m_i}\right|\right|^2. \tag{4.43}$$

In practice, to calculate $\mathbf{r}_{m_j m_i}$ (and $\mathbf{c}_{m_j m_i}$) a running average of the signal products are calculated over time (as shown by Benesty et al., 2000b):

$$\hat{\mathbf{r}}_{m_j m_i} = \sum_{\tau=0}^{T-1} \mathbf{m}_j(n-\tau) m_i(n-\tau). \tag{4.44}$$

The summation range T and the maximum value of $\mathbf{c}_{m_j m_i}$ as a function of τ gives an indication as to the degree of correlation between the signals. If this maximum value is equal to 1, then the signals are considered *coherent* (Cremer, 1976; Blauert, 1997).

To resume this enquiry as to how the correlation between the two mike signals effects the error output level, it is seen that for a given filter the error-signal power is proportional to the power of the unfiltered input signal and inversely proportional to the correlation between the two input signals. This is intuitively reasoned: if the correlation between the two input signals is unity then they would exactly cancel when subtracted. If the cross-correlation $\mathbf{c}_{m_1 m_2}$ was zero, the filter which minimizes the error signal would be a vector of zeros and the error signal would be the same as the unfiltered input signal, with Ψ=1. The relationship between inter-mike cross-correlation and the error-signal level is visualized in figure 4.13. This is empirically validated later in section 5.2.3.

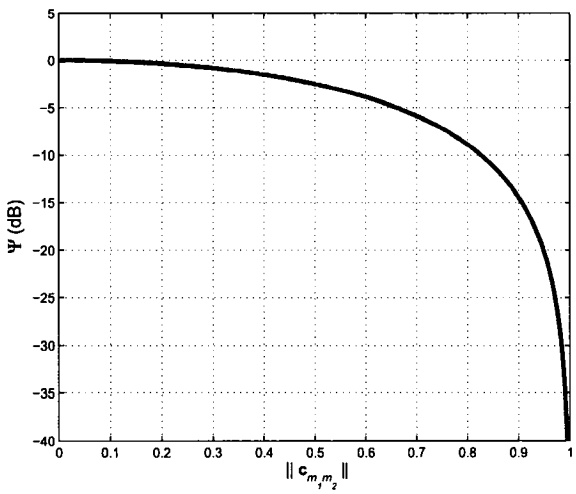


Figure 4.13: The energy of the error signal relative to the unfiltered mike signal (Ψ) is inversely proportional to the double-norm of the cross-correlation vector of the two mike signals $\mathbf{m}_j$ and m_i, ($\mathbf{c}_{m_j m_i}$).

4.4.3 Effect of IR correlation on algorithm performance

The direct-to-reverberant (or reverberant-to-direct) ratio is a commonly used term in acoustics defined as the relative energy densities of the early-arriving and reverberant sound (Kinsler et al., 1999, pg. 342). The distance from the source at which this ratio is equal to unity is called the *reverberation distance* (or radius), and for a room of volume V with reverberation time T, the distance r_h is given by (4.45) (Kutruff, 1991):

$$r_h = 0.1\sqrt{\frac{V}{\pi T}}, \tag{4.45}$$

which for Pollack Hall is about 2 metres.

The early, low-order reflections are sometimes used as part of the direct sound energy calculation, (e.g. the direct-sound energy measure is calculated

over the first 25-80 ms; Reichardt and Lehmann, 1978). Likewise, all sound arriving after the direct sound up to the mixing time is included in the calculation of the direct sound energy and the reverberant-to-direct energy ratio is defined as the energy ratio between this and that part of the sound which arrives via the reverberant path (these two signals are defined on page 148 as s_{di} and s_{ri}, respectively). This is calculated according to (4.46):

$$\gamma_i = \frac{E\left\{s_{ri}^2(n)\right\}}{E\left\{s_{di}^2(n)\right\}}. \tag{4.46}$$

Conveniently, Ψ_i can be expressed in terms of just the reverberant-to-direct ratio γ_i and the direct-path correlation $\mathbf{c}^2_{s_{di}s_{dj}}$ with a simple five-step procedure:

1. Starting with (4.38):

$$\Psi_i = 1 - \frac{\mathbf{r}^T_{m_j m_i}\mathbf{R}^{-1}_{m_j m_j}\mathbf{r}_{m_j m_i}}{\sigma^2_{m_i}},$$

and expanding the cross-correlation vector and autocorrelation vector gives:

$$\Psi_i = 1 - E\left\{\mathbf{m}_j(n)m_i(n)\right\}^T \left(E\left\{m_i^2\right\} E\left\{\mathbf{m}_j(n)\mathbf{m}_j^T(n)\right\}\right)^{-1} E\left\{\mathbf{m}_j(n)m_i(n)\right\}. \tag{4.47}$$

2. Using the definitions for the signals $\mathbf{m}_j(n)$ and $m_i(n)$ in (4.7) allows

(4.47) to be written as:

$$\begin{aligned}\Psi_i = 1 &- E\left\{\left(\mathbf{s}_{dj}(n) + \mathbf{s}_{rj}(n)\right)\left(s_{di}(n) + s_{ri}(n)\right)\right\}^T \\ &\times \left(E\left\{\left(s_{di}(n) + s_{ri}(n)\right)^2\right\} E\left\{\left(\mathbf{s}_{dj}(n) + \mathbf{s}_{rj}(n)\right)\left(\mathbf{s}_{dj}^T(n) + \mathbf{s}_{rj}^T(n)\right)\right\}\right)^{-1} \\ &\times E\left\{\left(\mathbf{s}_{dj}(n) + \mathbf{s}_{rj}(n)\right)\left(s_{di}(n) + s_{ri}(n)\right)\right\}.\end{aligned} \tag{4.48}$$

3. Expanding (4.48), we get:

$$\begin{aligned}\Psi_i = 1 &- \left(E\left\{\mathbf{s}_{dj}(n)s_{di}(n)\right\} + \underline{E\left\{\mathbf{s}_{dj}(n)s_{ri}(n)\right\}} + \underline{E\left\{\mathbf{s}_{rj}(n)s_{di}(n)\right\}} + \underline{E\left\{\mathbf{s}_{rj}(n)s_{ri}(n)\right\}}\right)^T \\ &\times \left(\left(E\left\{s_{di}^2(n)\right\} + \underline{2E\left\{s_{di}(n)s_{ri}(n)\right\}} + E\left\{s_{ri}^2(n)\right\}\right)\right. \\ &\quad \left.\times \left(E\left\{\mathbf{s}_{dj}(n)\mathbf{s}_{dj}^T(n)\right\} + \underline{E\left\{\mathbf{s}_{dj}(n)\mathbf{s}_{rj}^T(n)\right\}} + \underline{E\left\{\mathbf{s}_{rj}(n)\mathbf{s}_{dj}^T(n)\right\}} + \underline{E\left\{\mathbf{s}_{rj}(n)\mathbf{s}_{rj}^T(n)\right\}}\right)\right)^{-1} \\ &\times E\left\{\mathbf{s}_{dj}(n)s_{di}(n)\right\} + \underline{E\left\{\mathbf{s}_{dj}(n)s_{ri}(n)\right\}} + \underline{E\left\{\mathbf{s}_{rj}(n)s_{di}(n)\right\}} + \underline{E\left\{\mathbf{s}_{rj}(n)s_{ri}(n)\right\}}.\end{aligned} \tag{4.49}$$

With the assumptions given in section 4.2.2 that the reverberant and direct components are uncorrelated, and that the late-reverberation is normally distributed noise with a mean of zero, the coloured underlined terms in the above equation can be removed. Furthermore, according to (4.46) we can replace $E\left\{s_{ri}^2(n)\right\}$ with $\gamma_i E\left\{s_{di}^2(n)\right\}$, giving:

$$\begin{aligned}\Psi_i = 1 &- E\left\{\mathbf{s}_{dj}(n)s_{di}(n)\right\}^T \\ &\times \left(E\left\{s_{di}^2(n)\right\}(\gamma_i + 1)\left(E\left\{\mathbf{s}_{dj}(n)\mathbf{s}_{dj}^T(n)\right\}\right)\right)^{-1} \\ &\times E\left\{\mathbf{s}_{dj}(n)s_{di}(n)\right\}.\end{aligned} \tag{4.50}$$

4. As before in (4.41), the following assumption is made, which is an approximation since the relationship only holds for white noise (Haykin, 2001,

pg. 358):

$$\mathbf{R}_{s_{dj}s_{dj}} = \sigma^2_{s_{dj}}\mathbf{I}, \tag{4.51}$$

where $\sigma_{s_{dj}}$ is the variance of the signal $s_{dj}(n)$:

$$\sigma^2_{s_{dj}} = E\left\{s_{dj}(n)\right\}. \tag{4.52}$$

This generalization allows (4.50) to be written as:

$$\begin{aligned}\Psi_i &= 1 - E\left\{\mathbf{s}_{dj}(n)s_{di}(n)\right\}^T \\ &\quad \times \left(\sigma^2_{s_{di}}\left(\gamma_i + 1\right)\sigma^2_{s_{dj}}\right)^{-1} \\ &\quad \times E\left\{\mathbf{s}_{dj}(n)s_{di}(n)\right\} \\ &= 1 - \frac{E\left\{\mathbf{s}_{dj}(n)s_{di}(n)\right\}^T E\left\{\mathbf{s}_{dj}(n)s_{di}(n)\right\}}{\sigma^2_{s_{dj}}\sigma^2_{s_{di}}\left(\gamma_i + 1\right)}.\end{aligned} \tag{4.53}$$

5. By defining the cross-correlation coefficient vector between the direct-path vector $\mathbf{s}_{dj}(n)$ and the direct-path sample $s_{di}(n)$, which can be calculated using the same approach as in (4.44), as:

$$\mathbf{c}_{s_{di}s_{dj}} = \frac{E\left\{\mathbf{s}_{dj}(n)s_{di}(n)\right\}}{\sqrt{E\left\{\mathbf{s}^2_{dj}(n)\right\}E\left\{\hat{d}^2_i(n)\right\}}}, \tag{4.54}$$

the resulting solution for Ψ_i is found to be dependant on the direct-path IR correlation $\mathbf{c}_{s_{di}s_{dj}}$ and the reverberant-to-direct level γ_i:

$$\Psi_i = 1 - \frac{\left|\left|\mathbf{c}_{s_{di}s_{dj}}\right|\right|^2}{\gamma_i + 1}. \tag{4.55}$$

Combining (4.55) and (4.43), it is seen that the inter-channel correlation $\mathbf{c}_{m_j m_i}$ is proportional to the correlation of the direct-path IR's and inversely proportional to the relative reverberant energy in the IR:

$$||\mathbf{c}_{m_j m_i}||^2 = \frac{||\mathbf{c}_{s_{di} s_{dj}}||^2}{\gamma_i + 1}. \tag{4.56}$$

When the reverberation level is high, the denominator of (4.56) will dominate; the input cross-correlation will be low and the output error level large. This trend is visualized in figure 4.14, where it is seen that when the level of reverberation is 60 dB higher than the direct part, the correlation between the two microphones is approximately zero, irrespective of the correlation between the direct path sound. It can also be seen that when the direct sound level is 30dB greater than the reverb, the overall correlation between the two input signals is dominated by the correlation between the direct path signals s_{di} and s_{dj}. As typical two-microphone recordings are generally made with the mike diaphragms within the reverberant radius (Fukada et al., 1997), the most applicable part of figure 4.14 is to the left-hand side of the central line of the γ_i axis.

4.5 Real-Time Algorithm Design

4.5.1 Block-wise frequency-domain adaptive filtering

In this section the adaptive filtering in the ASUS system is described with functional operations which can be undertaken with audio signals on a computer.

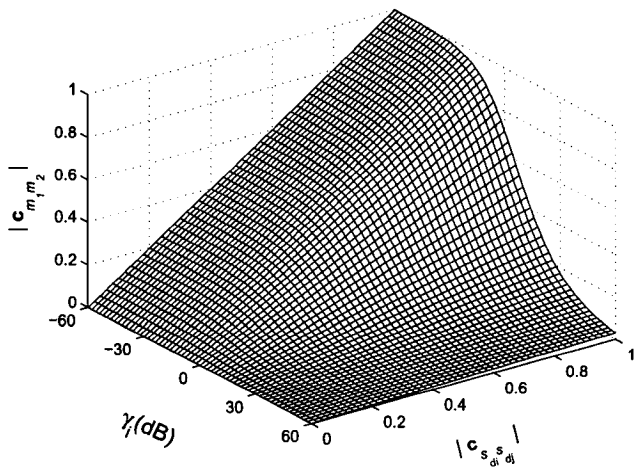


Figure 4.14: Theoretical input signal correlation (max) $\mathbf{c}_{m_j m_i}$ as a function of direct-path correlation $\mathbf{c}_{s_{di}s_{dj}}$ and the reverberant-to-direct ratio γ_i. Typical reverberant-to-direct ratios in two-microphone recordings in concert halls are 0 to -10 dB.

For every iteration of the NLMS algorithm a correlation and convolution of length M is required. Considering the filter $\mathbf{w}_{12}$: First, there is a linear convolution of the current input signal $\mathbf{m}_1(n)$ and the "old" filter coefficients $\mathbf{w}_{12}(n-1)$ (4.12). Second, in the calculation of the new filter coefficient vector $\mathbf{w}_{12}(n)$ (4.21), the gradient estimation requires a linear correlation between the other mike input vector $\mathbf{m}_2(n)$ and the current error output $e_1(n)$. These two calculations are implemented with fewer mathematical operations by transforming N samples of the signals into the frequency domain. When the input data is sectioned into blocks like this, the filter coefficients will remain constant for N samples.

The block convolution and correlation are most efficiently undertaken using the discrete Fourier transform (DFT) (Dentino et al., 1978; Shynk, 1992). This is because in the frequency domain, a convolution or correlation of two time sequences can be derived from a multiplication of two frequency-domain

vectors (Oppenheim and Shafer, 1999), and if we consider only stationary conditions the LMS algorithm is equivalent for the block and non-block formulation (Benesty and Duhamel, 1992). To implement the DFT using the Fast Fourier Transform (FFT) the input data must be sectioned into blocks. As long as we have the condition $1 \leq M \leq N$, the block size N can be different from the filter size M. Although any block length can be used, regardless of the filter length M (Benesty, 2001, chap. 8), the case $N = M$ is the most efficient block size for the FFT algorithms and the relative efficiency (in terms of operations per N) increases in the order of $O(Nlog_2N)$.[10]

4.5.2 Implementation details

Before discussing the convolution and correlation in the NLMS algorithm, the method is explained with the time sequence $x_1(n)$ which is sectioned into N-length blocks, and the stationary N-length filter vector $\mathbf{x}_2$:

- $\mathbf{x}_1(k) = [x_1(n), x_1(n-1), ..., x_1(n-N+1)]^T$.
- $\mathbf{x}_2 = [x_{2,0}, x_{2,1}, ..., x_{3,N-1}]^T$.

[10]Actually, due to cache limitations on a computer CPU, when the FFT block is increased the performance can decrease. For instance, using a freely available FFT software library (FFTW2: see Frigo and Johnson, 2005), on a 550 MHz Pentium III test system a 262144 point FFT was found to be 30 times slower than a 32768 point FFT, which theoretically should be only 10 times slower (Torger, 2001). In another benchmark study (see `http://www.fftw.org/benchfft/` for methodology) it was found that the FFT performance (measured in Mflops/sec) decreased when the block length was greater than 8192 samples. This problem can be overcome by *partitioned convolution* (Torger and Farina, 2001). The ASUS uses block sizes in the order of a few 1000 samples, so we will not look at partitioned convolution techniques.

The convolution of the two sequences for the block length N is:

$$\begin{aligned} x_3(n) &= \sum_{i=0}^{N-1} x_1(n-i)x_{2,i} \\ &\text{and} \\ \mathbf{x}_3(k) &= [x_3(n), x_3(n-1), \ldots, x_3(n-N+1)]^T. \end{aligned} \tag{4.57}$$

In the frequency domain, the signal block $\mathbf{x}_1(n)$ and filter $\mathbf{x}_2$ are expressed with capital letters at block time k:

$$\begin{aligned} \mathbf{X}_1(k) &= \mathsf{diag}\left\{\boldsymbol{F}\left\{\mathbf{x}_1(k-1), \mathbf{x}_1(k)\right\}\right\} \\ \mathbf{X}_2 &= \boldsymbol{F}\left\{\mathbf{x}_2, \underbrace{0, \ldots, 0}_{N}\right\} \\ &\text{and} \\ \mathbf{X}_3(k) &= \mathbf{X}_1(k)\mathbf{X}_2 \\ \mathbf{x}_3(k) &= \boldsymbol{F}^{-1}\mathbf{X}_3(k) \end{aligned} \tag{4.58}$$

where $\boldsymbol{F}$ represents the $2N \times 2N$ Fourier matrix, $\boldsymbol{F}^{-1}$ is the inverse matrix (IFFT), and $\mathsf{diag}\{\cdot\}$ is an operator which forms a diagonal matrix.

The multiplication of two $2N$-length frequency domain vectors $\mathbf{X}_1(k)$ and $\mathbf{X}_2$ is equivalent to a circular convolution of their time domain counter-parts (Oppenheim and Shafer, 1999, pg. 571). A circular convolution differs from a linear convolution in that one of the time sequences is circularly time-reversed and circularly time-shifted with respect to the other sequence. The result of this is that when $\mathbf{X}_3(k)$ is converted back to the time domain, the first half of the new time sequence has data which is corrupted by wraparound "aliasing" (Pelkowitz, 1981). Therefore, only the last N samples in the output block $\mathbf{x}_3(k)$ in (4.58) are equivalent to $\mathbf{x}_3(k)$ as defined in (4.57). This means that

for $2N$ output samples at block time k, the last N old samples of the time sequence $\mathbf{x}_1(k-1)$ are needed as well as N new samples $\mathbf{x}_1(k)$: in other words, an overlap of N coefficients of the filtered signal is required to avoid aliasing error. As the FFT size is $2N$, this is referred to as a 50% overlap, and the filtering style is the *overlap-save* method (Proakis and Manolakis, 1996, pg. 430). The overlapping factor is arbitrary, but will determine how long it takes for the filter to converge to the optimal solution and the tracking ability for non-stationary conditions (that is, the "real-world" time is inversely proportional to the overlap, but the number of iterations require to converge on the optimal solution will remain constant).

Referring back to the convolution in the NLMS algorithm (4.12), the M-sample input signal block $\mathbf{m}_j$ at sample time $n = kN$ is concatenated to form a block size of $2N$ before being converted into the frequency domain:

$$\mathbf{M}_j(k) = \mathsf{diag}\left\{\boldsymbol{F}\left\{\mathbf{m}_j^T(kN-N), \mathbf{m}_j^T(kN)\right\}^T\right\} \tag{4.59}$$

and the $M = N$ length adaptive filter coefficients are zero padded to make a $2N$ length sequence:

$$\mathbf{W}_{ij}(k) = \boldsymbol{F}\left\{\mathbf{w}_{ij}^T(kN), \underbrace{0, \ldots, 0}_{N}\right\}^T. \tag{4.60}$$

And now the convolution between the microphone signal and adaptive filter becomes the frequency domain multiplication:

$$\mathbf{Y}_i(k) = \mathbf{M}_j(k)\mathbf{W}_{ij}(k). \tag{4.61}$$

As a linear convolution is required, the resulting $\mathbf{Y}_i(k)$ block is converting back to the time domain using the IFFT and the last half of the data is retained by multiplying the time sequence by a $N \times 2N$ *window function* $\mathscr{W}_{N\times 2N}^{01}$ (Mansour and Gray, 1982; Clark et al., 1983; Sommen et al., 1987; Benesty and Duhamel, 1992):

$$\begin{aligned} &\mathbf{y}_i(k) = \mathscr{W}_{N\times 2N}^{01}\boldsymbol{F}^{-1}(\mathbf{Y}_i(k)) \\ &\text{where} \\ &\mathscr{W}_{N\times 2N}^{01} = \{\mathbf{0}_{N\times N}\mathbf{I}_{N\times N}\}, \end{aligned} \tag{4.62}$$

is a window function and $\mathbf{I}$ is the $N \times N$ identity matrix. The calculation of the gradient operator $\nabla J_1(\mathbf{w}_{ij})$ in (4.19) involves a linear correlation of the vector $\mathbf{m}_j(n)$ and the error signal $e_i(n)$. As a correlation is just a convolution except neither data sequence is time-reversed, the M sample error signal block $\mathbf{e}(k)$ (which is calculated by subtraction in the time-domain) is augmented with zeros (Shynk, 1992).

Defining:

$$\bullet\mathbf{e}_i(k) = [e_i(n), e_i(n-1), \ldots, e_i(n-N+1)]^T,$$

the time-domain error signal is augmented with N zeros and this new $2M = 2N$ length block is then converted into the frequency domain with a $2N$-point FFT:

$$\mathbf{E}_i(k) = \boldsymbol{F}\left\{\underbrace{0,\ldots,0}_{N}, \mathbf{e}_i^T(k)\right\}. \tag{4.63}$$

And the gradient estimate becomes:

$$\nabla J_i(\mathbf{W}_{ij})(k) = \mathbf{M}_j^H(k)\mathbf{E}_i(k), \tag{4.64}$$

where ^{H} denotes the complex conjugate transpose of a vector (i.e. Hermitian transposition).

Again, (4.64) results in a circular rather than linear behaviour. With a correlation, however, the wraparound error occurs for the *last* N samples of the $2N$ gradient estimate. A window, or *constraint*, is applied in the time-domain by multiplying the result of the IFFT of the gradient with a window function $\mathscr{W}_{2N\times 2N}^{10}$. The result is then converted back into the frequency-domain to give the constrained gradient estimate $\nabla_c J_i(\mathbf{W}_{ij})(k)$:

$$\begin{aligned} \nabla_c J_i(\mathbf{W}_{ij})(k) &= \boldsymbol{F}\mathscr{W}_{2N\times 2N}^{10}\boldsymbol{F}^{-1}\mathbf{M}_j^H(k)\mathbf{E}_i(k), \\ &\text{where:} \\ \mathscr{W}_{2N\times 2N}^{10} &= \begin{bmatrix} \mathbf{I}_{N\times N}\mathbf{0}_{N\times N} \\ \mathbf{0}_{N\times N}\mathbf{0}_{N\times N} \end{bmatrix}. \end{aligned} \tag{4.65}$$

The block-wise frequency domain filter update equation, in the style of the time-domain update equation (4.20), now becomes:

$$\mathbf{W}_{ij}(k) = \mathbf{W}_{ij}(k-1) + \mu\nabla_c J_i(\mathbf{W}_{ij})(k). \tag{4.66}$$

The normalization of the gradient vector is undertaken in the time domain by dividing the gradient with a power estimate of the filtered input signal according to (4.21). According to Parseval's theorem (e.g. Proakis and Manolakis, 1996; Oppenheim and Shafer, 1999), the energy density spectrum

of the signal block $\mathbf{M}_i(k)$ is equal to the time domain power estimate of the composite signals (which consists of two blocks of $\mathbf{m}_i(k)$ (4.59)). The energy density spectrum is simply the square of the (complex) signal spectrum. The power estimate is smoothed over time using a sliding exponential window which weights past power estimates progressively less (Sommen et al., 1987; Hänsler and Schmidt, 2003), so a $2N$ vector normalizes the gradient estimate, with this power estimate vector for input signal m_i defined as:

$$\begin{aligned} \mathbf{P}_i(k) &= \lambda \mathbf{P}_i(k-1) + (1-\lambda)\left|\left|\mathbf{M}_i(k)\right|\right|^2, \\ \text{with} \quad & 0 < \lambda < 1. \end{aligned} \tag{4.67}$$

The scaler λ is the smoothing constant of the power averaging network. Multiplying the gradient estimate normalizing vector $P_i(k)$ by a scaler μ, as shown in (4.68), gives a dynamic frequency-dependent adaptation vector $\boldsymbol{\mu}_i(k)$, which allows the filter taps of $\mathbf{W}_{ij}$ to converge at a constant rate as a function of frequency:

$$\boldsymbol{\mu}_i(k) = \mu \ \ \text{diag}\left\{\mathbf{P}_{i,0}^{-1}(k), \ldots, \mathbf{P}_{i,2N-1}^{-1}(k)\right\}. \tag{4.68}$$

The normalized version of the filter update equation (4.66) now becomes:

$$\begin{aligned} \mathbf{W}_{ij}(k) &= \mathbf{W}_{ij}(k-1) + \boldsymbol{\mu}_i(k)\boldsymbol{\nabla}_c J_i(\mathbf{W}_{ij})(k) \\ &= \mathbf{W}_{ij}(k-1) + \boldsymbol{F}\mathscr{W}_{2N\times 2N}^{10}\boldsymbol{F}^{-1}\boldsymbol{\mu}_i(k)\mathbf{M}_j^H(k)\mathbf{E}_i(k). \end{aligned} \tag{4.69}$$

The filtering process (4.61) and filter update (4.66) involves five order-$2N$ FFT's.[11] By omitting the gradient constraint in (4.65), we have the unconstrained frequency domain filter (Mansour and Gray, 1982) which is

[11] Actually, 2 IFFT's and 3 FFT's, but the FFT and IFFT are equivalent in terms of computational cost (Zolzer, 1997, pg. 155).

computationally cheaper but requires approximately twice the number of filter-update iterations to converge to the optimal solution (Sommen et al., 1987; Lee and Un, 1989).

A computational recipe for the frequency-domain implementation of the NLMS algorithm is summarized graphically in figure 4.15, which is similar to the overlap-save system given by Sommen et al. (1987) and Shynk (1992) though the algorithm used in this thesis is adapted for an arbitrary overlapping factor (Benesty, 2004). The forgetting factor for the power-estimate of the filtered input signal (λ) is calculated according to (4.70) (Benesty, 2001, pg. 166):

$$\lambda = [1 - \frac{1}{6N}]^S, \tag{4.70}$$

where S is the size of the output block ($S = 2N/\alpha$). The choice of algorithm parameters is chosen by an empirical study reported in the next chapter.

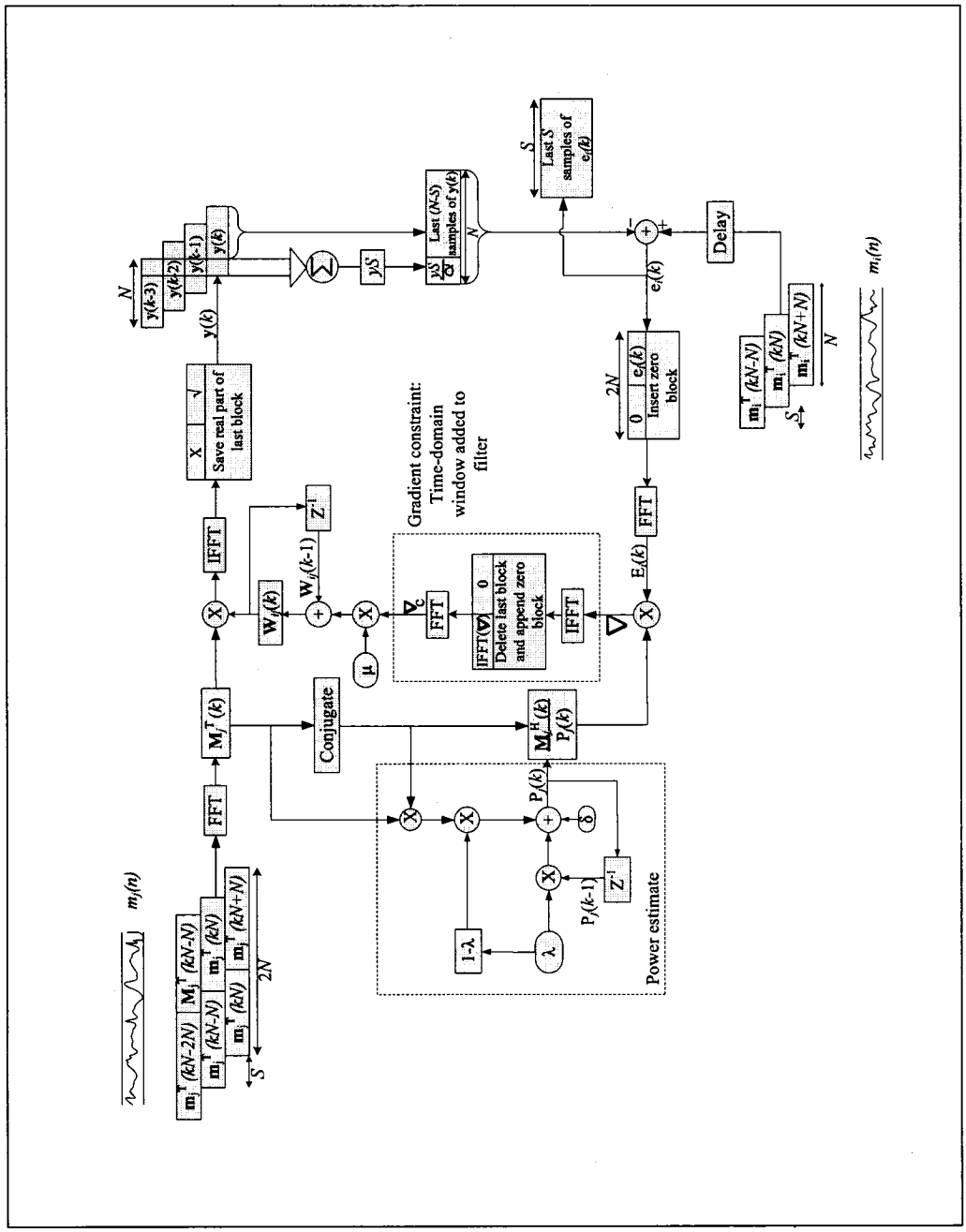


Figure 4.15: Flow chart showing implementation of the adaptive filter using overlap-save sectioning architecture (Benesty, 2004). System inputs are: the two microphone signals $m_i(n)$ and $m_j(n)$; adaptation step μ $(0 < \mu < 1)$; forgetting-factor λ; regularization constant δ; and overlapping factor α (where $\alpha = \frac{N}{S}$). Shown here, $\alpha = 4$. All (I)FFT's are of order $2N$. Z^{-1} represents a delay of one iteration, and the delay for the unfiltered input signal $m_i(n)$ is typically 500 samples.

Chapter 5

Performance of the New System: Electronic Measurements

5.1 Chapter overview

In this chapter a number of investigations are reported which looked at how the electronic performance of the new upmix system is affected by the algorithm parameters and by different recording techniques.

The following parameters of the algorithm were investigated:

- Regularization constant (δ). This is the denominator constant for the calculation of the power estimate in the (frequency domain) NLMS algorithm. It is used for ensuring against computational errors if the

spectral power at a certain frequency is too low, as mentioned in the NLMS algorithm derivation in section 4.3.2.

- Step-size constant (μ). This constant controls the amount by which each frequency tap can change from iteration-to-iteration (it is also normalized by an estimate of the spectral power).

- Overlapping factor (α). This is a fraction of the input block which is overlapped from iteration-to-iteration which affects how many new output samples are created every iteration and how often the filter is updated. Of course, the greater the overlap the more computational power is required (in terms of operations per second).

- Filter size ($2N$). This is the size of each FFT and IFFT, and the block size for the two input channels.

The effect of these parameters on the algorithm performance was measured empirically using two-microphone recordings of a single sound source in a concert hall. Later in the chapter there is a report on a brief comparative study with two commercial upmixers (Circle Surround II and Dolby Pro-Logic II) using "off the shelf" commercial recordings, where output level and correlation properties are analysed.

Performance shall be measured according to a variety of criteria:

- Misadjustment: the energy ratio of the output signal $e_i(n)$ to the unfiltered input $m_i(n)$, as defined in (5.1):

$$\Psi_i = \frac{E\left\{e_i^2(n)\right\}}{E\left\{m_i^2(n)\right\}}. \tag{5.1}$$

- Rate of Convergence: the time taken for the misadjustment to reach a certain level.
- Early misadjustment: misadjustment 1 second after initialization.
- Final misadjustment: misadjustment 30 seconds after initialization.
- Inter-channel correlation properties; cross-correlation between output and input channels of the new upmixer (e.g. front-left speaker and rear-right speaker signals).

5.2 Electroacoustic measurements in a concert hall

In this section, the effect of different audio signals, recording techniques and algorithm parameters on *misadjustment* is investigated. (Misadjustment is defined as the level ratio between the output (rear loudspeaker) signal and the input signals.) As mentioned, for the sake of an appropriately detailed thesis the scope focuses on an analysis of two-microphone recordings of solo musical performances. The recordings in this section were made in Pollack Hall at McGill University and reproduced using a single loudspeaker on the stage. The reason for this approach to creating the test stimuli was to control the study so that only the effect of the algorithm parameters with sources of different temporal-spectral properties can be investigated. Using "live" musicians would have meant that the recordings could not have been repeated at different locations; natural movement would have made identical repeated performances impossible. Also, musical instruments have very different directional radiation characteristics (Caussé et al., 1992), so this factor

was eliminated by using a single loudspeaker to radiate the different audio signals. Furthermore the recordings made here were also used in two listening tests (reported in the next chapter), allowing a comparison between the subjective evaluation and electroacoustic measurements.

5.2.1 Method

Six stimuli were used in this investigation; a white-noise source (30 seconds long) and five musical excerpts. Each of the five music recordings were of a different solo musical instrument recorded in an anechoic chamber, and mixed to a single channel.[1] They were chosen for their variety of temporal and spectral properties- details of these recordings are shown in figure 5.1.

The stimuli were reproduced using a single Genelec 1032 loudspeaker at stage-centre in Pollack Hall (physical details about this 600 seater, 2000 m^3 hall are given in appendix B.2). A photograph of the recording setup which is shown in figure 5.2.

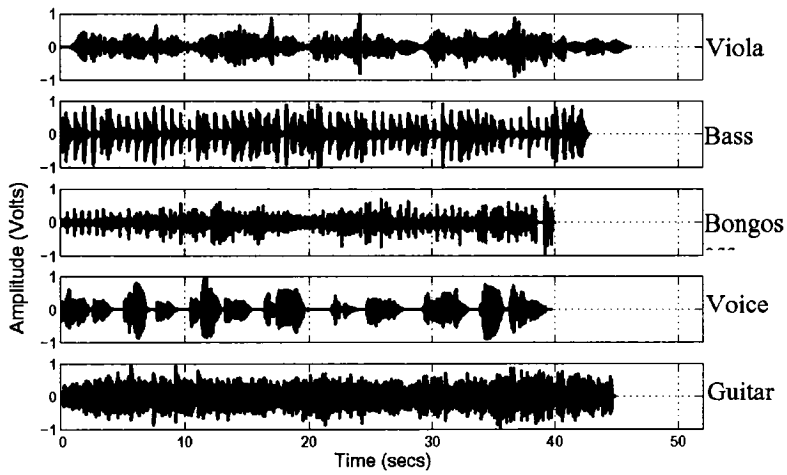
Source configurations:

- Five loudspeaker locations were recorded; equidistant to each microphone ("on-axis", i.e. along the central axis bisecting the stage) and at 1.5 metre intervals to the left and right.

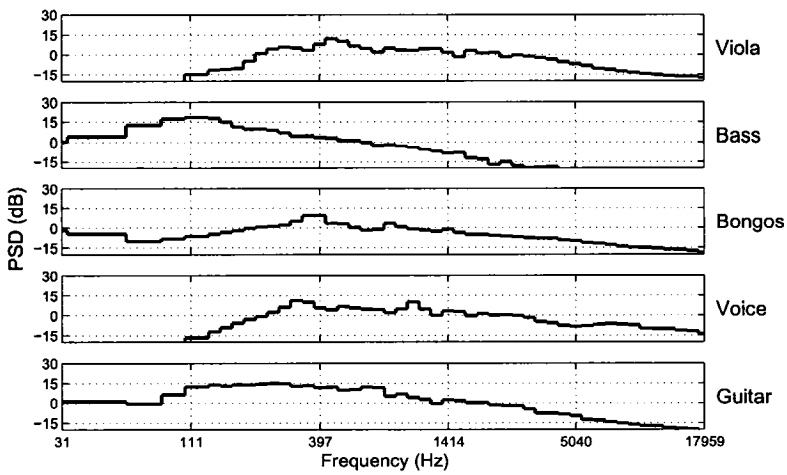
Microphone configurations:

- All recordings used B&K type 4011 cardioid microphones.
- All microphones at the same height and the same distance to the stage.

[1] These recordings were kindly donated by different people from the SURSOUND email list.



(a) Time-domain response of stimuli.



(b) Power spectral density: 1/6 octave smoothing.

Figure 5.1: Details of the stimuli used to create test stimuli for the electronic and subjective experiments with the new upmixer (a white noise source was also used). These anechoic recordings of solo instruments were reproduced with a single loudspeaker on the centre of the stage and recorded using various microphone-pair configurations.

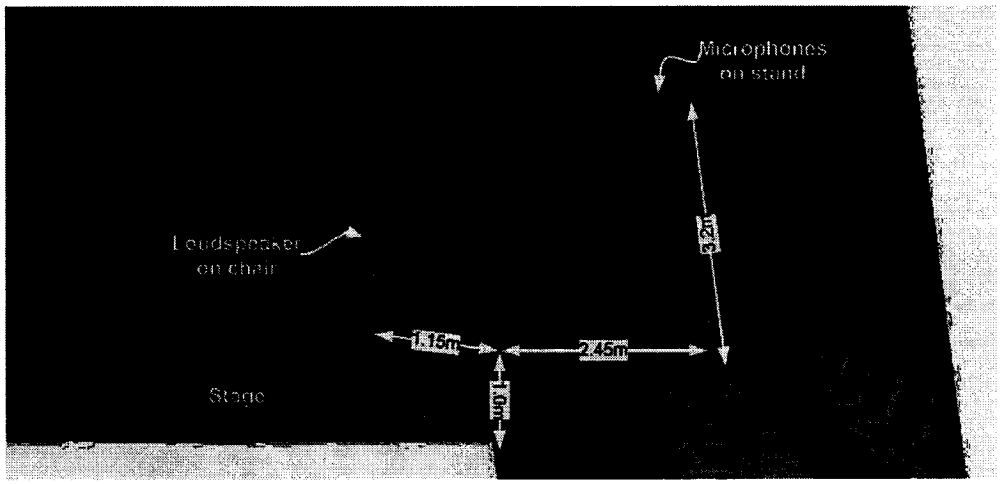


Figure 5.2: Loudspeaker and microphone set-up in Pollack Hall for recordings used to evaluate the new upmix system. Details of the hall are given in appendix B.2.

• Three microphone arrangements were used, which by convention are called:

- AB (50 cm spacing diaphragm, normal to the stage front).

- XY (coincident pair, 90°).

- ORTF (110°, 16 cm spacing of diaphragm).[2]

The recordings were made using a Grace pre-amp direct to a hard-drive (using Pro-Tools), with A/D and D/A conversion undertaken with a 44.1 kHz sample rate at 16-bit resolution.

For each of the three recording techniques, the left and right channel of the on-axis source recordings were normalized.[3] This ensured that for the

[2]The ORTF technique is so-called because it was developed by the French national radio agency: Office de Radiodiffusion Télévision Française.

[3]±0.5 dB RMS, measured using a 100 ms Hanning window with 50% overlap, averaged

on-axis recording, the source image should be located in the direction of the central listening axis if reproduced with a conventional 2/0 front loudspeaker pair.

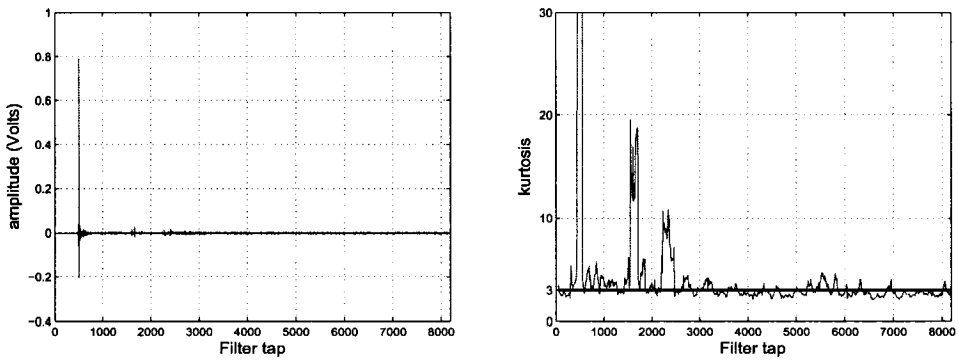
5.2.2 Filter convergence properties

The purpose of the ASUS is to remove the correlated direct-component from the input signals, so the length of the adaptive filter is designed to be the length of the direct-sound components (i.e. up to the mixing time of the inter-microphone impulse response, which is equal to L_r in figure 4.5). To find the onset of the reverberation in the inter-microphone IR, the local kurtosis of the adaptive filter was calculated as per the investigation in section 4.2.3. Kurtosis was used as a measure of normality, as the reverberant component of an impulse response can be defined as that part where the local distribution is normal (Gaussian) (Schroeder, 1987; Abel and Berners, 2004). Kurtosis was calculated according to (4.10), averaged over 96 samples and then the window was advanced by 12 samples (i.e. the overlapping factor was 8). A population of samples with a normal distribution has a kurtosis of 3; which defines the reverberant component of an IR. Deviation from a kurtosis of three can be used to show the effect of the direct sound and strong low-order reflections, as can be seen in figure 5.3(b).

•Effect of filter length.

The input signal delay (i.e. delay 1 in figure 4.4) used in the analysis was 500 samples, and as the source was on-axis the impulse response is centred about tap number 500 of the adaptive filter. As shown in the kurtosis plot in

over the duration of the recording made using a 30 seconds white noise signal reproduced with a loudspeaker.



(a) Time domain adaptive filter coefficients.

(b) Kurtosis for adaptive filter.

Figure 5.3: Filter characteristics for AB microphone recording with on-axis loudspeaker reproducing white noise in a concert hall. Adaptive filter length (N) was 8192 taps, overlapping factor α=4, μ=0.08 and regularization parameter was 1.

figure 5.3(b) that the impulse response distribution is non-normal for about 2000 samples after the direct-sound onset. The two reflections at about 29 ms and 40 ms are at least 20 dB lower than the direct-sound peak at tap 500, which explains why system performance in terms of misadjustment (see figure 5.4) is not significantly improved for filter lengths greater than about 512 taps (but it must be at least 500 samples to account for the input delay).

•Effect of step-size and regularization constant.

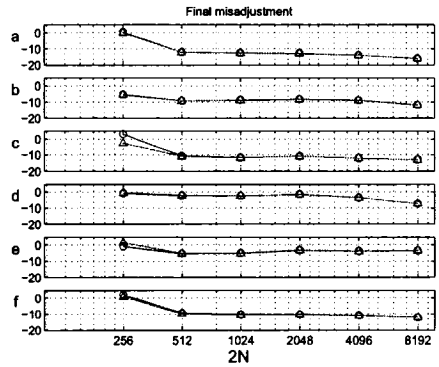
The affect of step-size (μ) and regularization constant δ on misadjustment is visualized in figure 5.4 and 5.5. It can be seen in figure 5.4(c) that a high μ caused the filter to become unstable if the regularization constant was too low (at least for the guitar solo with a filter size of 256 taps). The importance of a suitably large δ is returned to when the performance characteristics for pure sine waves are considered. However, the cost of a high regularization constant can be seen in figure 5.5 in terms of a slower rate of convergence; it seems like a constant of about 1 is a good compromise for both fast convergence

rate and insurance of stability.

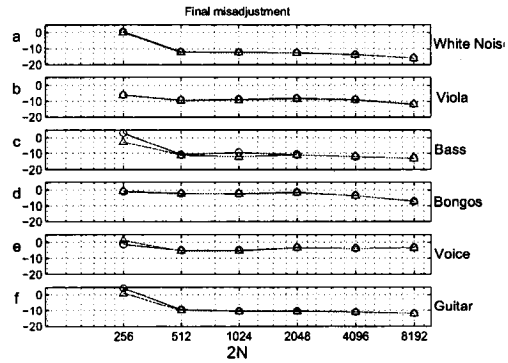
•Effect of microphone configuration on filter convergence.
Looking at the adaptive filters in figure 5.6, it can be seen that as the loudspeaker source is moved from the -3 m to the 3 m position the peak grows from about 0.3 V to 1.0 V (for the ORTF arrangement); a change of over 10 dB. This is expected as the source is moving closer to the right-hand microphone, so the left-hand channel must be boosted to compensate to cancel the direct sound components. The steering servo in the Dolby Pro-Logic II system responds in a similar way by boosting the lower-level channel before the difference signal is calculated (as discussed on page 75). With the ASUS, this gain is applied on a frequency-by-frequency basis as if there is an N-band "equalizer" (i.e. not $2N$ as there are only N frequency bands obtained from a $2N$ FFT). Figure 5.6(a) shows that the high-frequency boost is larger for the XY microphone arrangement when the source is located 3 m off-centre, due to the cardioid directivity pattern and microphone angle which created an level difference larger at high than low frequencies.

Besides aligning the two input signals *spectrally*, the adaptive filter also aligns them *temporally*. This can be seen in the time-domain filter response shown in figure 5.6(b). For the XY configuration, there is no time shifting of the main peak of the adaptive filter when the sound source moves from the 3 m to the -3 m position. This is because there is no change in time-of-arrival change due to the coincident diaphragms. The largest change is observed with the AB configuration (about 100 samples) as this has the largest diaphragm spacing.

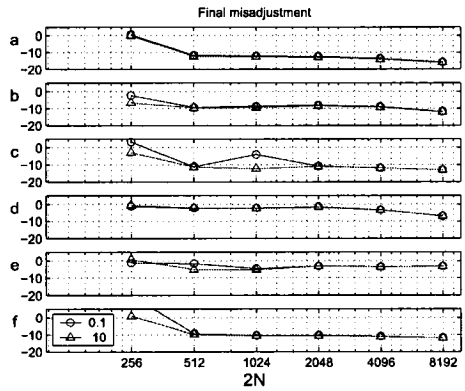
Effect of step-size and filter length.



(a) $\mu = 0.08$



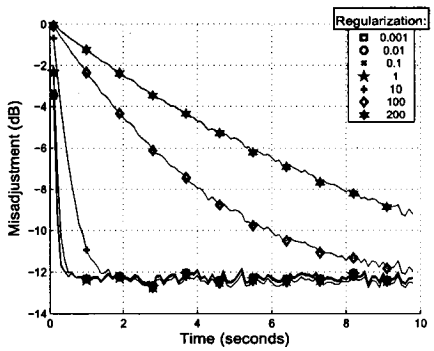
(b) $\mu = 0.12$

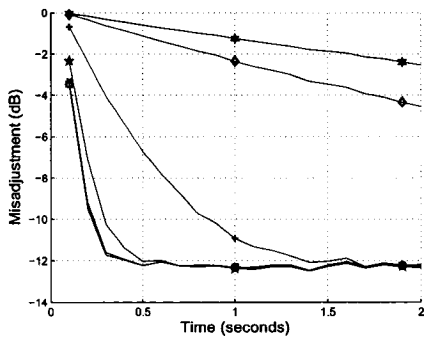


(c) $\mu = 0.16$

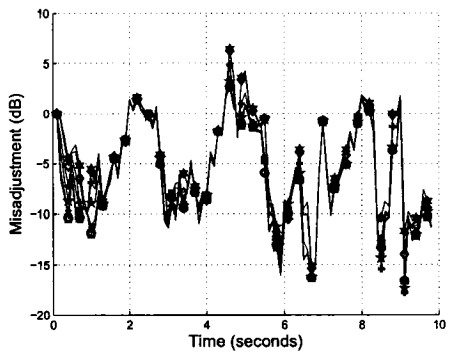
Figure 5.4: Affect of step-size (μ) on final misadjustment (averaged from 25 to 35 seconds, shown in dB) for a centrally located sound source radiating a variety of stimuli: **a**: White noise; **b**: Viola; **c**: Bass; **d**: Bongos; **e**: Voice; **f**: Guitar. Adaptive filter length investigated from 256 to 8192 samples. In all simulations $\alpha = 4$. Regularization parameter was 0.1 or 10. The delay on each input channel was 500 samples.

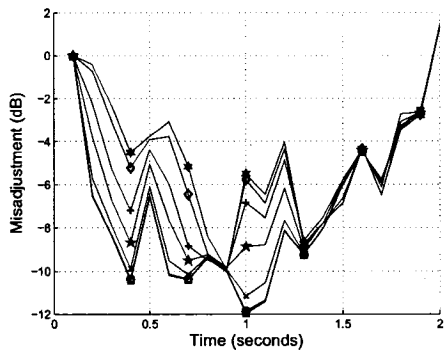
Effect of regularization constant.



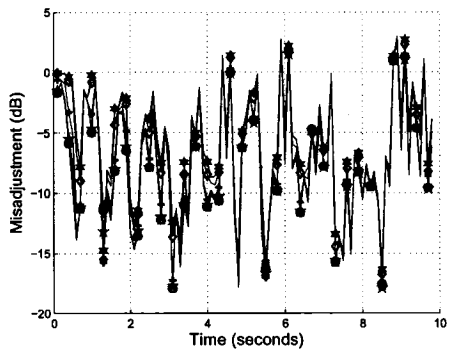


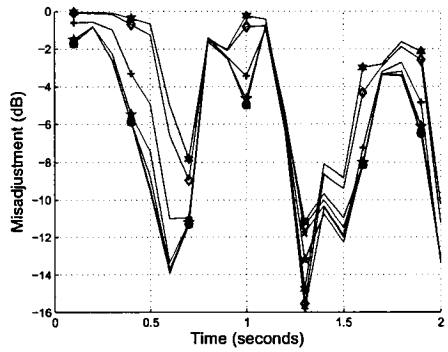
(a) White noise source





(b) Solo voice

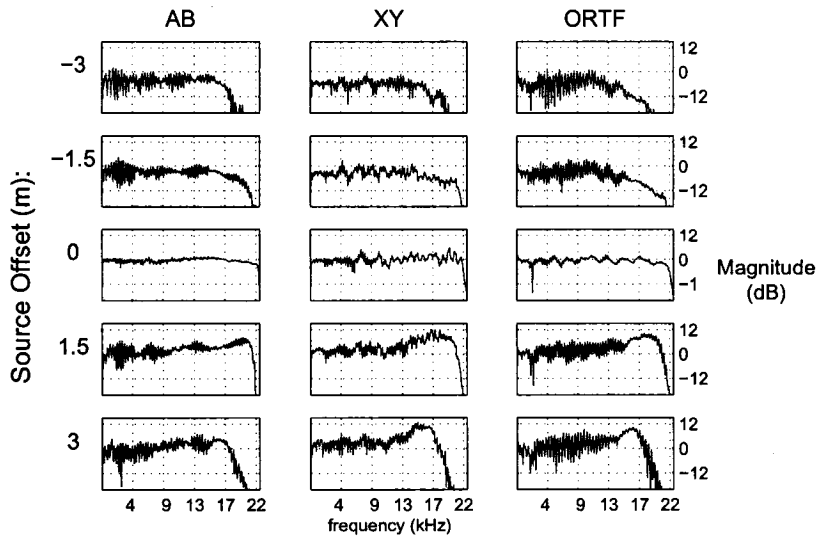




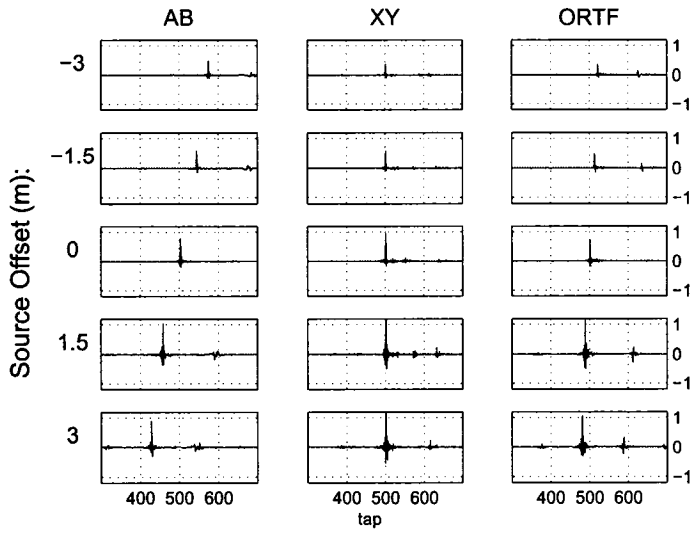
(c) Solo bongos

Figure 5.5: Effect of regularization constant (δ) on misadjustment for on-axis recording of reproduced white noise and anechoicly recorded solo voice and bongos reproduced with single loudspeaker using a spaced (AB) microphone configuration. Algorithm parameters were: μ=0.08, α=4, $2N$=1024. A zoomed-in plot is shown on the right-hand side.

Adaptive filter condition for different microphone configurations.



(a) Frequency domain adaptive filter coefficients.



(b) Time domain adaptive filter coefficients.

Figure 5.6: Frequency and time domain representation of one adaptive filter ($\mathbf{w}_{12}$) for a white-noise sound source reproduced by a single loudspeaker in Pollack Hall, at five different locations. The 0 m offset location is when the sound source was equidistant to each microphone. The loudspeaker source was moved ± 3 m from the central 0 m location. The AB microphone pair was spaced 50 cm, the XY was a coincident pair, and the ORTF was spaced 16.5 cm and angled at 110°. The input signal delay (see figure 4.3) was 500 samples.

5.2.3 Effect of microphone spacing

In this section, the theoretical model for the system developed in the previous chapter is tested. This model describes the behavior of the new system for input signals of different electronic correlations and predicts the level of the output signal. In summary, the developed theory shows that the misadjustment[4] is related to the cross-correlation between the two input signals ($\mathbf{c}_{m_j m_i}$) according to (5.2):

$$\Psi_i = 1 - \left|\left|\mathbf{c}_{m_j m_i}\right|\right|^2, \tag{5.2}$$

where

$$\mathbf{c}_{m_j m_i} = \frac{\mathbf{r}_{m_j m_i}}{\sigma_{m_i}\sigma_{m_j}} \tag{5.3}$$

is the cross-correlation vector between $\mathbf{m}_j$ and m_i (see e.g. Benesty et al., 2000b) and

$$\mathbf{r}_{m_j m_i} = E\left\{\mathbf{m}_j(n) m_i(n)\right\} \tag{5.4}$$

is the M-length cross-correlation vector between the input microphone signals.

The algorithm parameters used for the analyses were:

- Overlapping factor (α): 8.
- Block size ($2N$): 2048 samples.

[4] Misadjustment is defined as the level ratio between the output (rear loudspeaker) signal and the input signals; Ψ_i.

- Delay of unfiltered channel: 500 samples.
- Regularization parameter (δ): 1.
- Update coefficient (μ): 0.05.

White-noise was reproduced from a single loudspeaker on the stage in Pollack Hall, equidistant to each microphone. The same cardioid microphones were used, but the spacing between their diaphragms was varied whilst keeping the distance from the source constant (about 3.5 m). The one exception to this was a case when the microphone pair was taken to the back of the hall, about 26 m from the loudspeaker source yet only 1 m apart. In such a case, even though the correlation of the direct sound and early reflections may be high, the reverberant-to-direct level ratio (γ) would also be high and as predicted by (4.56) (derived in section 4.4.3 and summarized in figure 4.14); when γ is larger than about 15 dB, the overall interchannel correlation $\mathbf{c}_{m_j m_i}$ is dominated by γ and is no larger than 0.4. This predicts a misadjustment of only about -1.5 dB; in other words, the rear loudspeaker signals of the ASUS would have a similar level to the front speaker signals when the microphone pair is far away from the source.

Looking at figure 5.7, the far-away microphone pair (i.e. case 1 m*) is highly correlated at low and high frequencies (about 0.95 at 100 Hz and 0.8 at 12 kHz) and therefore the misadjustment is low (about -15 dB). At mid-frequencies, however, the signals are less correlated (0.15 at 1.5 kHz) giving a higher misadjustment (close to 0 dB). A similar trend is also seen for the other microphone pairs. This can be explained by two factors: Firstly, if the microphone diaphragm spacing is small compared with the wave-length λ, there is little decorrelation effect as the sound pressure is similar at each microphone (Jacobsen and Roisin, 2000). This explains why for the 6 cm

spacing the microphone signals are highly correlated up to about 1.5 kHz, where λ=23 cm. Secondly, the reverberant-to-direct ratio reduces at high frequencies due to air absorption (as mentioned; about 0.1 dB per metre at 4 kHz but only 0.001 dB/m at 100 Hz; Kinsler et al., 1999, pg. 224) and sound absorption from objects in the room such as soft chairs and carpets.

The relationship of microphone signal cross-correlation and misadjustment shown in figure 5.9 looks promising to support the new theory developed in the last chapter (the accuracy is within $\pm$5 dB). The model is less robust when the correlation is high, but it must be remembered that the theoretical derivation assumed stationary impulse response statistics as well as a noise-free operating environment; both of which was obviously not the case (see discussion on page 143 for sources of IR variation). The noise floor was about 48 dB unweighted, yet only about 32 dB A-weighted; indicating a higher noise level at low frequencies. The source was reproduced at about 90 dB (measured 3 m away), so the signal-to-noise ratio was probably about 60 dB at high-frequencies and 40 dB at low frequencies. Also, because the filter update (step-size) was finite, the optimal solution could never be exactly met and results in *gradient noise* (Widrow and McCool, 1976) which would limit the minimum level of the error signal. Other studies (e.g. Elko et al., 2002) have remarked how it is very difficult to get misadjustment statistics for practical adaptive filtering applications (such as echo canceling in teleconferencing) less than about 20 dB.

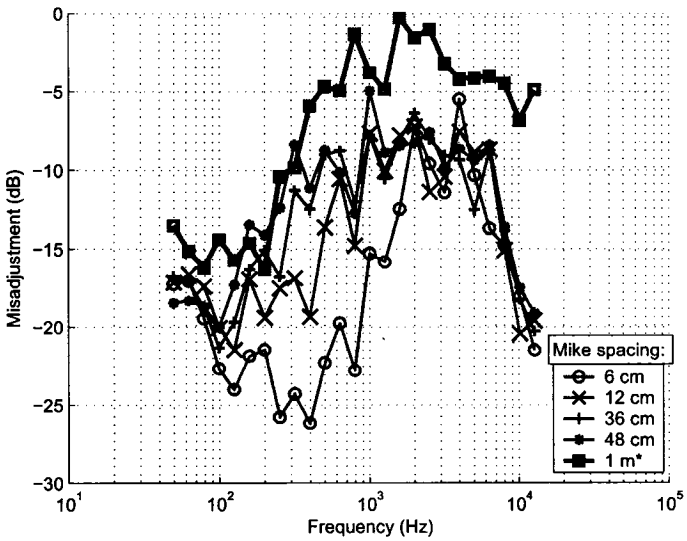


Figure 5.7: Final misadjustment as a function of frequency (1/3rd octave band average) for different microphone spacings. White noise was reproduced with a loudspeaker in Pollack Hall and recorded with microphone pairs with a variety of spacing. The "1 m*" recording was with a microphone-spacing of 1 m, but was 26 m from the source. All other recordings were made 3 m from the source.

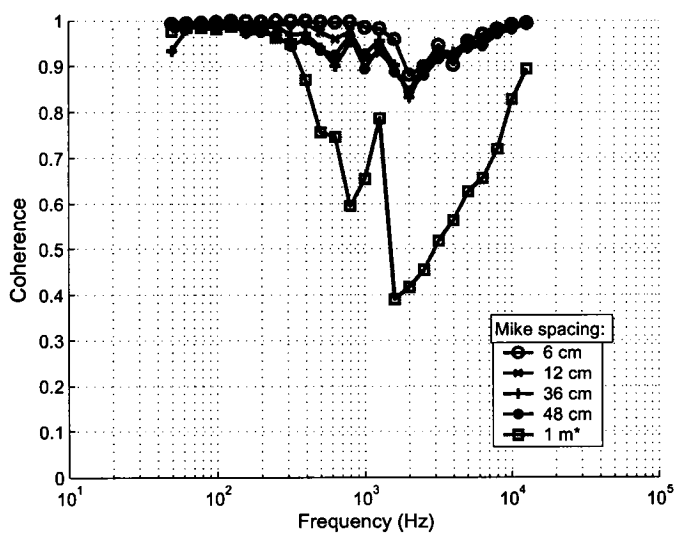


Figure 5.8: Coherence (maximum absolute cross-correlation in a 23 ms window, averaged over the recording) for different microphone spacings (as figure 5.7).

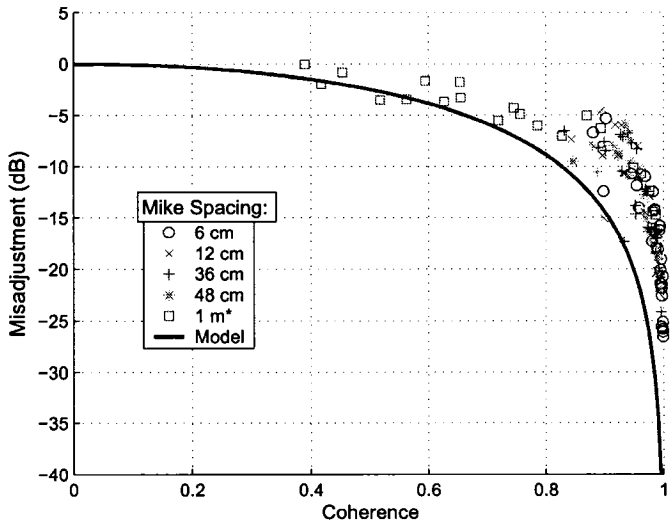


Figure 5.9: Effect of inter-microphone correlation on misadjustment- i.e. a combination of the data presented in figures 5.7 and 5.8. The theoretical derivation for the model is described in section 4.4.2. Different markers represent misadjustment and cross-correlation of input signals (measured within a 1024 sample window) averaged over a 5 second portion of the recording.

5.3 Response of the system to sine waves

Here the ASUS performance for pure sine-waves is investigated. This is primarily an exploratory investigation, as pure noise-free sine-waves are, of course, practically impossible for two-microphone recordings of natural instruments- though they are (unfortunately) not uncommon in computer-generated "electroacoustic" music.

A five second input file was created with a pure sine-wave of various frequencies- both the left and right channels had equal peak-to-peak voltages of ± 1 Volts. 400 randomly chosen frequencies were tested (ranging from 20 Hz- 20 kHz).

The parameters for the algorithm were as follows:

- Overlapping factor (α): 8.
- Block size ($2N$): 1024.
- Delay of unfiltered channel: 500 samples.
- Regularization parameter (δ): 0.1.
- Update coefficient (μ): 0.05.

As can be seen in figure 5.10, the rate of convergence (RoC) for the misadjustment to reach -60 dB was fairly constant across frequency at approximately one second (the misadjustment was calculated every 0.1 seconds- hence the time quantization). However, it was found that for some sine-wave inputs the ASUS went unstable- the output signal actually grew in magnitude and would eventually produce a NaN error (these unstable frequencies

are shown with crosses in figure 5.11). Also, there are obvious peaks in the RoC which led the investigation to look at the relationships between these poor convergence rate frequencies. Convergence was poor (i.e. slow or unstable) if the sine wave period was closely related to the number of overlapping samples (S)- i.e. the filter length $2N$ divided by the overlapping factor α (S is the number of new samples created for every iteration of the algorithm). In this case, the number of overlapping samples was 128- i.e. the same period as a 344 Hz sine wave. Fortunately, as shown in figure 5.11 the instability problem can be mitigated by using a regularization constant (δ) of at least 100. The disadvantage of using a such a high regularization constant is the increase in convergence time (see figure 5.5(a)). However, with a typical convert-hall musical recording, a regularization constant as low as 0.001 is sufficient for filter convergence with a suitable RoC (see figures 5.5(b) and (c)).

5.4 System performance with commercial two-channel recordings

As mentioned in the introduction, this thesis is principally concerned with two-microphone recordings of solo musical instrument performances in live spaces. However, we will briefly look at a study on how the ASUS performs for arbitrary "off the shelf" music recordings, e.g. using the left and right output channels from a commercially available CD. This revealed some interesting limitations of the ASUS, which in turn led to a practical solution. A full investigation into this is beyond the scope of this thesis, as there is such a huge range of recording, mixing and editing techniques used in commercial

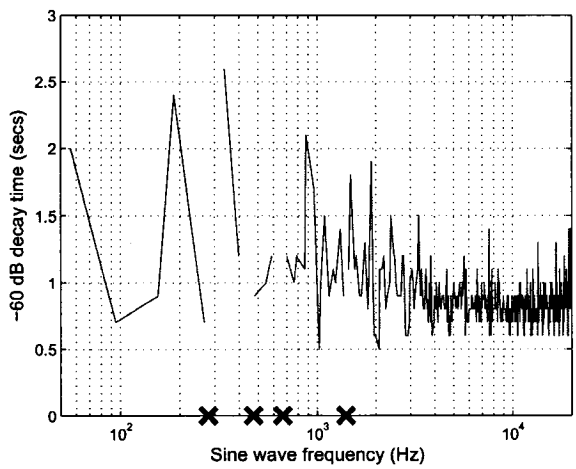


Figure 5.10: Time taken for misadjustment to decay to -60 dB for different pure-tone inputs (magnitude of ± 1 V). The crosses indicate frequencies which were unstable (i.e. the algorithm didn't converge with these input signals). In all tests, regularization parameter δ=0.1.

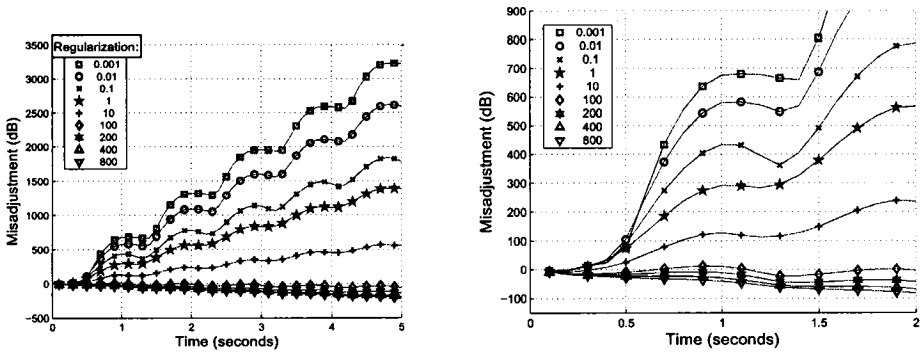


Figure 5.11: Output signal level relative to input signal (i.e. misadjustment) as a function of time in response to a 344 Hz sine-wave input for different regularization constants. Note that the regularization parameter must be larger than 100 for the filter to converge. (A positive misadjustment indicates that the output signal is larger than the input- i.e. the filter is unstable.)

recordings (especially "pop" music).

5.4.1 Using musical recordings with hard-panned images in the mix.

A major problem with using the ASUS was found when dealing with music which contains a sound source which is only present in one channel- so called *hard-panned* sources. "Hard-panning" is a commonly-used term in sound engineering for musical recordings to describe the process whereby a single audio channel is mixed to only one channel; for instance a single vocal track mixed to just the left channel of a two-channel mix for a CD. Of course, hard-panning is impossible for "live" recordings made with two-microphones due to the acoustic cross talk from the sound source to both microphones, but hard-panning is very common in audio mixes for pop-music.

A good example of a recording with hard-panning is the song *Her Majesty* by The Beatles. Looking at the time-domain waveform of the left and right channels of this two-channel pop-song in figure 5.12(a), four regions are identified. The output of the ASUS in response to this input is shown in figure 5.15, and is discussed with reference to the music (times given are referenced to the start of the CD track, which begins with 0.75 seconds of silence).

- Region A: 0.75-2.45 seconds. Introductory "crash" with cymbals, snare-drum and guitars. The decay is 2.8 seconds in the left channel but is faded to zero after 0.7 seconds in the right channel. A quick inspection of a Lissajous phase plot (figure 5.13(a)) of this region shows that the two signals are weakly correlated- and as expected, the ASUS output level is high for this region (see figure 5.15(a)).

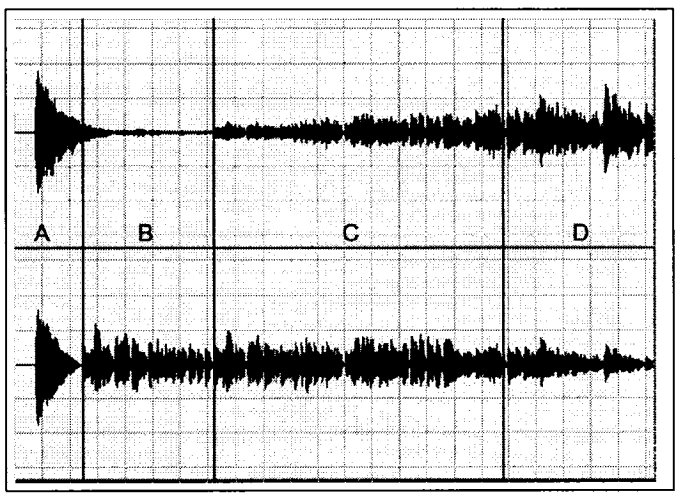
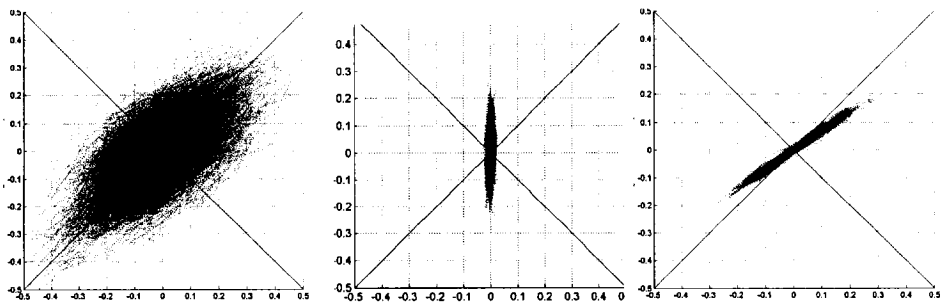


Figure 5.12: Time-domain plot of a recording which contains both hard-panned music and dynamic amplitude panning: *Her Majesty* by The Beatles. X axis is time (0-23 seconds) and Y axis is Voltage- left channel on the top.



(a) Region A (b) Region B (c) Region D

Figure 5.13: Lissajous phase plots for various regions of *Her Majesty*. x axis is the left channel voltage, and y axis is the right channel.

- Region B: 2.45-7.24 seconds. An example of hard-panning: voice and acoustic guitar in the right channel only, with what sounds like plate-reverb of voice in the left channel. The left channel level is much lower, as can be seen in the Lissajous plot in figure 5.13(b). Also on the left channel is a strange female-voice or sped-up male voice vibrato note at 3.7 seconds.[5] If there is any reverb on the right channel, then it is very low level. The ASUS would not be able to cancel the direct-sound components of the guitar in the right channel, as they do simply not exist in the left channel. Also, as the reverb in the left channel seems to be "100% wet" (i.e. no direct sound), the voice can not be cancelled and would also appear in the surround-right channel. Analysis of the ASUS output for section B in figure 5.15(a) shows this is the case. Therefore the source image would be pulled in the direction of the rear loudspeakers, which conflicts with the aims that the source image should be undistorted in the upmixed 2/2 audio scene (i.e. compared with the 2/0 scene) to respect the mixing engineers' intentions of a frontal source stage. When listening to this, the source image would be spread over this side.

- Region C: 7.24-16 seconds (approx.). Voice and guitar cut-in on left channel (i.e. no hard-panning) and slowly increase in level. Level of voice and guitar slowly decrease in right channel.

- Region D: 16 seconds-end. Voice and guitar are higher in level in the left channel than the right. Amplitude-panning effect moves image to left channel. A Lissajous plot (figure 5.13(c)) reveals that the left and right channel are highly correlated for regions C and D, so the misadjustment would be low.

[5]Googling with the words "Her Majesty beatles left channel" on 4-11-05, all ten hits on the first page mentioned this odd sound.

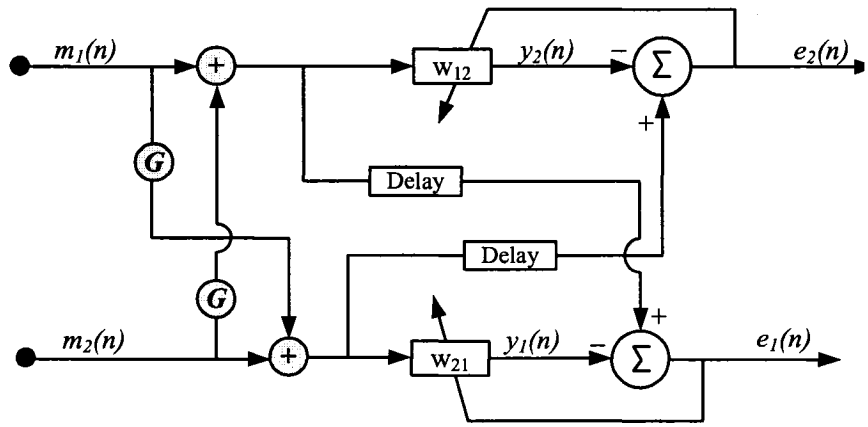


Figure 5.14: Modified ASUS with cross-talk set by gain G between input channels to help convergence for input signals with hard-panned sources.

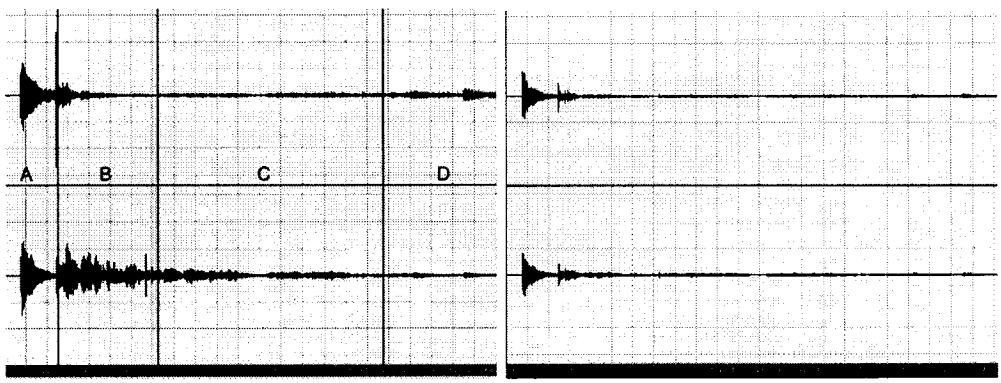
To overcome this problem associated with hard-panning in music recordings, a cross-talk between the two input channels is introduced as shown in the signal schematic in figure 5.14. This modified form of the ASUS dramatically improves the performance for music with hard-panned sound. The ASUS was implemented with the cross-talk added off-line to the entire stereo signal (though it could of course be added on a sample-by-sample basis). The algorithm parameters used throughout section 5.4 are:

- Overlapping factor (α): 8.
- Block size (L): 1024.
- Delay of unfiltered channel: 500 samples.
- Regularization parameter (δ): 0.1.
- Update coefficient (μ): 0.05.

The effect on the cross-talk on the output signals of the ASUS can be

clearly seen by comparing figures 5.15 (a) and (b). The misadjustment for the system with different cross-talk gain values is shown in figure 5.16. The effect of the cross-talk is most dramatic in section B of the *Her Majesty* recording; i.e. when the voice and guitar are both hard-panned in the right channel. As expected, when there is no cross-talk the speaker channel RS output (i.e. error signal e_2) is the same level as the input giving a misadjustment of approximately 0 dB. With a cross-talk of -5 dB, the misadjustment decreases to about -15 dB; when all that can be heard in the left and right channels is the plate-like reverberation from the voice and guitar, plus the strange female vibrato note.

The effect of adding the cross-talk is to increase the correlation between the input channels- as shown in figure 5.17. When the cross-talk is unity (i.e. G= 0 dB), then both input channels are identical so there would be no output from the ASUS (i.e. the misadjustment would be -∞ dB). As predicted by the theory discussed in section 4.4.2, the misadjustment (Ψ) can be predicted for a given cross-correlation between the input signals ($\mathbf{c}_{m_j m_i}$)- as shown in (5.2). For instance, if two uncorrelated white noise signals are mixed with a cross-talk (G) of -5 dB, figure 5.17 shows that the interchannel correlation is now 0.85. Another way of phrasing the effect of the cross-talk modification is that it bounds the maximum level of the output (error) signal (i.e. in terms of misadjustment).



(a) No cross-talk. (b) -5 dB cross-talk.

Figure 5.15: ASUS output for amplitude-panned input signal (*Her Majesty* by The Beatles). x axis: time (0-23 seconds); y axis: Voltage, same scale in (a) and (b).

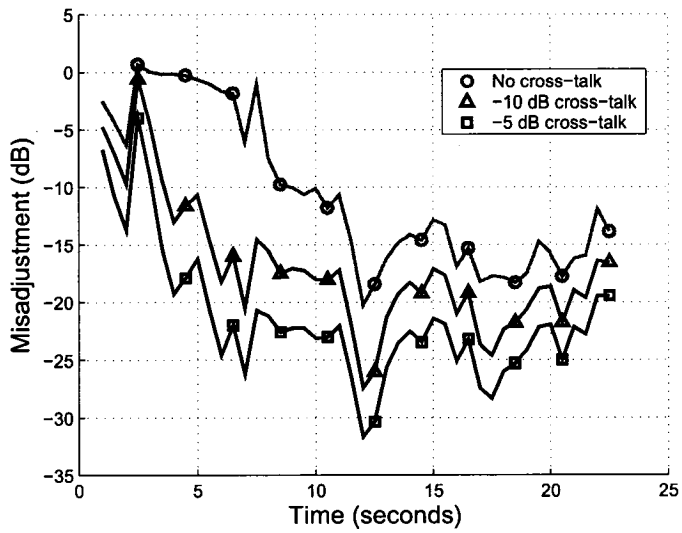


Figure 5.16: Misadjustment for right channel of ASUS output with and without cross-talk for the *Her Majesty* piece.

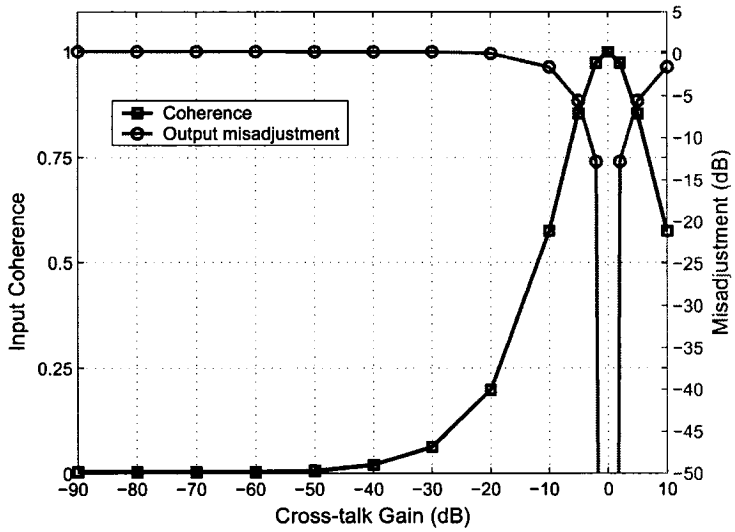


Figure 5.17: For two uncorrelated noise inputs: Showing the effect of interchannel input cross-talk on interchannel coherence and output misadjustment (misadjustment is the dB ratio of output to input level). Data from empirical measurements using the new upmix system.

5.4.2 Using musical recordings with time-delay panning.

We shall now look at how the cross-talk modification affects the system performance when dealing with input signals with time-delay panning.[6] The stimuli used here were the same as in previous sections in this chapter (see section 5.2 for recording details). In summary: white noise was reproduced from a single loudspeaker on a stage in Pollack Hall and recorded using a spaced pair of B&K type 4011 cardioid microphones, mounted on 3 metre

[6]In two-channel audio, time delay panning occurs when the direct sound from a source is reproduced from one loudspeaker before the other, so the cross-correlation between the two signals has a maxima at time $t \neq 0$. As long as the lag of the cross-correlation peak is between ± 1 and ± 10 ms, no separate echo is heard and the source image is heard in the direction of the leading loudspeaker. For delays less than 1 ms the image is heard between the loudspeakers (i.e. there is a single source image). Time-delay panning is discussed in section 2.1.8.

high microphone stands facing the loudspeaker (see photograph in figure 5.2). The microphones were spaced by 50 cm and the sound source was 3 m off-axis- i.e. 3 m to the left of the central axis. This gave a time-of-arrival delay between the two microphones of about 1.7 ms (77 samples at 44.1 kHz sample rate); as shown by the cross-correlation response in figure 5.18. The two-channel recording was then processed with the cross-talk modified ASUS (as shown in figure 5.14), and the adaptive filter coefficients and output error level were measured for three levels of cross-talk gain (G): -5 dB; -10 dB; and with no cross-talk (i.e. $G = -\infty$ dB). The affect of G on filter adaptation and misadjustment are summarized in figure 5.19.

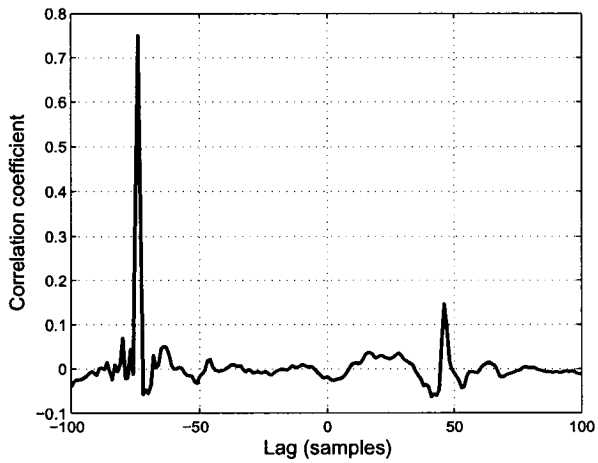
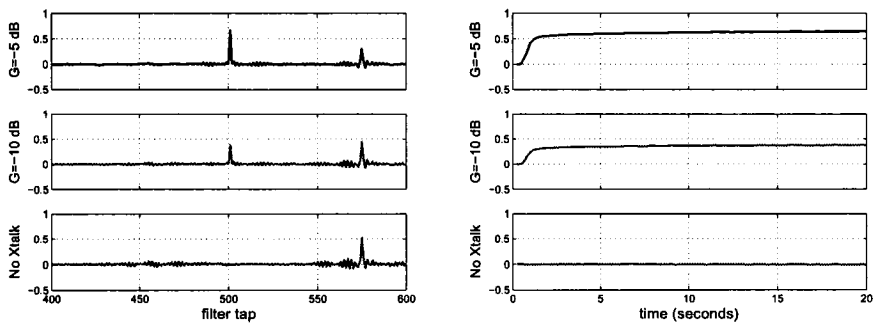


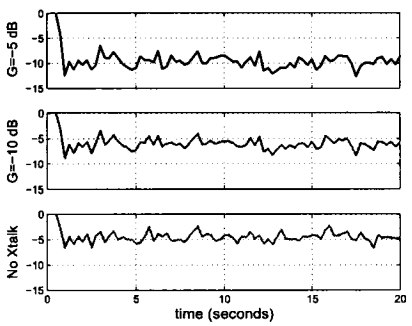
Figure 5.18: Electronic interchannel cross-correlation of spaced two-microphone recording. The noise source (loudspeaker reproducing white noise on stage in Pollack Hall) is offset by 3 m from the central microphone axis. The smaller peak (which is symmetrical about the main peak) is probably due to a first-order floor reflection.

Figure 5.19 shows that the adaptive filter compensates for the cross-talk with a peak in the filter coefficients at zero delay. Due to the 500 sample offset between the unfiltered input signal and the filter signal (considering the $\mathbf{w}_{21}$ filter system, this means signals m_1 and m_2 respectively), this peak



(a) Adaptive filter coefficients $\mathbf{w}_{21}$ (at 20 seconds). The peak at 501 samples is due to the cross-talk and the second peak due to the time-lag between the microphones due to the off-axis source.

(b) Change in tap 501 of $\mathbf{w}_{21}$ adaptive filter with time.



(c) Misadjustment (dB ratio of left input channel to LS output channel e_1).

Figure 5.19: Effect of cross-talk gain G on filter convergence for a spaced-microphone pair recording of an off-axis noise source in Pollack Hall. There was a 500 sample delay for the unfiltered input channels.

occurs at the 501^{st} coefficient. As can be seen from figures 5.19(a) and (b), the height of this peak is proportional to the amount of cross-talk; hence it is less for the G= -10 dB case (as there is less cross-talk to cancel). For all levels of G, a peak occurs at a delay corresponding to the delay between the two channels due to the source offset- i.e. 77 samples.

5.5 Comparison of the new upmixer with two commercial upmixers

In this section a comparison is reported with the output signal properties of the ASUS and two commercially available upmixers. Various two-channel input signals were used, all of which were un-encoded (i.e. have not been processed with a down-mix algorithm). As mentioned in the introduction of the dissertation, the focus of the thesis is on a thorough understanding of the new system, specifically for dealing with simple two-microphone recordings of solo musical performances; a thorough comparison of the subjective attributes of the new system in relation to extant systems is simply beyond the scope of the present work. Furthermore, the new upmix system described in this thesis is a two-to-four channel upmixer, whereas commercial upmix systems are nearly all two-to-five (or more) channel systems which utilize the centre channel of the ITU-R BS 775-1 (1994) loudspeaker configuration. However, a basic insight into the performance of the two other upmixers can still be gleaned by an analysis of the level and correlation properties of the output signals.

5.5.1 Selection of upmixers

Besides the different versions of the ASUS (with different cross-talk gains), the two upmix systems looked at were Dobly Pro-Logic II (DPLII) and Circle Surround II (CSII) (DPLII is described in section 2.2.3). These were chosen due to availability. Also available was Dolby Pro-Logic I; this was not chosen because it is generally considered an obsolete technology (especially for un-encoded input signals). As with the ASUS, both DPLII and CSII are so-called *natural spatialization algorithms* (Rumsey, 1999) or *multichannel converters* (Avendano and Jot, 2004) as they do not simply add reverberation to create the new channels but utilize information already within the original signals.

The algorithm parameters used throughout this section are:

- Overlapping factor (α): 8.
- Block size ($2N$): 1024.
- Delay of unfiltered channel: 500 samples.
- Regularization parameter (δ): 0.1.
- Step-size (μ): 0.05.
- Cross-talk gain (G): -∞ dB and -5 dB (i.e. there are two ASUS versions).

5.5.2 Method and stimuli

This comparative study was conducted at Philips Natlab research facilities using a kindly provided surround-sound processer (manufactured by Marantz, model SR4400 "AV Surround Receiver") with a DPLII and CSII upmixer. The Marantz unit was fed a two-channel input signal and the five processed channel outputs were recorded onto a PC using MATLAB software and an RME Hammerfall soundcard. The processor was configured so that the sub-woofer channel was not used and both input and output signals were analog with RCA-type connectors. The two-channel audio source material was also from the same PC, with the soundcard clocked at 44.1 kHz sample rate and 16-bit resolution for recording and playback. The "music mode" was selected for each upmixer, but there were no other options available (such as changing the time delay for the rear channels).

Three two-channel stimuli were used, all of which have been introduced earlier in this chapter:

1. White noise reproduced from a single loudspeaker in Pollack Hall. The recording was made using a forward-facing spaced microphone pair (50 cm spacing, B&K type 4011 cardioid microphone), 3.5 metres from the source, with the source equidistant to each microphone (see photo-graph in figure 5.2).

2. 40 second anechoic recording of singing female voice reproduced from a loudspeaker in Pollack Hall and recorded in the same way as described for the noise source above, except the loudspeaker was 3 metres off-axis (in the direction of stage-right). (See figure 5.1 for time and spectral envelope details of original source.)

3. 23 second two-channel recording of *Her Majesty* by The Beatles, from a CD- see figure 5.12 for waveform. This was chosen because it contains hard amplitude-panned sources, as an example of a mixing technique common in pop-music.

5.5.3 Analysis of output signals

The level and correlation between audio signals radiated with loudspeakers affect auditory image formation and direction.[7] For the upmixers investigated (i.e. the commercial upmixers and the new system) the level was calculated in two ways; firstly, as the ratio between the front and rear loudspeaker signals (the front signal levels were summed and the rear signal levels were summed and the level ratio calculated); and secondly, as the ratio of the rear-right channel level to the front-right channel level (for the ASUS, this is the same as the misadjustment). These level analyses are summarized in figure 5.20.

The electronic correlation between the output channels was analysed by calculating the average cross-correlation in a 23 ms window between a variety of signal pairs;[8] as summarized in figure 5.21 for a noise signal and figure 5.22 for the voice signal- both recordings made in Pollack Hall as described. The time-averaged cross-correlation measurement undertaken is not ideal. As mentioned in section 2.1.8, variation in the **IACC** over time has been shown to affect the perceived width of source images (Mason et al., 2005). However, such an averaged statistic gives a basic insight into how the output channels

[7]It is the inter*aural* correlation which affects image formation (Chernyak and Dubrovsky, 1968), but this can be approximated from the interchannel correlation of two signals feeding loudspeakers in the listening room; as discussed in the section on panning in chapter 2.

[8]For two signals, say the L and R loudspeaker channel vectors, the cross-correlation was calculated using the following MATLAB incantation: `xcorr_LR=xcorr(L,R,1024,'coeff')`.

of the upmixers maybe be perceived if radiated by loudspeakers around the listener.

Looking at the level analysis in figure 5.20, we see that the rate of convergence for the ASUS was increased by adding the cross-talk (this was concluded by looking at the slope of the misadjustment with time). For the off-axis voice recordings, all upmix systems give a relatively high rear level output, though this is lowest with the CSII system (which gives a lower rear loudspeaker level in general). For the DPLII system and unmodified ASUS (i.e. when there was no cross-talk), there are occasions when the front-loudspeaker level is larger than the rear. Regarding the *Her Majesty* piece, it can clearly be seen that as well as the unmodified ASUS, the DPLII system can not cancel the hard-panned source and it appears in the RS channel with a high level. This suggests that DPLII does not use any cross-talk of the input channel for the upmixing process, though how CSII accomplishes this is unknown; it seems that there is a kind of level "smoother" (like a dynamic compressor) for the rear-loudspeaker channels with CSII as there is not much variation in front-rear level ratio.

As mentioned in the subjective design criteria in section 3.3: *Spatial distortion of source image (compared with 2/0 loudspeaker audition) should be minimized.* In the electronic design criteria in section 4.1 this was translated as: *Signal RS must be uncorrelated with signal L, and LS uncorrelated with R.* This was based on the assumption that it is only those correlated sound components between the microphone channels which contribute to spatial properties of the source image.[9] And if diagonally opposite signals are uncorrelated then they must not contain those sound components which affect source imagery. Looking at the correlation between signals R and LS for the

[9]At least, correlated within an approximately 20 ms lag time.

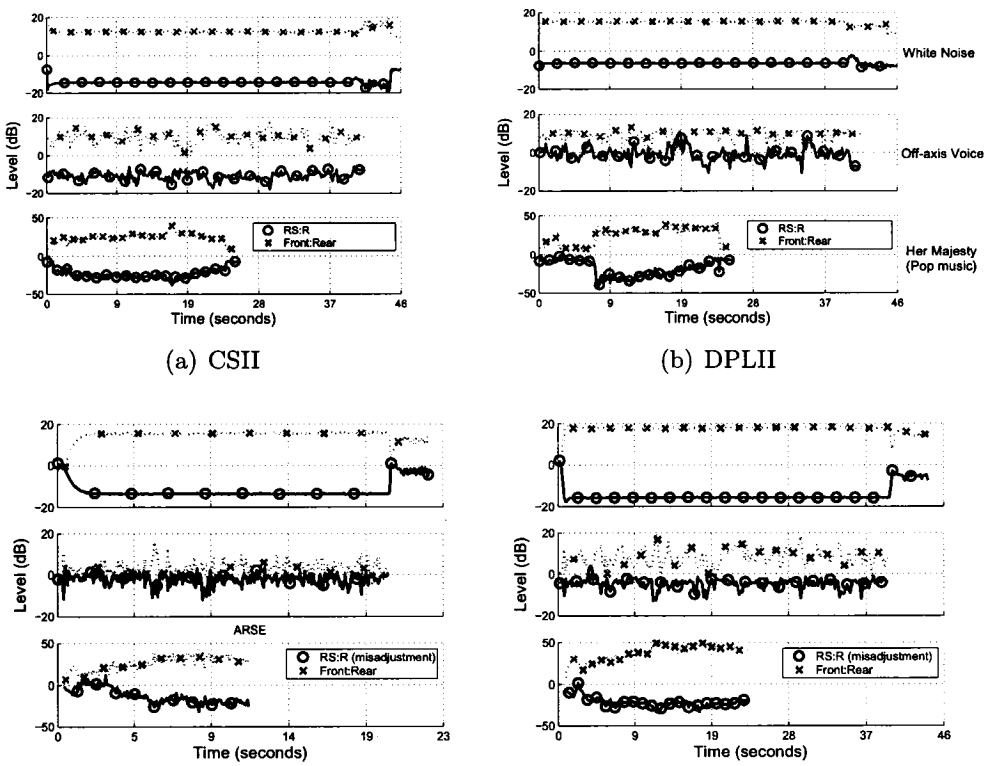
noise stimuli, it is seen that these signals are highly correlated for the CSII and DPLII systems (close to unity) yet are uncorrelated for the ASUS system: empirically confirming the principle of orthogonality (see section 4.3.3). For the voice source, the system with cross-talk slightly increases the cross-correlation here, but it is still very close to zero (a probably imperceptible difference, as the auditory system is less sensitive to changes in interaural correlation close to zero; Culling et al., 2001).

Comparing the left and right signal correlation does not give a representative idea about how the source image might be perceived with the different systems, as DPLII and CSII both produce a centre loudspeaker channel as well. However, looking at the correlation between the side loudspeaker channels (i.e $R - RS$), we can see that the rear channels for the CSII system are delayed by 440 samples- i.e. 10 ms. For musical audio, it is not recommended to implement any delay on the rear loudspeaker channels with DPLII (Dressler, 2000).

For the rear loudspeaker channels, the correlation was unity for the CSII system: in other words they were identical; a "mono" surround audio signal (this may have been due to experimental error or a problem with the particular decoder unit). Listening to this it was quite obvious; the reverberance seemed to decay to a point directly behind the listener, which seemed quite unnatural and at times irritating (although listening to one channel alone, the temporal structure of the reverberance seemed quite natural and pleasant). For the DPLII and ASUS systems there was a strong negative correlation in the rear loudspeaker channels ($LS - RS$). For the ASUS, this is expected as the two input signal were quite correlated (the microphone spacing was close, so even the reverberation would be highly correlated; as previously discussed). Therefore, the two error signals would be similar but

with a reversed polarity; for instance, the signal feeding the LS loudspeaker would be equal to the filtered right input channel subtracted from the unfiltered left input channel, and the signal feeding the RS loudspeaker would be the opposite.

Output signal level



(a) CSII

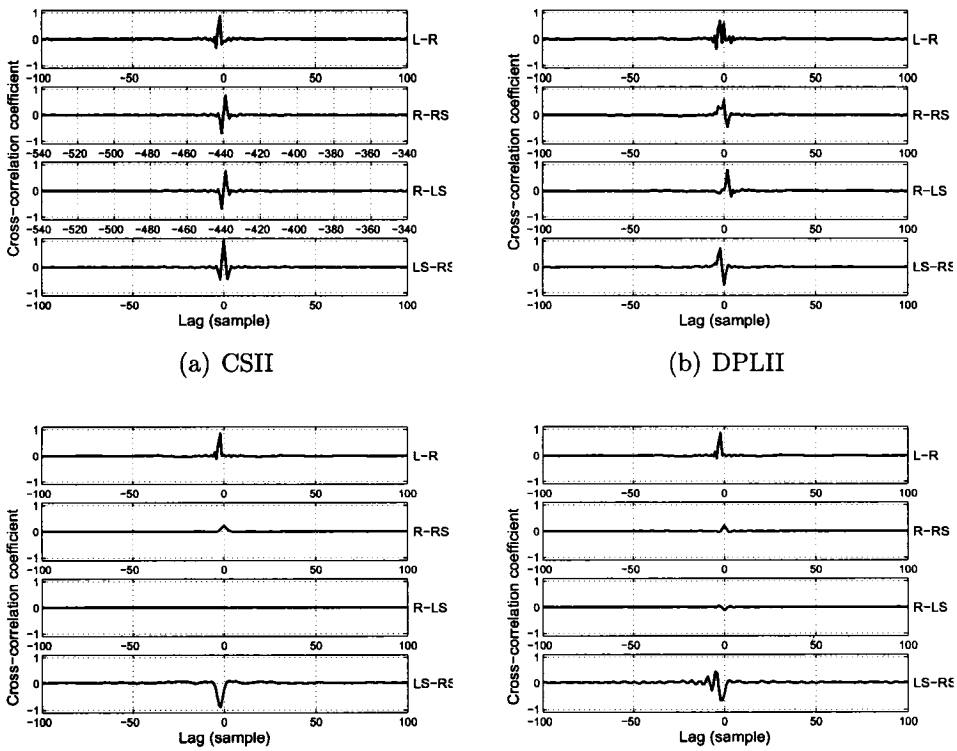
(b) DPLII

(c) ASUS : No cross-talk

(d) ASUS : G=-5 dB

Figure 5.20: Energy ratio of front:rear channels and side channels (the rear-right loudspeaker channel RS and the front-right channel R) for different upmix systems with three different two-channel input signals. In each sub-plot: Top plot is for noise source; middle plot is for off-axis voice recording; bottom plot is for *Her Majesty* by The Beatles.

Output signal correlation



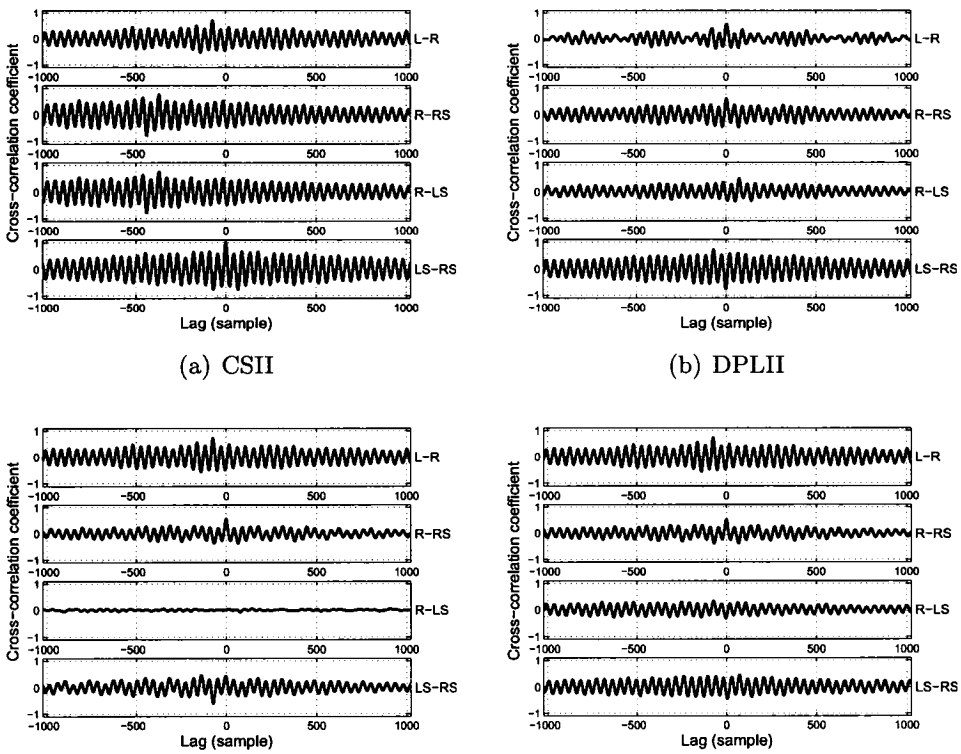
(a) CSII

(b) DPLII

(c) ASUS : No cross-talk.

(d) ASUS : G=-5 dB

Figure 5.21: Cross-correlation coefficient between front, side and rear loudspeaker-channel outputs of different upmix systems. The test stimulus was a recording of a white noise signal reproduced with a loudspeaker in concert Pollack Hall. Note the different lag range for the CSII system: this is due to a 10 ms delay of the rear speakers (DPLII did not have a delay for the music mode setting with the tested unit). Each graph is zoomed in around the main peak. The noise source was slightly off-axis, hence the main peak occurs at a lag of -4 samples.



(a) CSII

(b) DPLII

(c) ASUS : No cross-talk

(d) ASUS : G=-5 dB

Figure 5.22: As figure 5.21 but with the voice source instead of noise source.

Chapter 6

Subjective Evaluation of the New Upmixer

In this chapter, two listening tests are reported on which investigated auditory spatial imagery and preference for audio scenes created with the new upmixer (i.e. the ASUS) and conventional 2/0 audio scenes created by reproduction of an unencoded pair of audio signals with two front loudspeakers at $\pm 30°$. The results of these experiments are then used to provide a qualitative and quantitative response to the subjective design criteria outlined in section 3.3.

The approach to evaluating the degree to which the subjective design criteria have been met follows the general guidelines of ITU-R BS 1116 (1994), which outlines three general sound quality issues to investigate for sound quality evaluation of spatial audio systems:

1. Basic audio quality: *This single, global attribute is used to judge any and all detected differences between the reference and the object.*

2. Front image quality.

3. Impression of surround quality.

The standard recommends *basic audio quality* as the principle topic of investigation. This is to reduce the "response burden" for more complicated critical listening tasks; especially for non-expert ("naïve") listeners. It can be argued that basic audio quality is a more intuitive aspect of the sound scene to evaluate than a descriptive analysis of specific image properties (such as image width or distance). The ITU standards' distinction between front imaging and spatial impression seems to be strongly related to the distinction between source (S) and reverberance (R) imagery outlined in section 2.1.3.

In a study comparing commercial 2-to-5 channel audio upmix systems with a conventional 2/0 system, Rumsey (1999) adds three further research topics to the ITU standard's three ratings listed above, which are here paraphrased:

1. Effect of different program material on sound *quality* ratings (e.g. different recording techniques or sound sources).

2. Do naïve listeners prefer the upmixed or original 2/0 audio scene?

3. Interaction effects of signal processing method and program material on quality ratings.

To investigate how these factors relate to S and R imagery in the 2/0 and upmixed 2/2 sound scenes (i.e. upmixed using the ASUS), two formal

listening tests were undertaken. In the first, the GUI that was introduced in chapter 3 was used to enable the listener to give a quantitative description of perceived S and R image geometry. The preference issue was addressed in a second experiment, comparing the upmixed ASUS scene (and variants of it) with conventional 2/0 scenes.

6.1 Configuration of the new upmix system

The tests reported here were designed specifically to evaluate the new system in the context of the aims outlined in the dissertation introduction; for use with two-microphone recordings of solo musical instrument performances in a concert hall. Therefore, the system architecture is that which was introduced in chapter 4, without the cross-talk mechanism which was developed in chapter 5 for dealing with hard-panned sources found in pop-music.

The algorithm parameters for the adaptive filter in the ASUS were chosen as a result of the empirical parametric study reported in section 5.2.2:

- Overlapping factor (α): 4.
- Block size (N): 1024.
- Channel input delay[1]: 500 samples.
- Regularization parameter (δ): 1.
- Update coefficient (μ): 0.05.

[1]This is the delay on the $m_i(n)$ and $m_j(n)$ input channel shown in figure 4.15.

6.2 Auditory spatial imagery produced by the new system

In this section the first subjective experiment evaluating the ASUS system is reported. The experiment used the computer-driven graphical mapping system (i.e. the GUI) introduced in section 3.1 to allow listeners to describe the location and extent of source and reverberance imagery in both the new ASUS upmixed scenes (i.e. the 2/2 scenes) and the unprocessed two-loudspeaker audio scenes (i.e. the 2/0 scenes). Graphical mappings of source and reverberance (S and R) images were made by different subjects using the GUI, and the elicited plots were analysed using the methods described in the subjective experiments on S and R image interaction in section 3.2. The elicited images are investigated in terms of three spatial attributes: image width (i.e. extent in the horizontal plane, as observed from the listening position and measured in degrees); image distance (i.e. ego-centric range); and image direction (i.e. azimuth).

There were two basic hypotheses as to the change in S and R imagery between the 2/2 and 2/0 channel audio scenes:

1. It was anticipated that the new upmix system would create R images to the sides of the listener due to the correlated side channels (e.g. loudspeaker channels L and LS; as was measured in the electronic analysis in section 5.5.3). This hypothesis was evaluated using the GUI by looking at the spatial distribution of reported reverberance images (in the horizontal plane), using a measure such as image directional strength.[2]

[2]Image directional strength is described in section 3.1.3. It is a measure of how often a

2. The source image will remain spatially undistorted (at least, it will not be larger) for the upmixed (2/2) audio scene compared with the 2/0 scene (i.e. the source image will have a high spatial fidelity in the upmixed scene relative to the 2/0 scene).

Thus, the null and alternative hypotheses to be tested regarding attributes of auditory spatial imagery, when comparing between the 2/0 and the upmixed listening conditions, can be stated most generally as follows. Given an original two-channel recording of a solo musical performance in a concert hall, reproduction of the audio with the new 2/2 upmixer versus with a 2/0 system...

$\mathbf{H_0}$: makes no identifiable difference in spatial attributes of the source image.

$\mathbf{H_A}$: creates readily identifiable differences in spatial attributes of the source image (specifically; differences in the source image width and azimuth).

The reasoning for this is that (by definition) it is the correlated early-arriving sound components which primarily affect the auditory spatial imagery of a source image, so if the rear loudspeaker signals are uncorrelated with one of the front loudspeaker signals then the source image will be unaffected (at least, it won't be pulled in the direction of the rear loudspeakers). And as reported in the previous chapter (section 5.5.3) diagonally opposite loudspeaker signals (e.g. signals feeding the left-surround and front-right loudspeakers) had a very low correlation. This hypothesis can be evaluated

particular image type (such as a source image) is reported in a given direction for a given stimulus, generally expressed as a percentage.

by investigating whether the source image has a similar direction and similar width in both the 2/2 and 2/0 scenes.

6.2.1 Method

Set-up:

This experiment was conducted in the MARLAB (see appendix B.1). The only active loudspeakers were in the 2/2 ITU-R BS 775-1 configuration (i.e. there was no centre-speaker) as shown in figure B.1, with rear loudspeakers at $\pm 120°$. The listener sat on a rigid chair in the sweet-spot. The loudspeakers were calibrated so as to produce an equal SPL at the listening position (74 dB, unweighted, slow time averaging, using pink noise). All sound stimuli was reproduced from PC1 using the program PD (see appendix C for hardware and software details).

The GUI version was the same as for the investigation of S and R image interaction in section 3.2, and also ran on PC1. However, unlike in the previous experiments the listeners' heads were "softly" fixed by instructing them to touch the nape of their neck on a soft piece of foam (held in place using a mike-stand) behind the listening position as can be seen in figure 6.1. This ensured that the centre of the listeners' head (defined as the halfway point between the two ears, on the interaural axis) was at the sweet-spot (i.e. the point in the listening room which is equidistant to all loudspeakers). Furthermore, when judging image azimuth the listeners were told to keep their head facing forward, to the zero-degree reference marker. This was explained both verbally and formally with the instructions (see appendix D).

The markers on the inner curtain ranged progressively from "0" at the

straight-ahead, on-axis location, to "13" at ±130° (the rear loudspeakers were at ±120°). To help with spatial correspondence between the listeners "real world" environment and the virtual environment of the GUI, the subjects used a laser pointer to mark the extent and direction of the perceived images and would then draw these images using the GUI. This was done in the following way; they would hold the laser pointer and point at the end points of the perceived image and then "read-off" the marker value this corresponded to, rounded to an accuracy of one thumb width (about 0.5°).

Stimuli:

The stimuli were taken from the measurements recorded in Pollack Hall discussed in the previous chapter. Two mike configurations were used: ORTF and coincident XY. To recap., the microphones used were all B&K model 4011 cardioids, with the ORTF diaphragms spaced 17 cm apart and angled at 110°, and the coincident XY pair at 90° (i.e. at 45° to the central axis, with the diaphragm-ends of the microphones touching). All recordings were made simultaneously as shown with the microphone rig in figure 5.2. The anechoic recordings of a solo musical performance were reproduced from a loudspeaker on the stage. The recordings were of a sung-voice and a viola (see figure 5.1 for temporal and spectral details). Two loudspeaker locations were used: on-axis and 3 metres to the left (which will be called "centre" and "-3 m" positions). These permutations are summarized in table 6.1.

In this experiment, the 2/2 scene were created by reproducing the two-channel microphone recordings with the front loudspeaker pair (i.e. as with the 2/0 scenes); therefore the level of the front loudspeaker signals in the 2/0 and 2/2 scene was kept the same for a given stimulus. The output signals of the ASUS were time-aligned with the original signals to account for the processing delay (256 samples; i.e. equal to the overlapping factor S).

	Mike Conf.	Source	LS position	Scene Conf.
1	ORTF	Viola	Centre	2/0
2	ORTF	Viola	Centre	2/2
3	ORTF	Viola	3 m left	2/0
4	ORTF	Viola	3 m left	2/2
5	ORTF	Voice	Centre	2/0
6	ORTF	Voice	Centre	2/2
7	ORTF	Voice	3 m left	2/0
8	ORTF	Voice	3 m left	2/2
9	XY	Viola	Centre	2/0
10	XY	Viola	Centre	2/2
11	XY	Viola	3 m left	2/0
12	XY	Viola	3 m left	2/2
13	XY	Voice	Centre	2/0
14	XY	Voice	Centre	2/2
15	XY	Voice	3 m left	2/0
16	XY	Voice	3 m left	2/2

Table 6.1: Stimuli permutations for subjective evaluation of imagery in the ASUS using the GUI. *Scene conf.* (the last column) refers to either the reference presentation of just the two mike signals with a front loudspeaker pair (2/0) or a configuration with the rear loudspeakers being fed from the output of the ASUS algorithm (2/2). The stimuli were randomly presented in two trials (i.e. each of 16 stimuli), and the order was different for every subject.

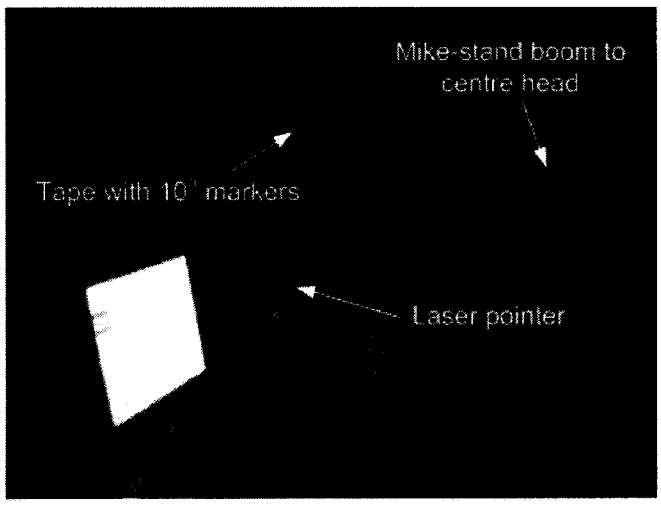


Figure 6.1: Photograph of a subject taking part in the MARLAB listening test; using the GUI to map perceived source and reverberance images experienced in the upmixed 2/2 scene and the original 2/0 scene. The inner-curtain had numbered markers at 10° intervals and was about 1 m in front of the listener. The mike-stand boom had a soft foam piece at the end which rested at the back (nape) of the neck so as to position the head at the listening position.

Subjects:

11 subjects took part in the experiment- all were members of the sound recording department at McGill University, all had experience with critical listening tests before, and 8 had taken part in previous experiments using the GUI. After reading the instructions, they had a practice session with at least 5 scenes (lasting 20-30 minutes). No feedback was given regarding the GUI response, though during this training session the subjects were often reminded by the experimenter not to move their heads from the head-rest and to use the laser-pointer. The three runs were generally done on the same day, with a break of at least 20 minutes between sessions. A typical session took about 2 hours (see response times in figure 6.2). Subjects were paid $20 for taking part in the listening tests.

6.2.2 Results

The data from the elicited scene descriptions were analysed in the same way as with the exploratory experiment using the GUI to investigate source and reverberance image interaction (reported in section 3.2). The image geometry attributes (width, range and direction) were calculated for each unique scene description for both elicited source and reverberance images; there were 33 descriptions for each of the four stimuli (11 subjects×3 runs) that were presented with both the upmixed 2/2 configuration and the original 2/0 scene (the listener had to complete the GUI description for a given stimulus before they could advance to the next stimulus). These attributes are summarized in figures 6.3 and 6.4. Density plots were also created by summing these scene descriptions, which are shown in table 6.2. Image directional strength plots were then created from these density plots; as shown in table 6.3. The time taken for a listener to draw a single sound scene was measured by the GUI; these results are shown for each subject in figure 6.2.

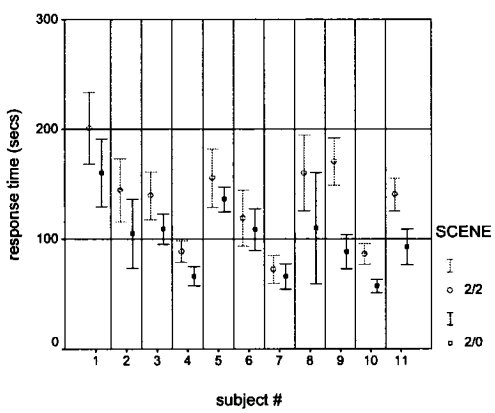


Figure 6.2: Response time for a single scene description using the GUI, grouped by subject and scene configuration. Mean and 95% confidence intervals shown.

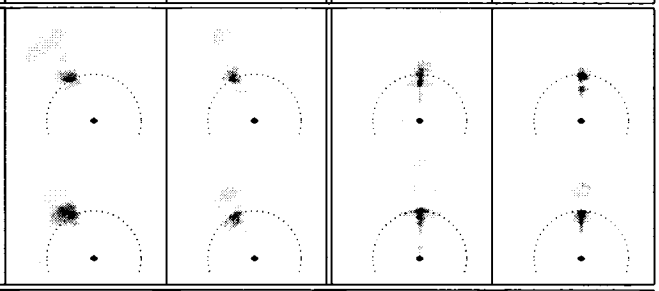
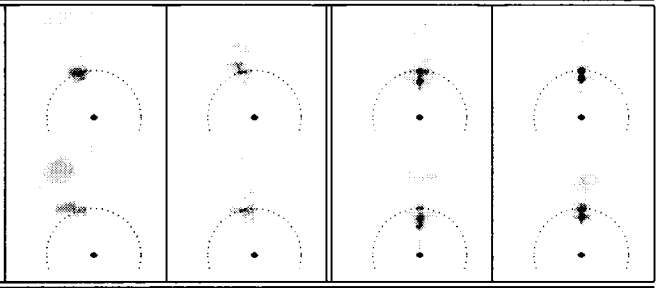
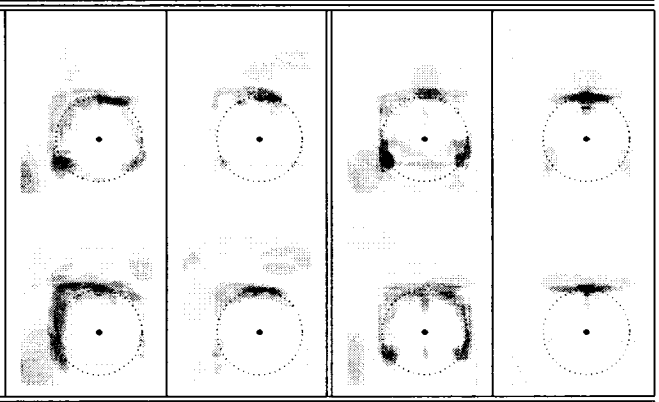
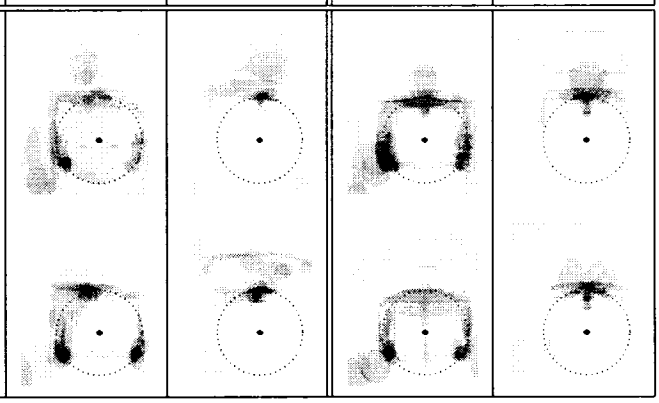
Mike conf.	Music	Source at -3 m: 2/2	Source at -3 m: 2/0	Source at centre: 2/2	Source at centre: 2/0
ORTF (S Image)	Voice				
	Viola				
XY (S Image)	Voice				
	Viola				
ORTF (R Image)	Voice				
	Viola				
XY (R Image)	Voice				
	Viola				

Table 6.2: Density plots showing elicited stable source (S) images and reverberance (R) images in the upmixed 2/2 scene and reference 2/0 scene. Each of the 16 unique audio scenes was graphically described 3 times by each of the 11 subjects- i.e. there are 33 scene descriptions for each density plot.

Image directional strength.

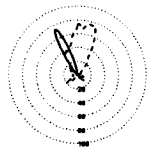
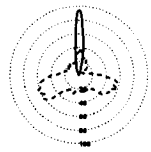
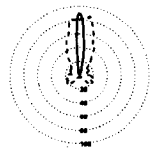
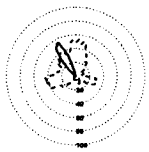
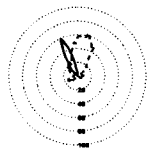
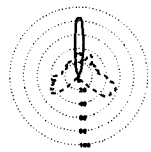
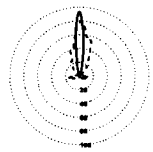
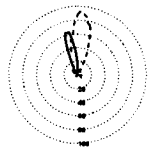
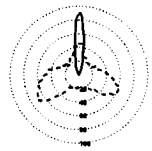
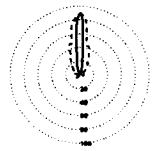
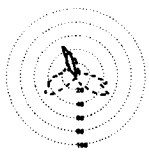
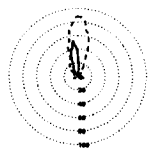
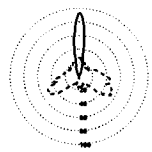
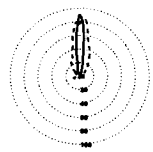
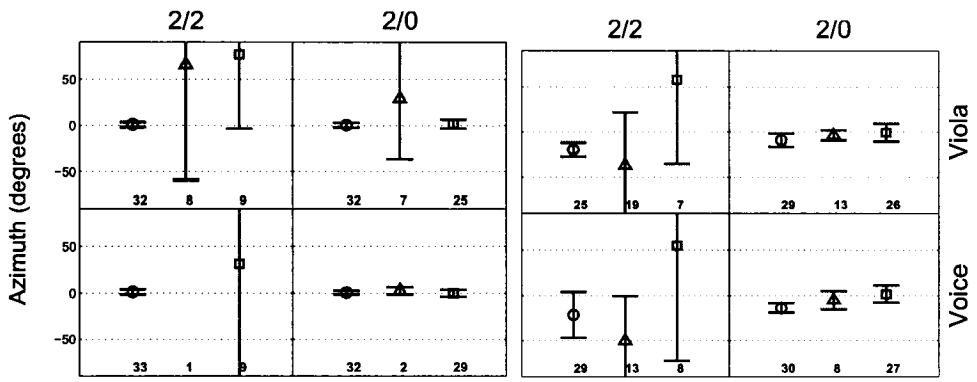
Mike conf.	Music	Source at -3 m: 2/2	Source at -3 m: 2/0	Source at centre: 2/2	Source at centre: 2/0
ORTF	Voice				
	Viola				
XY	Voice				
	Viola				

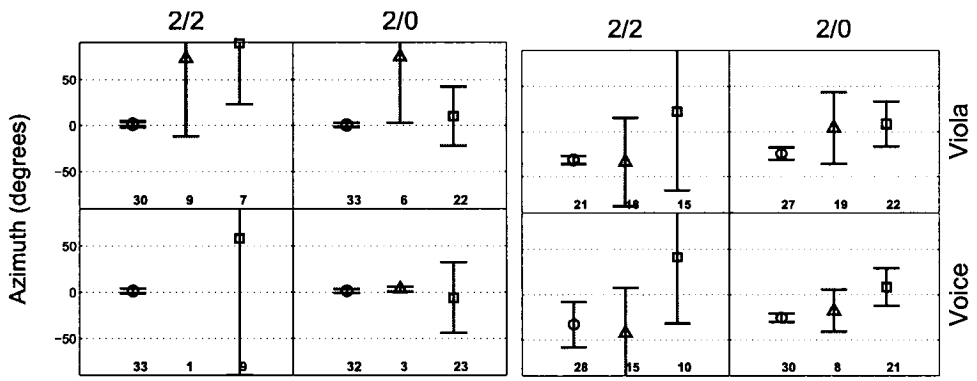
Table 6.3: Image directional strength showing the percentage of instances stable source images (solid line) or reverberance images (dashed line) were reported coming from a particular direction (see figure 3.5 for an explanation of how these are calculated). These plots ignore the reported image distance and depth. Data has been smoothed with a Hanning-shaped window of width 5° (i.e. ±2.5°) to account for the spatial resolution of the auditory system for sound localization in the horizontal plane, which has a localization accuracy for broadband sources reproduced with loudspeakers around the listener of 3°-10° (Blauert, 1997, pg. 41). The radial scale corresponds to the percentage of times an image was reported in a particular direction from all 11 (subjects)×3 (runs) graphical descriptions of each unique audio scene (the outer-most circle on each polar plot corresponds to 100%).

Statistical analysis of GUI responses.



(a) XY. Centre-source.

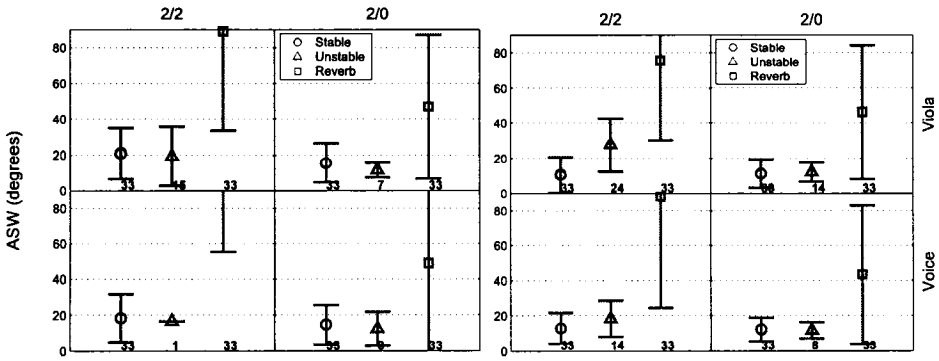
(b) XY. -3 m source.



(c) ORTF. Centre-source.

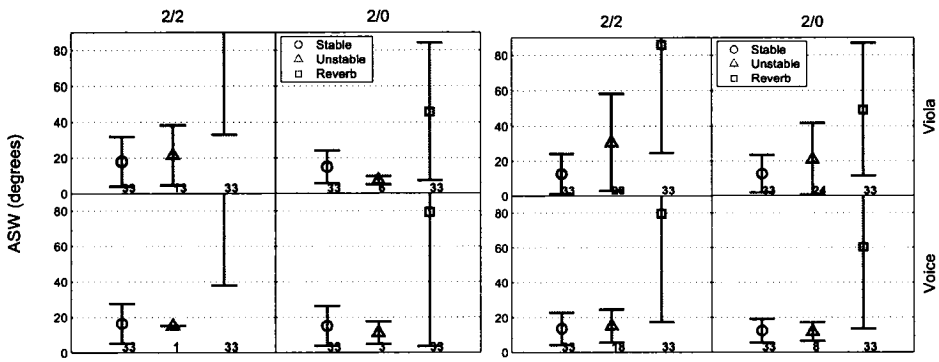
(d) ORTF. -3 m source.

Figure 6.3: Statistical analysis of image azimuth (i.e. direction of image centre, in degrees) for elicited stable source (○), unstable source (△) and reverberance (□) images. The number at the bottom of each data marker shows the number of cases used to calculate the mean and 95% confidence intervals; it was not the same each time since if more than one separate image could be elicited for a single scene description, a single azimuth could not be calculated for that image type.



(a) XY. Centre-source. Mean ASW for the R image in the bottom-left panel is 92°.

(b) XY. -3 m source.



(c) ORTF. Centre-source. Mean ASW for R image in top-left panel is 90°, in bottom-left panel it is 105°.

(d) ORTF. -3 m source.

Figure 6.4: As figure 6.3 showing reported image width (in degrees) for stable source (○), unstable source (△) and reverberance (□) images. Width was calculated as the sum of image spread in the horizontal plane for a given scene description. The number of source images varies because subjects could draw either (or both) a stable or unstable source image.

The dependant variables investigated using an ANOVA test and their abbreviations were:

- The location of the loudspeaker in the recording (either on-axis or 3 m to the side) [POSITION].
- Azimuth of reverberance image [AZM_REV].
- Width of stable source image [WIDTH_ST].
- Width of unstable source image [WIDTH_UNSTABLE].
- Width of reverberance image [WIDTH_REV].
- Distance of stable source image [DIST_ST].
- Distance of reverberance image [DIST_REV].

Distance and azimuth data for the unstable source images were not used because there was typically half as much valid data per stimulus because generally more than one image was drawn per stimulus. A 1-way ANOVA was conducted on all of the data with the dependant variable as the microphone type used in the recording- to see whether the ORTF or XY pair affected each of the above dependant variables. It was found that this only affected reported AZM_ST [$F(1,475) = 13.5, p < .001$]- which shall be discussed later.

The independent variables investigated were:

- The location of the loudspeaker in the recording (either on-axis or 3 m to the side) [POSITION].
- The original recording source either a viola or a sung voice [MUSIC].

- Scene configuration: either the "reference" 2/0 scene or the upmixed 2/2 scene.

The data was split into four groups; corresponding to the four quadrants of the graphs shown in this section (see figures 6.3 and 6.4). That is, the four groups were for the two recording locations [POSITION]- with POSITION= 1 corresponding to the loudspeaker at -3 m to the centre line and POSITION= 2 corresponding to the loudspeaker equi-distant to the microphone pair; and the two stimuli [MUSIC]- with MUSIC= 1 the voice and MUSIC= 2 the viola.

Four each of these four groups, the results for each of the seven dependant variables were analyzed using an ANOVA model with the SCENE CONFIGURATION as the fixed factor. The results are summarized in table 6.4.

	-3 m		Centre	
Dependant variable:	Voice	Viola	Voice	Viola
AZM_ST	$p = 0.000$	$p = 0.032$	$p = 0.219$	$p = 0.371$
AZM_REV	$p = 0.027$	$p = 0.004$	$p = 0.000$	$p = 0.024$
WIDTH_ST	$p = 0.243$	$p = 0.164$	$p = 0.056$	$p = 0.219$
WIDTH_UNSTABLE	$p = 0.000$	$p = 0.000$	$p = 0.000$	$p = 0.222$
WIDTH_REV	$p = 0.000$	$p = 0.001$	$p = 0.000$	$p = 0.000$
DIST_ST	$p = 0.918$	$p = 0.646$	$p = 0.251$	$p = 0.374$
DIST_REV	$p = 0.603$	$p = 0.337$	$p = 0.532$	$p = 0.663$

Table 6.4: Results of ANOVA analysis to see statistical significance of affect of scene configuration (2/0 or 2/2) on changes in S and R image spatial properties for different positions of loudspeaker used in the recording, and different stimuli.

6.2.3 Discussion

The differences between auditory spatial imagery in the reference 2/0 and upmixed 2/2 scenes are now discussed for each of the following dependant variables:

- Reported image direction.
- Reported image range.
- Reported image width.
- Image directional strength.

Reported source image direction

The microphone configuration factor significantly affected the reported direction of the source images. This trend can be seen by eye-ball inspection of the contour plots when looking at the sources from the -3 m location (i.e. the eight contour plots on the left of table 6.3). To summarize the trend: when the loudspeaker was at the -3 m location, the source (S) image was heard closer to the centre axis with the XY mike configuration than with the ORTF set-up. This is intuitively understood by considering the arrival time difference from the source to each mike in the XY or ORTF pair- there would be no time difference for the coincident (XY) pair. This is confirmed by analysis of the adaptive filter coefficients shown in figure 5.6(b) (page 199). Looking at the filter, it can be seen that for the source at -3 m, the inter-mike amplitude difference is similar for both the XY and ORTF arrangements, but the inter-mike time difference is about 0.45 ms for the spaced ORTF pair yet (obviously) zero for the coincident XY pair. It was found in a previous

experiment (Usher and Woszczyk, 2003) that a pure delay of 0.45 ms between a typical two-channel musical signal feeding a front loudspeaker pair at ±30° would give an image located at about 15° (i.e. towards the direction of the non-delayed loudspeaker). Therefore, the time-of-arrival cue pulled the source image farther in the direction of the source offset with the ORTF configuration.

There was an interaction effect of scene type and reported S image direction; but was only significant for the -3 m source. In other words, with the voice or viola recording made with the loudspeaker 3 m off the central microphone axis, the reported source image direction was significantly different in the 2/2 and 2/0 scenes. This was confirmed using an ANOVA and the trend held for both voice and viola. This is not immediately obvious from inspection of the contour plots, but can be seen by looking at subfigures (b) and (d) of figure 6.3. For voice, the variation in reported image direction is more for the 2/2 scene than for the 2/0 scene (95% confidence intervals of approximately 50° and 10°, respectively). Why the consistency is less for the 2/2 than for the 2/0 scene could be due to there being *occasional* components of direct and early sound (i.e. those sound components which affect spatial aspects of the source image) being radiated by the rear loudspeakers and distorting the S image due to principles which explain coherent amplitude panning (see section 2.1.8).

Reported source and reverberance image distance

As found in previous work discussed in this thesis, reported image range is very inconsistent both within and between subjects. The ANOVA results are given table 6.4 and show that the reported image distances were not

significantly affected by the upmixing process.

Reported source and reverberance image width

The width of the stable source image was not affected by the scene configuration. (This was statistically verified with the ANOVA summarized in table 6.4.) The unstable image width (WIDTH_UNSTABLE) was affected by the SCENE factor for all stimuli except the viola reproduced from the centre loudspeaker position (though it should be bared in mind that there was about half as much data for the unstable images than for elicited stable source images). Of course, it is not surprising that the width of the elicited reverberance (R) images were affected by the SCENE factor- it can clearly be seen from the raw density plots that the reverberance images are wider and more enveloping for the 2/2 scene than the 2/0 case. The increase in the size (i.e. area) of the elicited R images may have accounted for the extra time it took for subjects to describe the 2/2 scenes compared with the 2/0 scenes- a trend consistent across subjects as shown in figure 6.2.

Interaction of source and reverberance images in 2/0 and 2/2 scenes

Interaction between S and R images was informally investigated using the image directional strength plots in table 6.3. These show the relative number of times an image was reported as a function of direction, and some basic trends between the upmixed (2/2) and original (2/0) audio scenes can be discerned. As would be expected, for the 2/0 scene the R images are nearly always localized between the front loudspeaker pairs, as was the S image.

For the upmixed 2/2 scene, S and R images were rarely reported from

the same direction (i.e. either the S or R image was dominant in a particular direction). Looking at the image directional strength polar plots, it can be seen that for the centre-sources the instances of R images heard at 0° is <50%. On the other hand, for the 2/0 scene R images are reported from this straight-ahead direction between 80% and 90% of instances (i.e. a similar amount of times that S images were heard from this direction). In other words, in the 2/0 scene the S and R images are co-localized more than with the upmixed 2/2 scene; though a further statistical analysis is needed to confirm this general trend.

The lack of co-location of S and R images for the 2/2 scene in this experiment supports the idea that the reproduction of reverberation from the rear speakers increased the perceptual separation of the source and reverberance image streams compared with the 2/0 case. However, the reverberance imagery was not evenly distributed to the side of the listeners; there was a noticeable detent in the direction of the rear loudspeakers. A preference experiment was conducted to see whether reducing the level of the rear loudspeakers could help create an enhanced overall sound quality.

6.3 Global preference experiment

A final experiment was conducted to see if two modifications of the ASUS can positively enhance the listening experience. This experiment will also address one of the original aims; that the new system should be preferred to a conventional 2/0 reproduction. Rather than a descriptive analysis of spatial imagery, as with the previous test, a simple preference paradigm was used where all sound scenes were compared in a blind pairwise manner. The

modifications were time delayed rear channels and attenuated rear channels. The two other sound scene configurations were the unmodified ASUS (as was used in the GUI experiment just described) and a conventional 2/0 scene where just the front loudspeaker pair are active; as summarized in table 6.5.

Config. #	Scene configuration
1	ASUS (unmodified)
2	2/0
3	ASUS : rears with 10 ms delay
4	ASUS : rears with -6 dB gain

Table 6.5: Audio scene configurations for preference experiment. Configuration #1 is with the unmodified ASUS configuration used in the previous GUI experiment (i.e. there was no cross-talk introduced to the input signals). Configuration #2 is with only the front loudspeakers active (i.e. a 2/0 scene). In configurations #3 and #4, the rear speaker signals are modified with either a time delay or a signal attenuation. All ASUS scenes are created using the algorithm parameters as in the GUI subjective experiment described earlier in this chapter.

As discussed in the literature review in section 2.2, delay of the rear loudspeakers relative to the front is common in audio upmixing systems. The idea behind this is to minimize perceptual fusion of sound components present in the front and rear channels which contribute to source imagery, which could distort both the timbral and spatial properties of the source image. This is related to the precedence effect; when a sound is reproduced from a loudspeaker in front of the listener and a delayed copy from a side loudspeaker, distortion of the source image (e.g. in terms of colouration or spatial geometry) is minimal when the delay is about 10 ms (e.g. Olive and Toole, 1989; Bech, 1998) and will help ensure the source image is located in the front of the audio scene. Typical rear-loudspeaker channel delays used in blind audio upmixing systems are about 10 ms (Rumsey, 1999; Irwan and Aarts, 2002b), though this is variable with Pro Logic II from 0-30 ms

(Dressler, 2000).

The attenuated rear-channel scene was investigated because in the GUI experiment just reported, subjects sometimes commented that the rear-loudspeaker level seemed too loud. This may account for the reverberance images often being reported in the direction of the rear loudspeakers, as shown by the image directional strength plots in table 6.3. A 6 dB attenuation was chosen following extensive informal listening by the author, as this seemed to create a more natural-sounding sense of envelopment (the reverberance imagery seemed more homogenous and did not distract from the source imagery).

6.3.1 Method

Set-up:

This experiment was conducted at the Banff Centre in an acoustically treated editing and mixing room (approximately 50 m^3, estimated RT_{60}<0.5 s). As with the previous experiment, four loudspeakers were used arranged in the conventional 2/2 ITU-R BS 775-1 configuration (no centre-speaker), with rear loudspeakers at ±120° (the loudspeakers used in this experiment were all model type 1031 manufactured by Genelec). The listener sat on a non-rotating chair at the sweet-spot, 2.3 m from each loudspeaker. The loudspeakers were calibrated so as to produce an equal SPL at the listening position (74±0.5 dB, unweighted, slow time averaging, using pink noise). The stimuli were burnt onto a DVD disc (recorded at 44.1 kHz, 16 bit), and the music was presented with four loudspeakers from a DVD-A player.

The method of paired comparison was used to evaluate the ASUS in terms of overall preference (as recommended by IEC 268-13). The subject

was presented with a two-way switch labeled A or B, which allowed the subject to freely switch between the two audio scenes using the switching device shown in figure 6.6 and to report which sound scene was preferred. Stimulus A or B corresponded to one of four scenes; a variant of the ASUS or the 2/0 scene. This AB preference method was the same as used in the evaluation of various upmix systems in two previous studies: Irwan and Aarts (2002b) investigated preference for four surround sound configurations (but did not have a stereo 2/0 configuration). Rumsey (1999) also investigated front-imaging and spatial impression (as suggested in ITU-R BS 1116) with two commercially available upmix systems and two custom systems (only one of which used the centre channel) with a double blind AB comparison method, where the listener was asked to rate the front-image and spatial impression of each stimulus using a 10 point scale.

In this preference experiment, the subjects were presented with two sound scenes, A or B, which they could freely select using the signal switcher shown in figure 6.6. With a computer program, they were asked: "`Which sound scene do you prefer: A or B?`", and responded by selecting either the "A" or "B" icon on the screen (and were prompted to confirm their choice). It was emphasised that the audio system they were evaluating was intended for use in a domestic home environment for entertainment purposes, so they should think about the preference task as if they were evaluating a product in a shop which they were going to buy and bring home.

Two groups of people undertook the experiment: 5 audio engineers and 11 musicians. The engineers were all past Tonmeister students, each with at least three years of experience with sound recording practice. The musicians were enroled on an intensive music performance or composition program at the Banff Centre (most of whom are professional). How these groups can be

distinguished in terms of experience will be discussed later.

Stimuli:

The stimuli were the same as used in the previous GUI experiment just described, except only the ORTF recording configuration was used. Thus, there were four possible combinations of source-recordings and reproduction position, as shown in table 6.6.

Fragment #	Source	Loudspeaker recording position
1	Viola	3 m left
2	Viola	3 m left
3	Voice	Centre
4	Voice	Centre

Table 6.6: Stimuli used in preference experiment. These recordings are the same as used in the GUI experiment and electronic analyses in chapter 5; made using an ORTF arranged mike pair in Pollack hall. The sound source was a loudspeaker on stage reproducing an anechoicly recorded solo music performance. The loudspeaker position was either equidistant to each microphone (i.e. "centre") or 3 m off-centre.

Spectral and temporal details of the stimuli are given in figure 5.1. The subject could advance to the next DVD track at any time by pressing the "next track" button, or could replay the track by pressing the "repeat track" button (i.e. the subject could take as long as they want to listen to the tracks and decide whether A or B was preferred). The DVD play-mode was random, so the trial order was randomized (the subject would write down the track playing order on a piece of paper). Each of the four fragments were presented in four scene configurations, as shown in table 6.5.

The A-B comparison was made for each of the four fragments with stimulus A corresponding to this fragment presented with one of the four scene configurations in table 6.5, and stimulus B with the same fragment presented

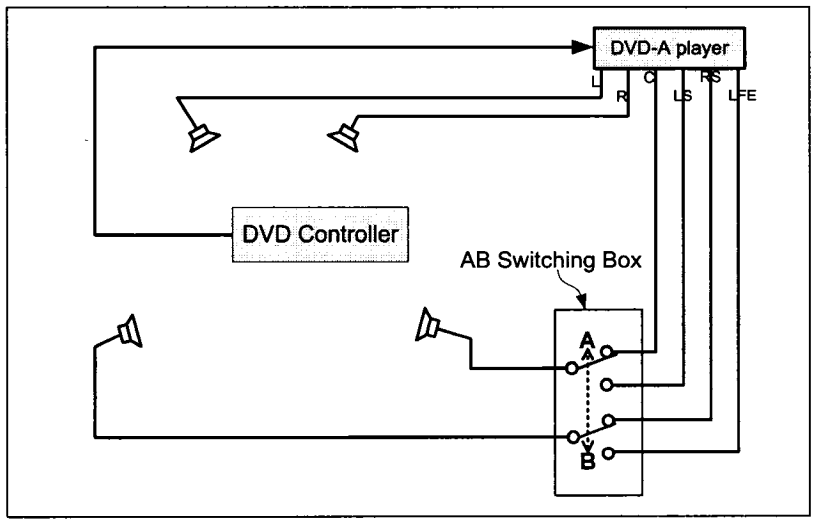


Figure 6.5: Signal schematic for preference experiment. All audio outputs are from a DVD-A player (the LFE channel is a normal, full-bandwidth signal). The rear-loudspeaker channel switching device is a passive switching box with an "A-B" clicking switch (see photograph in Fig. 6.6). The channel gain-trims are not shown.

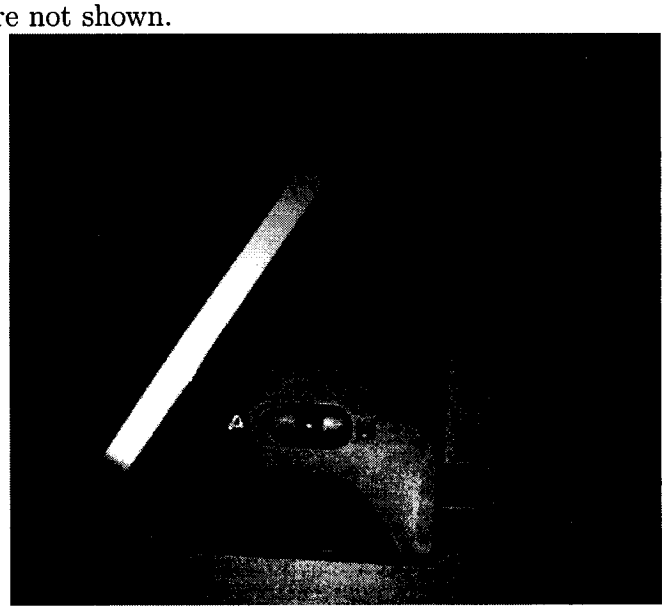


Figure 6.6: Photograph of the AB switching box used in experiment.

with a different scene configuration. Therefore there were 6 pair-wise comparisons for each of the four fragments, giving 24 unique paired comparisons (see table 6.7). This was presented twice to each subject, with the A-B stimuli order reversed for the second presentation.

Permutation #	Fragment #	Config. # (A)	Config. # (B)
1	1	1	2
2	1	1	3
3	1	1	4
4	1	2	3
5	1	2	4
6	1	3	4
7	2	1	2
8	2	1	3
9	2	1	4
10	2	2	3
11	2	2	4
12	2	3	4
13	3	1	2
14	3	1	3
15	3	1	4
16	3	2	3
17	3	2	4
18	3	3	4
19	4	1	2
20	4	1	3
21	4	1	4
22	4	2	3
23	4	2	4
24	4	3	4

Table 6.7: The 24 stimuli permutations used in the preference experiment. Config. #(A) or #(B) is the upmix configuration for options A or B on the switch-box. Scene 1 is the unmodified ASUS arrangement; 2 is the original 2/0 scene; 3 is the ASUS with delayed rear loudspeaker channels; and scene 4 is the ASUS with the rear loudspeaker channels attenuated by 6 dB. The actual presentation order was randomized, and the test was repeated once for each subject.

Subject task:

The subject was presented the twenty-four excerpts of music twice, with a 5-20 minute break in between. Using a computer GUI, they were asked: "`Which sound scene do you prefer: A or B`?" Once they selected either option, a pop-up window prompted the subject to confirm and advance the DVD to the next track. The response time was measured using the GUI program. As mentioned, it was emphasised that the audio system they were evaluating was intended for use in a domestic home environment for entertainment purposes, so they should think about the preference task as if they were evaluating a product which they might purchase.

6.3.2 Results

A type III sum of squares general linear model analysis was conducted to investigate how various independent variables affected the response time. This method was chosen as it is robust to non-normal distributions of data (SPSS, 1995), which is common in time-series data as the lower limit is always bound to zero. The fixed factors were; GROUP (musicians or engineers); SUBJECT (which is of course nested within the GROUP factor); MUSIC (viola or voice); and POSITION (of loudspeaker during recording- i.e. -3 m or centre). Because there were only 5 subjects in the engineer group but 10 in the musician group, the GROUP factor had to be investigated separately using an ANOVA with just GROUP and TIME; the GROUP factor was found to be significant $[F(1,689) = 6.14, p < .013]$. All other factors significantly affected the response time except POSITION, as summarized in table 6.8.

Results for the paired comparisons are shown in figure 6.9; which shows how often a particular stimulus was (dis)preferred over an other. The data

Factor	DoF	F	Sig.
SUBJECT	13	40.5	<0.01
MUSIC	1	24.1	<0.001
POSITION	1	2.9	0.089

Table 6.8: Type III ANOVA analysis of response times for preference experiment.

is split into the musician and engineer group. The total number of trials (nTrials) was 24 scene pairs×2 runs×(the number of subjects), which was 240 for the engineer group and 480 for the musician group. 95% confidence intervals ($\pm 2\sigma$) were calculated according to (6.1):

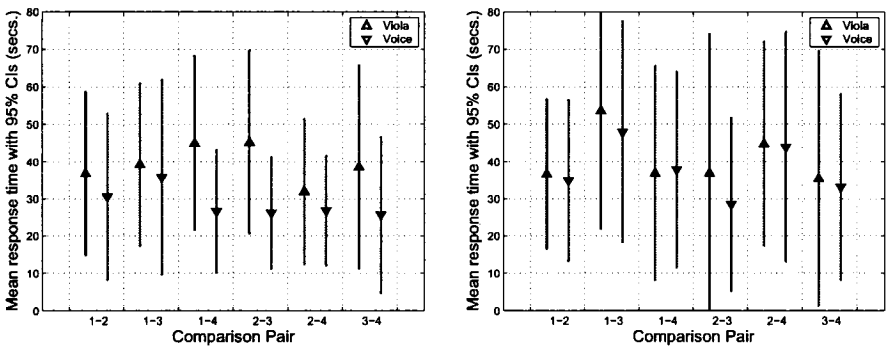
$$\sigma = \sqrt{\text{nTrials} \times P(A)(1 - P(A))}, \tag{6.1}$$

where $P(A)$ is the probability of the subject picking a scene configuration A by chance (which was 0.25 as there were four scenes).

6.3.3 Discussion

The response time data summarized in figure 6.7 show an interesting and clear trend for both the musician and engineer responses: that it took longer to report the preference for the viola piece than the voice piece. A common comment from experiment participants after the test was that the voice was more "revealing" about undesirable timbral and spatial features of the stimuli (such as spatial distortion of the source image). This may be related to the more temporally complex nature of the voice recording compared with the viola. Also, as shown by the spectral envelope in figure 5.1, the voice stimulus has more high frequency energy than the viola and Bech (1998) showed that sound reflections with a brighter timbre (particularly with frequencies

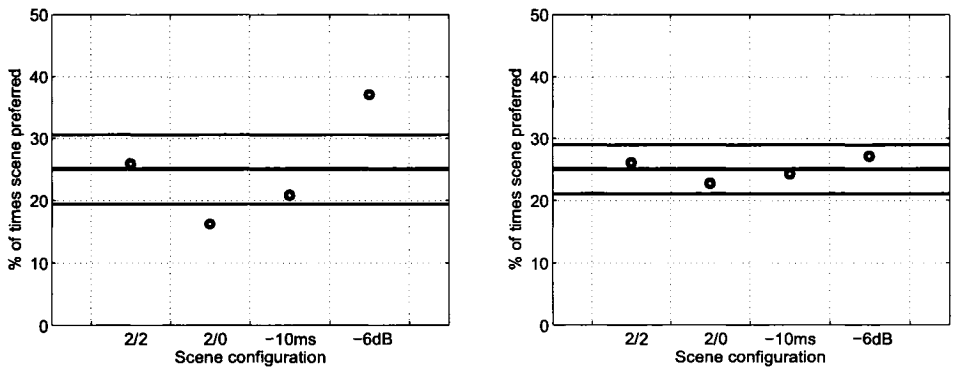
Response time



(a) Audio engineer group (5 subjects). (b) Musician group (11 subjects).

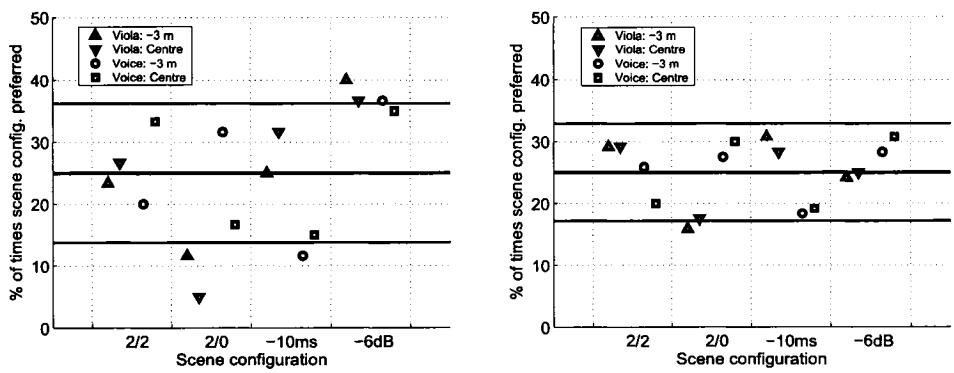
Figure 6.7: Raw response times for different scene-pair comparisons. The recording source position is ignored as it was found to have no statistical significance on response times. Scene 1 is the unmodified 2/2 upmix system, with no cross-talk on the input signals; scene 2 is the "reference" original 2/0; scene 3 is with a -10 ms delay on rear loudspeaker channels; scene 4 is with a -6 dB gain on the rear loudspeaker channels. Due to the non-normal distribution of response times, the confidence intervals look misleading; in fact, timing data was affected by both subject and the recorded source (i.e. viola or voice).

Preference analysis



(a) Audio engineer group (5 subjects). (b) Musician group (11 subjects).

Figure 6.8: Plot of percentage of preferred choices out of 4 possible scene configurations. Scenes were presented as a paired AB comparison and results are grouped by scene configuration. All scenes were the 2/2 ASUS except scene 2/0 (which was just the front loudspeaker pair). Central solid line shows likelihood of preferring a scene by chance (25%, i.e. if the subjects randomly pressed A or B) and flanking lines are 95% CI's. If the marker is above the upper-most line, then this scene configuration was preferred significantly more than the others. If the marker is between the upper and lower lines; this scene was neither preferred nor less preferred. And if it is below all lines, then this scene was preferred less than the other scenes. There were 4 stimuli and 2 presentations; so 240 responses for the audio engineer group and 480 for the musician group.



(a) Audio engineer group (5 subjects). (b) Musician group (11 subjects).

Figure 6.9: As figure 6.8 but stimulus effect is shown.

above 2 kHz) are heard easier (i.e. at a lower level) than dull-sounding reflections. Also, transient sources are more accurately located than smooth-envelope sources (Moore, 1997, pg.241), so the voice may reveal the presence of loudspeakers more than the viola making the preference decision easier and quicker. Referring back to the electroacoustic analysis of the ASUS, it can be seen that the misadjustment is slightly larger (i.e. the rear-loudspeaker signal magnitudes are relatively larger) for the voice than the viola signals by about 4 dB (figure 5.4); so maybe this was due to source-image components ("direct sound") which were reproduced from the rear loudspeakers.

From the results of the preference choice analyses shown in figures 6.8 and 6.9, it can be seen that for the musician group there is only one statistically significant trend. The only significant trend for the musician group was for the viola recording at -3 m, which for the 2/0 scene was reported as being less preferred than the other scenes (see figure 6.9); this stimulus was also generally dispreferred by the engineer group. This lack of clear preference for the musician group was surprising, considering the large difference in sound scenes (such as the 2/0 and 2/2 comparisons). However, even though the audio-engineer group preferred the upmixed scene significantly more than the 2/0 scene (except for the voice at -3 m), which is a principle subjective design criteria (section 3.3), the result should be not be interpreted too generally as the engineer group are simply not "average listeners" in terms of experience. This finding is different from Rumsey (1999), who found that from a listening panel of 22 experienced listeners (Tonmeister students and professional audio engineers), a conventional 2/0 audio scene was generally preferred *more* than an upmixed 2/2 or 3/2 scene (created using four different commercial upmixers) for music recordings, though there was no interpretation of the statistical significance in that study.

Other studies have found that the response variation for timbral and spatial sound quality evaluations of loudspeaker audio scenes are higher for non-experienced listener groups than with experienced groups (Bech, 1992; Olive, 2003; Rumsey et al., 2005b). How to clarify the distinction between these groups is not clear; Bech (1992) suggests the following factors should be considered:

- Familiarity with "live" sound, namely, concerts or experience in playing an instrument.
- Experience in critical listening to live or reproduced sound.
- General aptitude for detecting sonic differences in reproduced sound.

Both groups who undertook this preference experiment could therefore be considered experienced by the first two criteria. However, the last is less applicable for the musician group and as this category is the most applicable for the paired-comparison method employed, it seems reasonable to call the audio-engineer group the "more experienced" group of critical listeners.

In an recently reported experiment by Rumsey et al. (2005b) with 16-21 expert listeners and 40 non-experienced ("näive") listeners, it was found that the overall (basic) sound quality was affected by timbral and spatial ratings in a different way for each group. Although both groups considered timbral features to be the dominant factor influencing their overall sound quality rating, näive listeners seemed not to consider the frontal spatial imagery at all but instead just the rear imagery. For expert listeners, generally speaking the reverse was true; it was the spatial sound quality of the front images which seemed to matter. This contrasts with the results in the preference experiment reported here, where rear spatial image quality was obviously a

factor for the experienced listener group. If rear spatial image quality did not influence overall preference, then one would expect all preference ratings for the upmixed sound-scenes to be equal (at least for the unmodified and -6 dB ASUS configuration, as the timbral character of the delayed sound scene may have been modified due to the acoustic interference between the front and rear loudspeaker signals).

It is particularly interesting that the engineer group generally preferred the 2/2 scene with the 10 ms delay *less* than the other scenes; significantly less for the -3 m viola stimuli. This might be because the 10 ms delay destroyed the pair-wise amplitude panning between the (correlated) reverberation components in the front and rear loudspeakers, with the R image collapsing to the front speakers due to the precedence effect. Looking at the cross-correlation analysis of the side loudspeaker channels $R - RS$ (figure 5.22(c)), it can be seen that these channels are quite correlated (with a coherence close to 0.8).

As figure 6.9 shows, the scene with the -6 dB rear channels was significantly preferred over all other scenes for the engineer group, though this was close to the level of chance for the centre-voice stimuli. Rumsey (1999) reported from informal listening that one of the upmixers in his study could create an enhanced listening experience by reducing the output level of the rear loudspeaker channels from its default setting. Using the cross-talk mechanism introduced in chapter 5 (section 5.4), the gain of the output signals would be reduced. For instance, considering the voice recording at -3 m, the mike signal coherence is about 0.8 (see figure 5.22), and the misadjustment is between -10 and -5 dB (figure 5.5); i.e. the level from the rear loudspeakers is about half the loudness of the front loudspeakers. According to the empirical analysis summarized in figure 5.19(c), introducing an input cross-talk of -5 dB would reduce the output level of the ASUS by a further 5-10 dB. This

further supports the case for using the cross-talk mechanism; i.e. for dealing with recordings containing hard-panning, and also to reduce the level of the signals radiated by the rear loudspeakers.

Chapter 7

Conclusions

7.1 Context of work

Multichannel loudspeaker audio is a ubiquitous phenomenon in home theatre and car entertainment systems throughout the world today. The benefits of reproducing recorded music with surrounding loudspeakers are two-fold: Allowing a more immersive and natural sounding impression of the recording environment- an increased sense of *presence* of the musician and their environment- and with secondary effects of making it easier to discern spatial nuances in the musical performance and reveal subtle aspects of the recorded sound which would be otherwise hidden with two-loudspeaker reproduction; *"A good [audio] system not only allows listeners to hear sources at different directions, but also improves a listeners' ability to understand simultaneous sources and monitor a complex auditory scene"* (Shinn-Cunningham, 2006).

A device that can create from our blooming two-channel music collections an additional set of signals which can be reproduced with surrounding loud-

speakers is therefore more relevant than ever. This thesis is about such an *audio upmixer* and addresses the major shortcomings of existing upmixers. These shortcomings have been previously ignored due to certain assumptions about the music intended to be upmixed: That the sound recordings are pop music created with intensity-panned audio mixes. The new upmix system is designed for use with both these recordings and recordings made using multiple spaced microphones; a method generally employed for recording concert-hall musical performances. It was shown in electronic measurements that music recorded with spaced microphone pairs could not be processed by extant upmixers in the same way as for intensity-panned recordings and led to undesirable processing artifacts, such as rear loudspeaker signals which were correlated with both front loudspeaker signals and could therefore distort the frontal source image. In the new system, time-alignment of the input signals ensures that the rear loudspeaker signals minimally affect source imagery. The new upmixer provides a consistent audio quality for a variety of recording methods (including those with time and amplitude panning) using a novel implementation of an adaptive signal processing algorithm.

Results of listening tests reported in this dissertation using non-verbal description techniques have corroborated conclusions of previous work using free verbal elicitation (Berg and Rumsey, 1999) to show that two aspects of imagery in reproduced sound scenes have meaningful perceptual relevance to listeners: the perception of sound relating to spatial aspects of the recorded sound source (source imagery) and sound relating to spatial aspects of the recording environment (reverberance imagery). Other work has shown that these spatial aspects of perceived sound significantly affect overall sound preference (Zacharov and Koivuniemi, 2001b) and sound quality ratings (Rumsey et al., 2005a). The temporal factors affecting the distinction between source and reverberance imagery are thought to be the same principles relating to

the precedence effect; whereby early arriving, high level sound (direct sound and early reflections) primarily affect the source image (Morimoto, 2001) and later arriving low level sound (reverberation) affect the reverberance image (Soulodre et al., 2003). A graphical mapping technique was used to show how the perceptual distinction between source and reverberance images is affected by their perceived spatial separation. Findings of the new work support an explanation of spatial interaction effects between source and reverberance images in terms of a target/ masker paradigm. This is substantiated by the results of previous studies looking at how spatially separating (Plomp, 1976; Begault and Erbe, 1994; Shinn-Cunningham et al., 2001) or spatially redistributing (Hawley et al., 2004) a masking sound (i.e. the reverberance image) can increase the intelligibility of a target sound (i.e. the source image).

7.2 Summary of work (original contributions)

7.2.1 Methods for describing spatial sound quality

Various words used to describe perceived sound were reviewed: auditory objects, auditory streams, perceived spatial quality and auditory spatial impression. An *auditory image* was deemed to be a useful term for describing experiences with reproduced sound scenes. An image was defined as a representation of a perceived sound object in terms of its timbral and spatial attributes. Furthermore, auditory spatial imagery was defined as those parts of a perceptual sound object that can be described in terms of physical coordinate space, such as its perceived location and size. A variety of methods for investigating auditory spatial imagery in reproduced sound scenes was summarized, and it was concluded that a graphical mapping system enables

an effective, direct translation of the the spatial attributes of auditory images into a form which can be statistically analysed.

7.2.2 Existing approaches to audio upmixing

A dozen electroacoustic approaches to increasing spatial sound quality were reviewed. The study was restricted to devices which used loudspeakers to reproduce sound. These can use both passive analog implementations and digital signal processing techniques (i.e. signal processing solutions rather than new transduction approaches which use novel loudspeaker construction). Five audio upmixers for converting unencoded two-channel audio recordings to four or five signals for reproduction with conventional surround loudspeakers were discussed in depth. It was found that these "blind" or "unsupervised" upmixers rely on a common fundamental assumption; that a single dominant image direction exists at a given time and has been created using amplitude-panning techniques.

The methods for finding the principal image direction vary; e.g. by an estimate of the level of each input channel (Choi et al., 1995; Griesinger, 1996a), or using a bootstrapped mechanism which aligns the magnitude of the input signals using a comparator to minimize the difference signal (Gundry, 2001; Irwan and Aarts, 2002b). None of the methods investigated had a mechanism to deal with input signals where the direct sound component arrived in one channel before the other (such as occurs with time-delay panning). Furthermore, the level alignment procedure was generally a global level adjustment rather than a detailed frequency-dependent gain (an exception was the Avendano and Jot (2002) system, which could discriminate between correlated and uncorrelated sound components as a function of frequency, but could not dis-

criminate between correlated and uncorrelated components within the same band).

The new upmixer is best described as an "ambiance extractor" (Avendano and Jot, 2002) using a "natural spatialization algorithm" (Rumsey, 1999), as it is designed to extract those sound components which affect reverberance imagery rather than source imagery, and radiate these components with loudspeakers around the listener. This is in contrast to "repanning" systems such as "Trifield" (Gerzon, 1992a) which aim to affect primarily the source image components by re-radiating them with at least three loudspeakers. The new upmixer would therefore compliment such repanning systems if, for example, the listener wanted to use all five loudspeakers in a conventional 3/2 ("5.1") loudspeaker system.

7.2.3 Interaction of source and reverberance imagery in reproduced sound scenes

In order to provide a sensitive means for reporting auditory spatial imagery in multichannel loudspeaker audio scenes, a new computer-driven graphical mapping system was developed. This enabled the spatial geometry of auditory source and reverberance images to be visualized with a two-dimensional drawing. Ellipses drawn with the GUI to represent the location and spatial extent of the auditory image in the horizontal plane could be analysed to determine image width, distance and range. The computer-driven graphical user interface (GUI) provides a means to investigate the spatial distribution of auditory images as a function of azimuth with a new measure called the *image direction strength* (IDS). IDS can give an indication as to the homogeneity of imagery in audio scenes; the reported distribution of source or

reverberance images over space around the listener. Other image attributes which have been identified in previous work as being perceptually salient (such as image width, azimuth and distance; Mason et al., 2001) were also measured from the elicited image descriptions.

A subjective experiment using the GUI was undertaken to investigate the spatial interaction of source and reverberance images in multichannel loudspeaker audio scenes:

1. Auditory spatial imagery of source and reverberance images as they are pair-wise amplitude panned around a listener.
2. Interaction of source and reverberance images as a function of their perceived spatial separation.

The main findings were:

- Control of both reverberance image direction and width involves similar amplitude panning principles as for source images. Pair-wise panning of *source* images to the side of the listener has been investigated in previous work (e.g. Theile and Plenge, 1977; Pulkki and Karjalainen, 2001) and the new work in the thesis suggests that generalization to *reverberance* imagery is, at least to some extent, possible.
- Perceived width and direction of a source image panned at 0° azimuth was significantly affected (i.e. distorted) by the panned direction of a reverberance image. This spatial distortion was reduced as the perceived separation between source and reverberance images was increased. These findings support the analogy between source and reverberance imagery and a target/ masker paradigm; whereby a source

(target) image can be unmasked to increase semantic understanding (increased "readability") of the sound scene by creating a reverberance (masking) image which seems to originate from a different or distributed direction.

7.2.4 Subjective and electronic design criteria

From a review of the literature relating to "spatial release from masking" (Plomp and Mimpen, 1981; Shinn-Cunningham et al., 2001; Hawley et al., 2004) and the results of the experiment on source and reverberance image interaction, a set of design criteria for the new upmixer were proposed.

The three criteria are here summarized as they relate to both a subjective and electroacoustic evaluation:

1. *Spatial distortion of the source image in the upmixed audio scene should be minimized.*

   This principal aim was to maintain a similar spatial sound character of the source image in both the upmixed and reference 2/0 audio scenes (i.e. a high *fidelity* of the source image). It was proposed that if those sound components that affect only the reverberance imagery could be electronically extracted, then these could be radiated from the loudspeakers behind the listener to create new side reverberance virtual images; leaving the source image spatially undistorted. The efficacy of the upmixer for satisfying this was measured using the graphical mapping system to compare elicited source image geometry in the upmixed and reference 2/0 scenes.

As mentioned, it is the short-term correlated sound components in a pair of audio signal which affect source imagery (i.e. when these signals are radiated with a loudspeaker pair), so if just one of the rear loudspeaker signals is uncorrelated with one of the front loudspeaker signals, then the source image should not be affected. A way of measuring this in electronic terms is to ensure that the four loudspeaker signals are *diagonally uncorrelated.*

2. *Reverberance imagery should have a homogenous distribution in the horizontal plane; in particular, reverberance image directional strength should be high from lateral ($\pm 90°$) directions.*
   The implication of this statement regarding subjective imagery is that in order to create new reverberance images to the side of the listener, the side loudspeaker channels (e.g. R and RS) should have some degree of correlation. In reproduced sound-scenes, lateral-arriving reverberation has been shown to be positively correlated with a sense of listener envelopment (Hiyama et al., 2002), so ideally the reverberance images should be localized in a direction normal to the central loudspeaker axis (i.e. along the interaural axis for a forward-facing listener).
   This explains why the side loudspeaker signals should have a non-zero cross-correlation.

3. *A conventional 2/0 system should not be preferred to the new upmix system.*
   In the context of a home musical listening experience (i.e. listening for *pleasure*- not as part of a critical listening experience)- ideally the new system should be preferred over a reference 2/0 reproduction created using the same recording. The listening tests were actually undertaken

in a laboratory setting, but the listeners were asked to imagine they were listening for pleasure.

7.2.5 Design of the new upmix system

The new upmix system proposed in chapter 5 used a pair of frequency domain adaptive filters, updated according to the Normalized Least Mean Square (NLMS) algorithm (Shynk, 1992; Haykin, 2001). Each input channel was filtered so as to reduce the level of the difference (or error) signal when it was subtracted from the other channel. It was mathematically and empirically shown that the minimization of this error signal in the mean-square sense is equivalent to minimizing the cross-correlation between the error signal and one of the filtered input signals (according to the principle of orthogonality; Haykin, 2001). This satisfies the design criteria that diagonally opposite loudspeaker signals have minimal cross-correlation.

Due to a delay on the unfiltered input to each of the adaptive filters, the system can time-align input signals which have a relative delay of up to ± 10 ms. Such time delays occur when a sound source is recorded with multiple microphones which are at different distances to the source (the time delay could be anything, but delays over 5 ms are rare in natural concert-hall recordings). The system was optimized for real-time operation on a conventional computer (Benesty, 2004); though the processing was done off-line (what this means is that it took less than x seconds to process a two-channel recording of length x seconds, but the input signal was read and written to a file). Due to the overlap-save implementation (Sommen et al., 1987), the filter was updated every 3 ms and could converge to the near-optimum solution within 0.1 seconds from initialization.

A new model for predicting the "rear-to-front" energy ratio of the signals feeding the front and rear loudspeakers in the upmixed sound scene was derived. This ratio is called the *misadjustment* (Widrow and McCool, 1976); and the model was developed in the context of a two-microphone recording of a single source in a concert-hall; based on a stochastic impulse response model (Schroeder and Kuttruff, 1962; Polack, 1993; Jot and Chaigne, 1997; Jacobsen and Roisin, 2000) which assumes a time invariant impulse response and stationary source statistics.

The misadjustment (Ψ) is the energy ratio of the "error" signal ($e_i(n)$) feeding a rear loudspeaker channel to one of the original audio signals ($m_i(n)$) which is radiated from a front loudspeaker on the same side:

$$\Psi_i = \frac{E\left\{e_i^2(n)\right\}}{E\left\{m_i^2(n)\right\}},$$

and can be predicted from the cross-correlation of the two input signals (m_i and m_j):

$$\Psi_i = 1 - \left|\left|\mathbf{c}_{m_j m_i}\right|\right|^2,$$

where

$$\mathbf{c}_{m_j m_i} = \frac{\mathbf{r}_{m_j m_i}}{\sigma_{m_i}\sigma_{m_j}}$$

is the cross-correlation coefficient vector (see e.g. Benesty et al., 2000b) and

$$\mathbf{r}_{m_j m_i} = E\left\{\mathbf{m}_j(n) m_i(n)\right\}$$

is the M-length (un-normalized) cross-correlation vector between the input

signals ($\mathbf{m}_j(n)$ is a block of samples from the input signal up to sample time n, and $m_i(n)$ is a single sample of the other input signal at time n).

The model was further developed to show how system performance depends on the direct-to-reverberant sound energy ratio in the recording environment and the correlation between the early arriving sound components (i.e. the impulse response from source to each mike up until the mixing time, above the Schroeder frequency).

The signal model represented by the above equations was empirically validated using concert-hall recordings of a noise source reproduced with a single loudspeaker and recorded with a variety of microphones configurations and source locations. The model was shown to be robust within a ± 5 dB accuracy; the inaccuracy is accounted for by the fact that the length of the adaptive filter is intentionally less than the actual inter-microphone acoustic impulse response length, plus gradient and environmental noise (Widrow and McCool, 1976; Elko et al., 2002).

7.2.6 Electroacoustic evaluation of system

There were a number of electronic performance measures for the new upmix system: the relative level between the input and output signals (i.e. the misadjustment); the rate of convergence of the adaptive filters (the time taken for the misadjustment to reach a certain level); and the correlation between the output signals. The effects of various algorithm parameters on these measures were evaluated.

The signals used for the analysis were taken from recordings of noise and music reproduced with a loudspeaker in a medium-sized concert-hall

and recorded with a microphone pair. The music recordings were of solo, anechoic performances with a non-electronic instrument. Three commonly-used microphone configurations were used:

1. Spaced pair (AB).
2. Coincident pair (XY).
3. Spaced, angled pair (ORTF).

Analysis of the adaptive filter coefficients in the time-domain show that for a variety of source signals (i.e. different solo musical instrument performances) and source locations, a filter length of about 1000 samples is sufficient to ensure that the part of the impulse response containing the direct sound and low-order reflections (i.e. the non-reverberant part) is modeled by the adaptive filter. This was concluded by looking at kurtosis as a measure for the degree of normality of the filter coefficients, as the reverberant component of an acoustic impulse response can be defined as that part which has a normal (Gaussian) distribution (Schroeder, 1987; Abel and Berners, 2004).

It was found that the introduction of cross-talk between the input signals helped convergence of the adaptive filters for recordings where the direct components of the sound source were only present in one channel; i.e. with recordings containing *hard-(intensity) panned* sources. Without this, direct-sound components would appear in the rear loudspeakers if hard-panning was present, and the source image may be spatially distorted by the upmixing process.

In contrast to the two commercial upmixers tested (Circle surround II and Dolby Pro-Logic II), the new system produced output signals which

were *diagonally uncorrelated* (e.g. signals L and RS); even with coloured signals such as voice. This satisfies the first design criteria (at least, the electronic interpretation of this) and is explained by the principle of orthogonality (Haykin, 2001). Also, side-loudspeaker signals (e.g. L and LS) were found to have a non-zero correlation, which supports the findings from the graphical-description experiment that side reverberance images can be created using the new upmixer. Rear loudspeaker signals LS and RS were negatively correlated, due to the high spatial-correlation for closely-spaced points in a reverberant field (Jacobsen and Roisin, 2000) (i.e. the reverberance image components were highly correlated in the original test recording).

7.2.7 Subjective evaluation of system

Two experiments were conducted to evaluate the new upmix system with regard to the subjective design criteria.

The first experiment used the GUI to describe the perceived spatial envelope of the source and reverberance images in the 2/2 and 2/0 scenes, and the following properties of the elicited images were compared: image width, image distance and image azimuth. 11 experience subjects took part in this test, with stimuli made from recordings of a solo voice and solo viola reproduced with a loudspeaker in a concert hall, and recorded with two different microphone-pair configurations (XY and ORTF). It was found that the stable source image geometry was affected in terms of *azimuth* only for sound sources recorded off-axis, but the source image *width* was not significantly affected by the new upmixer. This result helps to confirm the hypothesis driven by the cross-correlation measurements found in the electronic study that the source image distortion is minimal for the upmixed scene; a finding

in-line with the design criteria.

In a second subjective listening test, 5 experienced and 11 non-experienced critical listeners (audio engineers and musicians) judged various configurations of the new upmixer and a conventional 2/0 audio scene in terms of *preference*. Preference was explained in terms of an overall choice, as if the listener was deciding to purchase the audio systems which produced the sound scenes. It was found that the engineer group consistently preferred the upmixed sound scene with rear signals attenuated by 6 dB, and consistently reported that the 2/0 scene was *less* preferred; contrary to the finding by Rumsey (1999) who found that experienced listeners rarely preferred an upmixed sound scene to the original two-loudspeaker scene.

7.3 Limitations of the new system

7.3.1 Timbral colouration of output signals

As was shown in the theoretical system description in section 4.4 and the empirical validation in section 5.2.2, the output level of the error signals (i.e. the "extracted ambiance" signals) is inversely proportional to the correlation between the input signals. It was also shown using spaced microphone pair measurements in a concert-hall that the correlation is generally much higher at low frequencies (<500-1000 Hz) and can also increase at high frequencies (>4 kHz) (as mentioned; lower frequency reverberation is highly correlated for closely-spaced points in a hall; Jacobsen and Roisin, 2000, and the high frequency increase is related to the reduced high-frequency reverberation due to air and boundary absorption). There is therefore a tendency for the mid

frequencies of the upmixer output signals to be about 10 dB higher in level than low and high frequencies. This is unfortunate (though not perceptually very obvious) because enveloping low frequencies are thought to contribute to an increased sense of immersion in reverberance imagery (i.e. an increased "listener envelopment" or increased "intimacy"; Martens, 1999). However, using the theoretical model developed in the thesis the relative spectral attenuation can be predicted given the signal correlation at a given frequency. Therefore a "spectral re-balancer" could be implemented to account for this whereby the low frequencies of the output signals could be boosted if the input signals were highly correlated at low frequency.[1]

7.3.2 Generalizability to more complicated sound scenes

Sound recordings made with more than two sources are not covered in the mathematical model because the optimum filter condition (which approximates the early part of the inter-microphone acoustic impulse response) would then be time variant. This is especially true when the sound sources are active at different times (such as with music).

To allow faster convergence for a changing optimal solution, multiple filters could be used in parallel; each with different adaptation parameters (e.g. different step-sizes). The filter which performs best according to a particular criteria (e.g. output signal level) is chosen at a given time. This has been used in acoustic echo cancelation systems with two (Ochiai et al., 1977) or more (Usher et al., 2004a) simultaneous filters, and although computation-

[1]The input correlation could be calculated "on the fly" using a running estimate, as suggested in a computationally efficient way by Aarts et al. (2002).

ally more expensive, there can be a dramatic improvement in audio quality for complicated sound scenes (e.g. with multiple, moving sound sources).

7.3.3 Detent of reverberance imagery

As can be seen from a quick inspection of the density plots and image directional strength plots for the evaluation of the new system (i.e. tables 6.2 and 6.3), the reverberance imagery is clearly not evenly distributed between the side loudspeakers- contrary to what was intended according to the subjective design criteria. Besides experimental factors which affected this result (for example; even though the subjects couldn't see the rear loudspeakers, most were familiar with surround-sound reproduction and had an expectation of where the rear loudspeakers were located), this result could be explained by transient components in the rear loudspeaker signals which aided localization of the loudspeakers. This could perhaps be mitigated by adding a delay to the rear loudspeaker channels; though in a later experiment (albeit with different subjects in a different listening environment) it was found that doing so did not significantly affect overall preference (an attenuated rear-output level was, in fact, preferred). To reduce transient localization cues, a temporal envelope smoother may help to reduce this detent affect- which could be implemented using a dynamic compressor with a fast attack time (ideally using a "look-ahead" analysis system- similar to that shown in figure 7.1).

7.3.4 Generalizability of listening test results

The number of subjects who took part in the listening tests for the evaluation of the new upmix system was low; 16 in the preference experiment and 11 in

the descriptive analysis experiment using the graphical mapping technique, with 2 or 3 experimental runs for these two tests. Furthermore, the number of stimuli were small; two different musical instruments, two microphone configurations, and two source locations. This latter factor would have increased the likelihood of *context effects* on the obtained data: how the rating of a stimulus (especially in terms of *preference* over another stimulus) is affected by stimuli which have preceded it.

The degree to which the data can be generalized to other groups of listeners and different stimuli is therefore limited:

> "In the final analysis, the only knowledge that a given scientific study may provide without doubt is the knowledge of what data has been collected in that study. Generalizing beyond these data to what data might be collected in future studies, in other contexts, relies upon acceptance of assumptions and/or models that go beyond the data. And if an experimenter wishes to draw implications for practical applications of the results, then the contextual effects must be addressed most clearly."
> (Martens, 2006)

7.4 Future directions

7.4.1 Look-ahead filtering

Figure 7.1 shows an implementation of the new system which allows for a "look-ahead" in time of the input signal so that the current filter state is

closer to the optimal solution. This should help filter-tracking of the optimal solution (i.e. the filter state which minimized misadjustment). A 5 ms "look-ahead" window is used in the upmix system described by Gundry (2001).

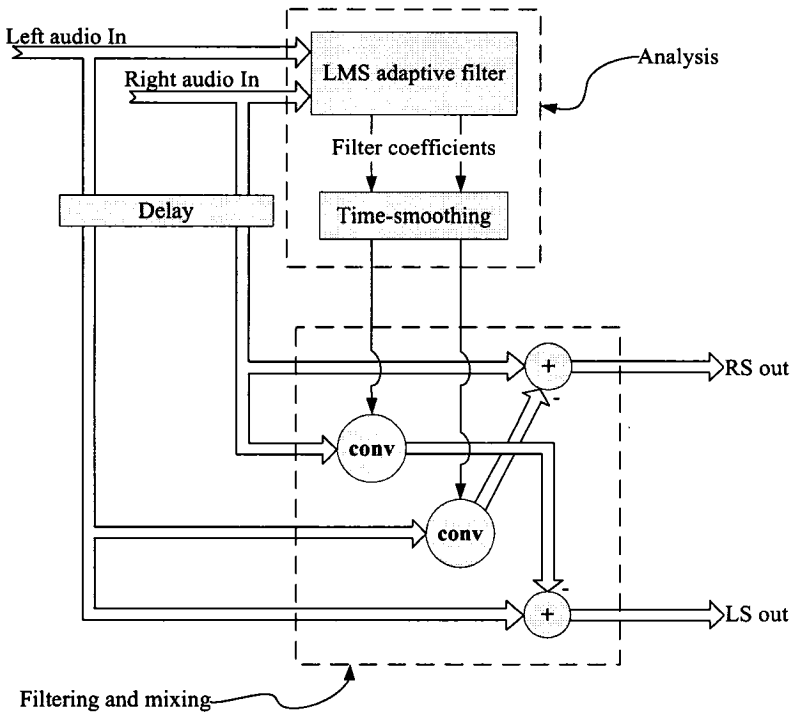


Figure 7.1: Overview of the look-ahead filtering method. The time-lag between the optimum filter conditions and the current state causes increased gradient noise; manifested as an increase in misadjustment (Widrow and McCool, 1976). The analysis and filtering systems are effectively independent (they could be undertaken on separate computers). The **conv** function represents a convolution operation in the time domain (which could be undertaken in the frequency domain, as with the present system). A lag of 5 ms maximum should be sufficient (as used on Dolby Pro-Logic II; Gundry, 2001). Cross-talk can also be added between the input signals, though this is not shown in the figure.

7.4.2 Use of different adaptive filters

To reiterate a comment which was made in section 4.3.1; it should be noted that the adaptive filtering algorithm which was used (i.e. based on the LMS algorithm) was chosen because of its relative mathematical and computational simplicity compared with others (such as the affine projection (Gay, 2000) or RLS (Haykin, 2001) algorithms). So the thesis (i.e. the idea behind the dissertation) is not so much about the particular embodiment presented here with the LMS-based algorithm but rather it should be treated as a general approach to upmixing using adaptive filters to realize the specific design criteria which were introduced.

Alternatively, the adaptive equalization of one channel before subtraction from the other could be accomplished using a frequency-domain transfer function; a method called "empirical estimation" (Carlile, 1996, pg. 157). Such methods are often used in room equalization (e.g. Fielder, 2003), and is very useful when the impulse response is long due to the efficiency of the FFT for large block sizes. However, empirical estimation gives a poor peformance when the noise level is only moderately high whereas methods using the LMS algorithm are very robust to noise (Carlile, 1996, pg. 157).

7.4.3 Adaptive parameters

By using smaller filter lengths the filter can be adapted more frequently at the same computational cost. In other words; convergence to the optimal solution could be faster and tracking of a time variant optimal filter for moving or multiple sources could be enhanced. The filter length was generally much larger than was necessary for the same filter performance (measured in terms

of misadjustment). This was to ensure that even if the musical instrument being recorded was closer to one microphone than the other, the adaptive filter could still time-align the two input signals using a non-minimum phase adaptive filter. Using the lookahead filtering method (figure 7.1), the delay could be estimated by finding the time lag at which the coherence (i.e. absolute cross-correlation) between the two input signals is largest. The filter length could be shortened further by investigating the filter statistics "on the fly", such as looking at the local kurtosis as suggested in section 5.2.2, and truncating the filter to a time when the filter coefficients become normally distributed. This would ensure that the early part of the inter-microphone impulse response would be removed from the rear loudspeaker channels, leaving the source image theoretically undistorted.

Another parameter which could be made dependant on the input signals is the cross-talk gain parameter G. This was the gain applied to one input signal before it was mixed with the other (if the gain was unity, then both input signals would be the same). As was empirically shown in section 5.4.1, if the input signals have a very low cross-correlation, then a cross-talk gain of up to G=-10 dB makes very little difference to the new inter-channel cross-correlation. The cross-talk gain could therefore be made dependant of the input signal correlation; for instance, it could be boosted at frequencies where the correlation was very low to reduce the rear-loudspeaker signal level.

7.4.4 Adaption of the GUI for periphonic and dynamic evaluation of imagery in audio

The reproduction of sound in a way which gives a sense of image *height* is increasingly relevant with spatial audio systems (Woszczyk et al., 2005), yet

the GUI in its present state can not report this. The height dimension could be reported by having a height label for each elicited image. Furthermore, a sense of image motion is only partially measured using the stable or unstable source image descriptor. Arrows showing a movement vector could be used to represent any changes in perceived image direction (e.g. showing where the reverberance image decays to).

7.4.5 Increased input-output channels

It is possible to adapt the upmixer for an arbitrary number of input signals. Such an implementation for five input signals (e.g. from a DVD-video soundtrack or SACD audio recording) is shown in appendix A, where the new signals could be reproduced with loudspeakers located above and below the listener. This would be particularly beneficial for teleconferencing systems for the transmission of multichannel audio signals, where the number of transmitted channels would be less than the number of reproduction channels; reducing transmission bandwidth and the need for expensive, high quality artificial reverberators to enhance the ambiance at the receiving end.

Appendix

Appendix A

ASUS "5 plus 8" arrangement

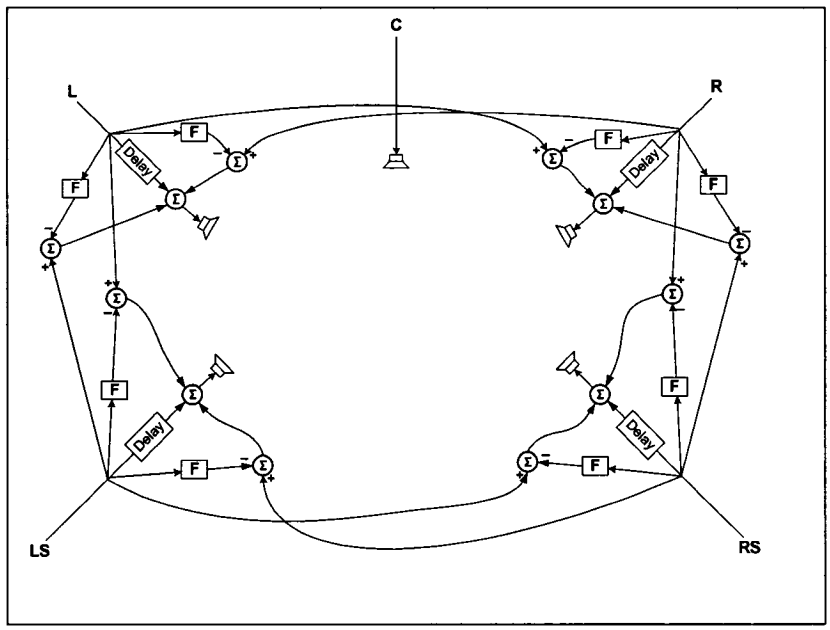


Figure A.1: "5 plus 8" arrangement. This implementation (which was not tested in the dissertation) uses five input audio channels (L, C, R, RS and LS; for example from a musical recording from a DVDA, SACD player or a five-channel output from a film sound-track) which are processed to produce eight new channels. Since the centre channel is not always used for five-channel musical recording mixes, the centre channel is not involved in the upmixing. The box with the "F" in represents the adaptive filter (the feedback path of the error signal is not shown). Rather than mixing the eight new signals to the five loudspeakers, they could be reproduced using other loudspeakers around the listener (such as with loudspeakers above and below the listener).

Appendix B

Recording and reproduction environments.

B.1 MARLAB

B.1.1 Physical details

Summary of MARLAB (adapted from Martin, 2001):

- Approximate volume: 79 m^3 (2.2 m high).
- Heavy velour curtain is hung approximately 10 cm from the walls
- Standard acoustical tiles on ceiling with 10 cm thick uncompressed mineral wool on upper side.
- Tuned membrane absorbers on side walls for low-frequency equalization.
- Separate machine room to house computers.
- Background noise level approximately 27 dB, A-weighted; 40 dB unweighted.

B.1.2 Acoustic analysis

From Martin (2001):
The measurement of the reverberation time of the MARLAB was performed using a DRA Laboratories Maximum Length Systems Sequence Analyzer or MLSSA. The analog output of the MLSSA was connected directly to the RCA input of a Bang & Olufsen Beolab 4000 two-way self-powered loudspeaker. The loudspeaker was placed in the North-East corner of the room, resting on an 11.5 cm high plywood cable trough at the cable entrance to

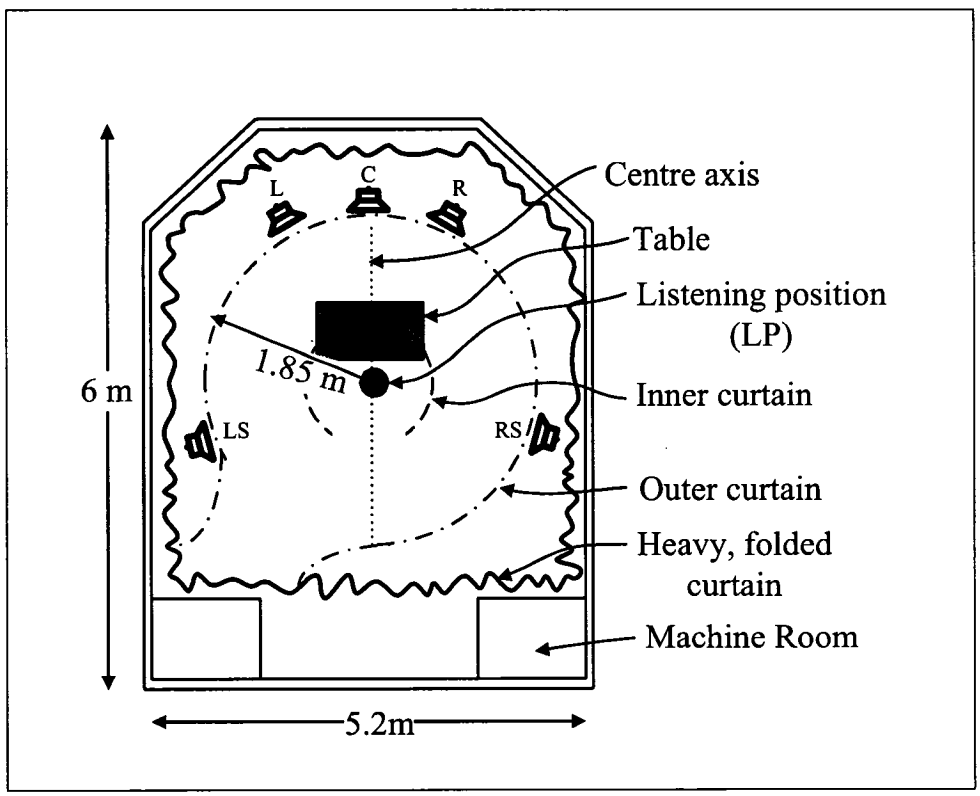


Figure B.1: Plan view of the MARLAB. The ceiling is 2.15 m high. The room is carpeted, with a floating acoustic-tile ceiling with approx. 30 cm of uncompressed mineral wool on top. The locations of the 5 loudspeakers, (L, C, R, RS and LS) are in accordance with ITU-R BS 1116. For the subjective evaluation of the new upmixer, the L and R speakers were at $\pm 30^o$ to the centre-axis and the RS and LS speakers at $\pm 120^o$ (for the new upmixer, the centre channel was not used).

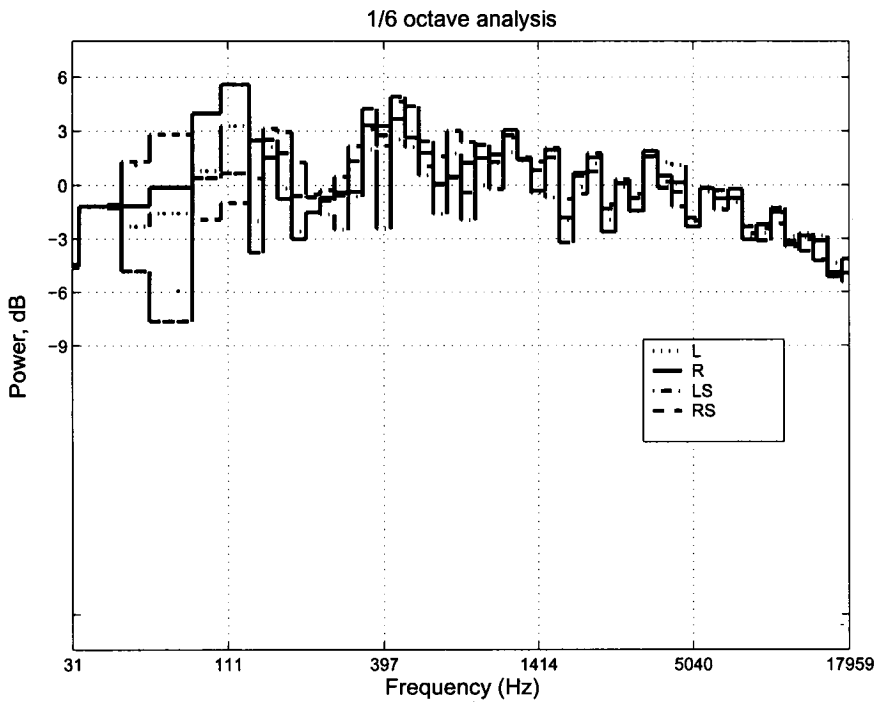


Figure B.2: Loudspeaker-room response of loudspeakers as arranged in figure B.1. Measured using a single B&K type 4191 capsule, 1/2 inch omnidirectional microphone, pointing towards the ceiling (height 1.30m), at the Listening Position (LP). Pink noise reproduced from each loudspeaker at 74 dB (at LP), averaged over 40 seconds. The speakers are 4 Beolab 4000's: 2-way, bass-reflex, 4 litre, active cross-over, built-in AB power amplifier, manufactured and kindly provided by Bang and Olufsen.

the machine room. The microphone was placed near the ceiling panels in the opposite corner of the room, at the junction of the South and South-West walls. The microphone used for this measurement was a Brüel & Kjær 4006 P48 microphone connected to a Tascam DA-P1 portable DAT machine whose output was connected directly to the input of the MLSSA.

Table B.1 shows a partial list of the results of the Calculate Acoustics command in the MLSSA. These data show the results for seven octave bands ranging from 125 Hz to 8 kHz. The Signal to Noise or S/N values are listed in dB to indicate the reliability of the data. The RT-20 values are extrapolated RT60 values (the time it takes the reverberation to decay 60 dB) calculated using the 20 dB window of -5 dB to -25 dB.

	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz
S/N (dB)	59.8	44.7	33.1	37.1	33.1	32.6	32.3
RT-20 (ms)	270	172	177	100	110	111	88

Table B.1: Signal-to-noise ratio and reverberation time measurement for the MARLAB (from Martin, 2001).

B.2 Pollack Hall

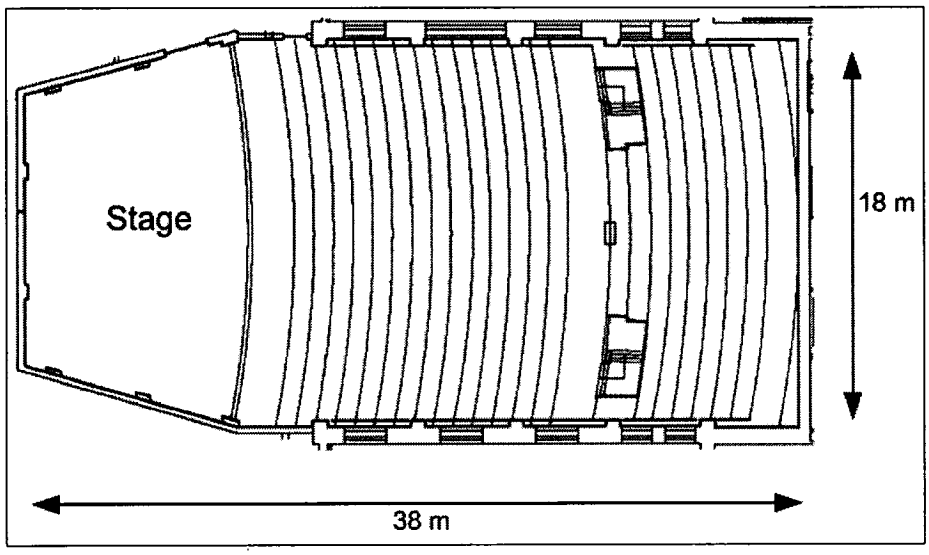


Figure B.3: Pollack Hall (Strathcona music building, McGill University) plan view.

Summary Pollack Hall's acoustical characteristics:

- Approximate volume: 2000 m^3.
- T30: 1.8 s (1 kHz), 2.3 s (63 Hz) (measured using a backward Schroeder integration method with a balloon-burst excitation).
- 600 seat concert hall (20 rows by 30 seats wide).
- Hardwood floor on stage.
- Diffusive wooden walls on stage.
- Sloped seating area raised towards back of hall.
- Diffusive elements on side plastered walls.
- Carpeted floor, upholstered seating.

Appendix C

Software and hardware details

C.1 Computers used in experiments

• The Macintosh computer mentioned in the early subjective studies was a G4 with the OSX operating system (OS).

•PC1:

Dell Poweredge server. Two 1 GHz Pentium III processors, 1.5 GB RAM. Windows 2000 Professional OS.

• PC2:

Dell Inspiron 1100 laptop. Pentium 4, 2.4 GHz processor, 1 GB RAM. Windows XP OS.

C.2 Soundcards

- MARLAB experiments: MOTU 1296.
- Philips experiments: RME Hammerfall.

C.3 Software

- Experiments conducted on the Macintosh: MAX/MSP.
- MATLAB: Same version used on both PC1 and PC2. Version 6.5, release 13. With DSP and statistics toolkits.
- SPSS: Used on PC2. Version 11.0.1.
- Pure Data (PD): Pd version 0.38.3.

C.4 Artificial reverberation generation

This section contains details of the artificial reverb generation used to create the stimuli in the listening tests in chapter 3.

Scene:	#127 "Church piano"
RT	4.0 s
Early Reflection level	-100 dB
Wet mix	100%
Pre delay	12 ms
Modulation width	90%

Table C.1: Configuration of tc electronics M3000 artificial reverberator for generation of single reverb channel used in the experiments.

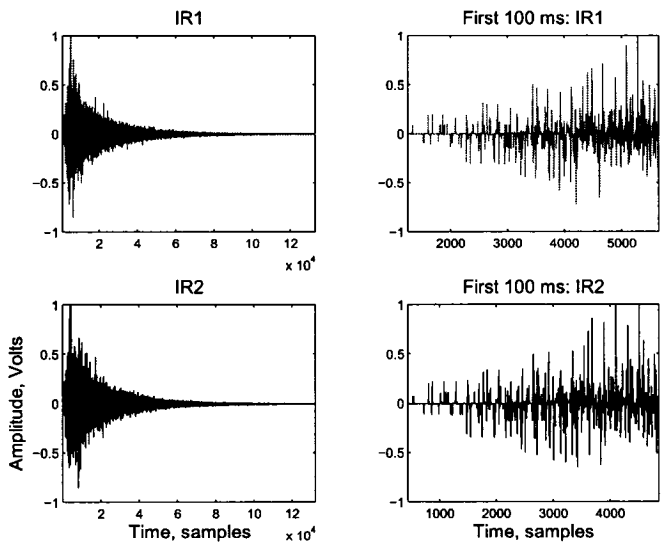


Figure C.1: Consecutive Impulse Responses of tc electronics M3000 artificial reverberator (as used in experiments). Created from a single Dirac function in the digital domain (44.1 kHz fs, 16-bit signal). The x-axis refers to sample time, and y-axis to voltage.

Appendix D

Instructions for participants in listening tests

Dearest Subject,
Please take a moment to read these comments for the listening test- there are a few differences with this test and previous tests.

This experiment is about describing where the sound images in a multi-channel "surround sound" audio experience seem to exist. You describe where you hear both the source images and reverberation images by drawing ellipses using a computer. To keep your head under the sweet-spot, you must keep your head touching the soft foam when you are reporting where you hear the image- you should also be facing forward when you decide where you hear the image (the image direction will change if you move your head only slightly!). You can take a break at any time during the test though.

You will hear a viola and a voice reproduced from loudspeakers behind the curtains around you. On the computer screen, you will see a top-down view

of the listening room, with the curtains and other numbered markers indicated. Use the markers on the curtain and the GUI to help match where you hear the sound images to the GUI. You should always use the laser pointer to decide where you hear the sound- sometimes this is best done by closing your eyes and just pointing, then looking at the markers to map where you hear the images. **It is the direction, width and distance of the image that you should describe- using the laser pointer to decide the direction and width.** Remember, you are describing where you hear the phantom sound image, not where you think the loudspeakers are! Therefore, if you think the image sounds like it is coming from in front of the curtains when you close your eyes, then draw the image there on the GUI. You may use as many ellipses as you wish to describe the image shape.

The images are described in two ways: either as a **source image** or a **reverberance image**. A source image is a sound image which seems to be where the (phantom) source exists. In this experiment, this means where the viola or singer seems to exist in the sound scene. You can describe the source image as being either stable or unstable. Unstable applies if the source image seems jumpy or fuzzy in a certain region of space. A reverberance image is a sound image which sounds like live-reverberation (reverberance is the perceptual equivalent of acoustical reverberation). If you listen to the output of an artificial reverberator, set to "100% wet", then you will hear a reverberance sound image. You can use as many ellipses as you wish to describe where you hear the source and reverberance sound images.

To reiterate, there are three things you have to describe:

1. The image direction.
2. The image width.
3. The image distance.

The laser pointer should be used for the first two. When describing the image

direction (i.e. the centre of the image), you should do this when facing forward and your head touching the foam. The sound is controlled by another program (PD) so when you have finished describing the images with the GUI you must hit the "next" button on the other program- I'll show you how!
Cheers, and thanks alot for giving your time and effort -jHon ☺

References

Aarts, R. M. (1995). System for deriving a center channel signal from an adapted weighted combination of the left and right channels in a stereophonic audio signal. US patent #5,426,702.

Aarts, R. M., Irwan, R., and Janssen., A. J. E. M. (2002). Efficient tracking of the cross-correlation coefficient. *IEEE Trans. on Speech and Audio Processing*, 10(6):391–402.

Abel, J. S. and Berners, D. P. (2004). Reverberation acoustics, analysis and synthesis. *Tutorial session T21, 117th Convention of the AES, San Francisco.*

Ando, Y. (1985). *Concert hall acoustics.* Springer-Verlag, Berlin ; New York.

Atal, B. and Schroeder, M. (1962). Apparent sound source translator. US Patent #3,236,949.

Avendano, C. and Jot, J.-M. (2002). Ambience extraction and synthesis from stereo signals for multichannel audio upmix. In *Proceedings IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Orlando, Florida.

Avendano, C. and Jot, J.-M. (2004). A frequency-domain approach to mul-

tichannel upmix. *Journal of the Audio Engineering Society*, 52(7/8):740–749.

Baird, J. C. (2001). *Sensation and judgement: complementarity theory of psychoacoustics*. Erlbaum, Mahwah, NJ.

Barron, M. (1971). The subjective effect of first reflections in concert halls - The need for lateral reflections. *Journal of Sound and Vibration*, 15:475–494.

Barron, M. and Marshall, A. H. (1981). Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure. *Journal of Sound and Vibration*, 77:211–232.

Bech, S. (1992). Selection and training of subjects for listening tests on sound-reproducing equipment. *Journal of the Audio Engineering Society*, 40:590–610.

Bech, S. (1998). Spatial aspects of reproduced sound in small rooms. *Journal of the Acoustical Society of America*, 103(1):434–445.

Bech, S. (1999). Methods for subjective evaluation of spatial characteristics of sound. In *Proceedings of the AES 16th international conference on spatial sound reproduction*, Rovaniemi, Finland.

Bech, S. (2001). Methods for the identification of primary subjective attributes of spatial sound quality. In *Proceedings of the International Workshop on Spatial Media*, Aizu-Wakamatsu, Japan.

Begault, D. R. (1992). Perceptual effects of synthetic reverberation on three-dimensional audio systems. *Journal of the Audio Engineering Society*, 40(11):895–904.

Begault, D. R. and Erbe, T. (1994). Multichannel spatial auditory display for speech communications. *Journal of the Audio Engineering Society*, 42:819–826.

Begault, D. R. and McClain, B. U. (2001). Early reflection thresholds for virtual sound sources. In *Proceedings of the International Workshop on Spatial Media*, Aizu-Wakamatsu, Japan.

Benesty, J. (2001). General derivation of frequency-domain adaptive filtering. In Benesty, J., Gänsler, T., Morgan, D. R., Sondhi, M. M., and Gay, S. L., editors, *Advances in Network and Acoustic Echo Cancellation*. Springer.

Benesty, J. (2004). Personal communication.

Benesty, J. and Duhamel, P. (1992). A fast exact least mean square adaptive algorithm. *IEEE Trans. on Signal Processing*, 40:2904–2920.

Benesty, J., Gänsler, T., and Eneroth, P. (2000a). Multi-channel sound, acoustic echo cancellation, and multi-channel time-domain adaptive filtering. In Gay, S. L. and Benesty, J., editors, *Acoustic Signal Processing for Telecommunication*, chapter 6, pages 101–120. Kluwer Academic Publishers.

Benesty, J., Morgan, D. R., and Cho, J. H. (2000b). A new class of doubletalk detectors based on cross-correlation. *IEEE Trans. Speech Audio Processing*, 8(2):168–172.

Beranek, L. L. (1996). *Concert and Opera Halls: How They Sound*. Acoustical Society of America through the American Institute of Physics, Woodbury.

Berg, J. and Rumsey, F. (1999). Spatial attribute identification and scaling

by repertory grid technique and other methods. In *Proceedings of the AES 116th international convention*, Rovaniemi, Finland.

Berg, J. and Rumsey, F. (2000). Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction. In *Proceedings of the AES 109th international convention*, Los Angeles.

Berg, J. and Rumsey, F. (2002). Validity of selected spatial attributes in the evaluation of 5-channel microphone techniques. In *Proceedings of the AES 112th international convention*, Munich, Germany.

Berkhout, A. (1988). A holographic approach to acoustic control. *Journal of the Audio Engineering Society*, 36(12):977–995.

Bernfeld, B. (1973). Attempts for better understanding of the directional sterephonic listening mechanism. In *Proceedings of the AES 44th international convention*, Rotterdam, Netherlands.

Blauert, J. (1997). *Spatial hearing: The psychophysics of human sound localization.* MIT Press, Cambridge, Mass.

Blauert, J. and Lindemann, W. (1986). Spatial mapping of intracranial auditory events for various degrees of interaural coherence. *Journal of the Audio Engineering Society*, 79(3):806–813.

Blech, D. and Yang, M.-C. (2004). DVD-Audio versus SACD: Perceptual discrimination of digital audio coding formats. In *Proceedings of the AES 116th international convention*, Berlin, Germany.

Blesser, B. (2001). Interdisciplinary synthesis of reverberation viewpoint. *Journal of the Audio Engineering Society*, 49(10):867–903.

Boone, M., de Bruijn, W. P. J., and Horbach, U. (1999). Virtual surround speakers with wave field synthesis. In *Proceedings of the AES 106th international convention*, Munich, Germany.

Braasch, J., Martens, W. L., and Woszczyk, W. (2004). Modeling auditory localization of subwoofer signals in multi-channel loudspeaker arrays. In *Proceedings of Audio Engineering Society 117th convention*, San Francisco, USA.

Bradley, J. S. and Soulodre, G. A. (1995). Objective measures of listener envelopment. *Journal of the Acoustical Society of America*, 98(2):2590–2597.

Bregman, A. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press, Cambridge, Mass.

Brungart, D. S. (1993). Distance simulation in virtual audio displays. In *Proceedings of the IEEE National Aerospace and Electronics Conference*, pages 612–617.

Carlile, S. (1996). *Virtual auditory space: Generation and applications.* Landes, Austin.

Caussé, R., Bresciani, J., and Warusfel, O. (1992). Radiation of musical instruments and control of reproduction with loudspeakers. In *Proceedings of the International Symposium of Music Acoustics*, Tokyo.

Chernyak, R. I. and Dubrovsky, N. A. (1968). Pattern of the noise images and the binaural summation of loudness for the different interaural correlation of noise. In *Proceedings of the 6th International Congress on Acoustics*, pages A53–A56.

Cherry, C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, 25:554–559.

Choi, Y., Han, S., Lee, D., and Sung, K. (1995). A new digital surround processing system for general A/V sources. *IEEE Transactions on Consumer Electronics*, 41(4):1174–1180.

Choisel, S. and Zimmer, K. (2003). A pointing-technique with visual feedback for sound-source localization experiments. In *Proceedings of the 115th Convention of the Audio Engineering Society*, New York.

Clark, G., Parker, S., and Mitra, S. (1983). A unified approach to time and frequency-domain realization of FIR adaptive digital filters. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-31(5):1073–1083.

Corey, J. and Woszczyk, W. (2002). Localization of lateral phantom images in a 5-channel system with and without simulated early reflections. In *Proceedings of the AES 113th international convention*, Los Angeles.

Corey, J. A. (2002). *An integrated system for dynamic control of auditory perspective in a multichannel sound field.* PhD thesis, Department of sound recording, McGill University.

Cremer, L. (1976). On the use of the terms "degree of correlation" and "degree of coherence". *Acustica*, 35:215–218.

Cremer, L. and Müller, H. A. (1982). *Principles and applications of room acoustics, volume 2.* Applied science, London and New York.

Culling, J. F., Colburn, H. S., and Spurchise, M. (2001). Interaural correlation sensitivity. *Journal of the Acoustical Society of America*, 110:1020–1029.

Damaske, P. (1967). Subjektive untersuchung von schallfeldern (subjective investigation of sound fields). *Acustica*, 19:199–213.

Damaske, P. and Ando, Y. (1973). Interaural crosscorrelation for multichannel loudspeaker reproduction. *Acustica*, 27:232–238.

Darlington, R. B. (1970). Is kurtosis really peakedness? *The American Statistician*, 24:19–22.

de Bruijn, W. (2004). *Application of wave field synthesis in videoconferencing.* PhD thesis, TU Delft.

de Vries, D., Hulsebos, E. M., and Baan, J. (2001). Spatial fluctuations in measures for spaciousness. *Journal of the Acoustical Society of America*, 110:947–954.

Dentino, M., Widrow, B., and McCool, J. (1978). Adaptive filtering in the frequency domain. In *Proceedings of the IEEE*, volume 66, pages 1658–1659.

Dolby (2005). Surround sound: Past, present, and future. Online article: `http://www.dolby.com/consumer/motion_picture/dolby_in_pictures3.html`.

Dressler, R. (2000). Dolby Surround Pro Logic II Decoder. Principles of operation. Dolby Laboratories Information.

Dunn, C. and Hawksford, M. (1993). Distortion immunity of MLS-derived impulse response measurements. *Journal of the Audio Engineering Society*, 41(5):314–335.

Ebata, M. (2003). Spatial unmasking and attention related to the cocktail party problem. *Acoustical Science and Technology*, 24:208–219.

Elko, G. W., Diethorn, E., and Gaensler, T. (2002). Room impulse response variation due to temperature fluctuations and its impact on acoustic echo cancellation. Technical report.

Evans, M. J. (1998). Obtaining accurate responses in directional listening tests. In *Proceedings of the AES 104th international convention*, Amsterdam, The Netherlands.

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

Farina, A. and Ugolotti, E. (1999). Subjective comparison between stereo dipole and 3D ambisonic surround systems for automotive applications. In *Proceedings of the AES 16th international conference on spatial sound reproduction.*

Fellgett, P. B. (1974). Ambisonic reproduction of directionality in surround sound systems. *Nature, London*, 252:534–538.

Fielder, L. D. (2003). Analysis of traditional and reverberation-reducing methods of room equalization. *Journal of the Audio Engineering Society*, 51(1/2):3–26.

Fisher, S. S. (1990). *Virtual Interface Environments.* The Art of Human-Computer Interface Design, New York: Addison-Wesley.

Ford, N. (2005). *Developing a Graphical Language to Represent Listeners' Experiences of Spatial Attributes in Reproduced Sound.* PhD thesis, Univerity of Surrey, England. School of Performing Arts.

Ford, N., Rumsey, F., and de Bruyn, B. (2001). Graphical elicitation techniques for subjective assessment of the spatial attributes for loudspeaker

reproduction: A pilot investigation. In *Proceedings of the AES 110th international convention*, Amsterdam, The Netherlands.

Ford, N., Rumsey, F., and Nind, T. (2002). Subjective evaluation of perceived spatial differences in car audio systems using a graphical assessment language. In *Proceedings of the AES 112th international convention*, Munich, Germany.

Fosgate, J. W. (1993). Surround sound loudspeakers and processor. US patent #5,199,075.

Fosgate, J. W. (2005). Method for deriving at least three audio signals from two input audio signals. US patent #6,920,223.

Frigo, M. and Johnson, S. (2005). The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231.

Fukada, A., Tsujimoto, K., and Akita, S. (1997). Microphone techniques for ambient sound on a music recording. In *Proceedings of the AES 103rd international convention*, New York.

Gabrielsson, R., Rosenburg, U., and Sjogren, H. (1974). Judgements and dimension analysis of perceived sound quality of sound-reproducing systems. *Journal of the Acoustical Society of America*, 55:854–861.

Gautier, P.-A., Berry, A., and Woszczyk, W. (2004). An introduction to the foundations, the technologies and the potential applications of the acoustic field synthesis for audio spatialization of loudspeaker arrrays. In *Proceedings of the Harvest Moon symposium on multichannel sound*, Montreal, Canada.

Gautier, P.-A., Berry, A., and Woszczyk, W. (2005). Sound-field reproduction

in-room using optimal control techniques: Simulations in the frequency domain. *Journal of the Acoustical Society of America*, 117(2):662–678.

Gay, S. (2000). The fast affine projection algorithm. In Gay, S. and Benesty, J., editors, *Acoustic Signal Processing for Telecommunication*, chapter 2. Kluwer Academic Publishers, Boston.

Gerzon, M. A. (1977). Design of ambisonic decoders for multi speaker surround sound. In *Proceedings of the AES 58th international convention.*

Gerzon, M. A. (1992a). Optimum reproduction matrices for multispeaker stereo. *Journal of the Audio Engineering Society*, 40(7/8):571–589.

Gerzon, M. A. (1992b). Panpot laws for multispeaker stereo. In *Proceedings of the AES 92nd international convention*, Vienna, Austria.

Gescheider, G. A. (1997). *Psychophysics, the Fundamentals.* Erlbaum, Hillsdale, NJ.

Griesinger, D. (1989). A directionality enhancement system for converting encoded stereo signals into four output channels. US patent #4,862,502.

Griesinger, D. (1996a). Multichannel matrix surround decoders for two-eared listeners. In *Proceedings of the AES 101st international convention.*

Griesinger, D. (1996b). Spaciousness and envelopment in musical acoustics. In *Proceedings of the AES 101st international convention*, Los Angeles.

Griesinger, D. (1997). The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces. *Acustica*, 83(4):721–731.

Griesinger, D. (1998). Multichannel active matrix sound reproduction with maximum lateral seperation. US patent #5,796,844.

Griffiths, T. and Warren, J. (2004). What is an auditory object? *Nature Reviews Neuroscience*, 5:887–892.

Guastavino, C. and Katz, B. (2004). Perceptual evaluation of multi-dimensional spatial audio reproduction. *Journal of the Acoustical Society of America*, 116(2):1105–1115.

Gundry, K. (2001). A new active matrix decoder for surround sound. In *Proceedings of the AES 19th international conference*, Schloss Elmau, Germany.

Hafler, D. (1972). Passive surround sound circuit. US patent #3,697,692.

Hamasaki, K., Hiyama, K., Nishiguchi, T., and Ono, K. (2004). Advanced multichannel audio systems with superior impression of presence and reality. In *Proceedings of the AES 116th international convention*, Berlin, Germany.

Hänsler, E. and Schmidt, G. (2003). Control of LMS-type adaptive filters. In Haykin, S. and Widrow, B., editors, *Least-Mean-Square Adaptive Filters*. Wiley.

Hänsler, E. and Schmidt, G. (2005). Single-channel acoustic echo cancellation. In Benesty, J. and Huang, Y., editors, *Adaptive signal processing*, chapter 3, pages 59–93. Springer, New York.

Hanyu, T. and Sekiguchi, K. (2004). Relationship between sound image and listener envelopment of sound fields in concert halls. In *Proceedings of the 18th International Congress on Acoustics*, Kyoto, Japan.

Harrison, R. W. (1995). Passive surround sound circuit. US patent #5,386,473.

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of America*, 115(2):833–843.

Haykin, S. (2001). *Adaptive Filter Theory.* Prentice Hall, Englewood Cliffs, N. J., 4th edition.

Haykin, S. and Widrow, B. (2003). *Least-Mean-Square Adaptive Filters.* Wiley, New-York.

Hertz, B. F. (1981). 100 years with stereo-the beginning. *Journal of the Audio Engineering Society*, 29(5):368–370.

Hiyama, K., Komiyama, S., and Hamasaki, K. (2002). The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field. In *Proceedings of the AES 113th international convention*, Los Angeles.

Holman, T. (1991a). New factors in sound for cinema and television. *Journal of the Audio Engineering Society*, 39:529–539.

Holman, T. (1991b). Sound system with source material and surround timbre response correction, specified front and surround loudspeaker directionality, and multi-loudspeaker surround. US Patent #5,043,970.

Holman, T. (2000a). *5.1 Surround Sound: Up and Running.* Focal Press, Boston, MA.

Holman, T. (2000b). Comments on "subjective appraisal of loudspeaker directivity for multichannel reproduction". *Journal of the Audio Engineering Society*, 48(4):314–321.

Houtsma, A., Rossing, T., and Wagenaars, W. (1987). Auditory demonstrations CD [track 35]. IPO and ASA.

IEC 268-13 (1985). Sound system equipment–part 13: Listening tests on loudspeakers. Technical report, International Electrotechnical Commission, Geneva, Switzerland.

Irwan, R. and Aarts, R. (2002a). Multi-channel stereo converter for deriving a stereo surround amd/or audio center signal. US Patent #6,496,584.

Irwan, R. and Aarts, R. M. (2002b). Two-to-five channel sound processing. *Journal of the Audio Engineering Society*, 50(11):914–926.

ITU-R BS 1116 (1994). Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Recommendation BS 1116, International Telecommunication Union Radiocommunication Assembly.

ITU-R BS 775-1 (1994). Multichannel stereophonic sound system with and without accompanying picture. Recommendation BS.775-1, International Telecommunication Union Radiocommunication Assembly.

Jacobsen, F. and Roisin, T. (2000). The coherence of reverberant sound fields. *Journal of the Acoustical Society of America*, 108(1):204–210.

Jot, J.-M. and Chaigne, A. (1997). Analysis and synthesis of room reverberation based on a statistical time-frequency model. In *Proceedings of the AES 103rd international convention*, New York.

Katz, B. (2002). Process for enhancing the existing ambiance, imaging, depth, clarity and spaciousness of sound recordings. US patent application # 09/877,158.

Keet, W. d. V. (1968). The influence of early lateral reflections on the spatial impression. In *Proceedings of the 6th International Congress on Acoustics*, pages E–2–4, Tokyo.

Kendall, G. (1995). The decorrelation of audio signals and its impact on spatial imagery. *Computer Music Journal*, 19(4):72–87.

Kim, S., Martens, W. L., and Marui, A. (2005). Discrimination of auditory source focus for musical instrument sounds with varying low-frequency cross correlation in multichannel loudspeaker reproduction. In *Proceedings of the AES 119th international convention*, New York.

Kinsler, L., Frey, A. R., Coppens, A. B., and Sanders, J. V. (1999). *Fundamentals of acoustics*. Wiley, New York, 4th edition.

Klipsch, P. W. (1958). Stereophonic sound with two tracks, three channels by means of a phantom circuit. *Journal of the Audio Engineering Society*, 6(2):118–123.

Kubovy, M. and Valkenburg, D. V. (2001). Auditory and visual objects. *Cognition*, 80:97–126.

Kurozumi, K. and Ohgushi, K. (1983). The relationship between the cross-correlation coeffiecient of two-channel acoustical signals and sound image quality. *Journal of the Acoustical Society of America*, 74:1726–1733.

Kutruff, H. (1991). *Room Acoustics*. Elsevier Science Publishers, Essex, 3rd edition.

Lauridsen, H. (1954). Experiments concerning different kinds of room-acoustics recording. *Ingenioren*, 47.

Lee, H.-K. and Rumsey, F. (2005). Investigations on the effect of interchannel crosstalk in three channel microphone techniques. In *Proceedings of the AES 118th international convention*, Barcelona, Spain.

Lee, J. and Un, C. K. (1989). Performance analysis of frequency-domain block LMS adaptive digital filters. *IEEE Trans. on Circuits and Systems*, 36:173–189.

Letowsky, T. (1989). Sound quality assessment: concepts and criteria. In *Proceedings of the AES 87th international convention*, New York.

Lund, T. (2000). Enhanced localization in 5.1 production. In *Proceedings of the AES 109th international convention*, Los Angeles.

Mader, A., Puder, H., and Schmidt, G. U. (2000). Step-size control for acoustic echo cancellation filters – an overview. *Signal Processing*, 80:1697–1719.

Madsen, E. R. (1970). Extraction of ambiance information from ordinary recordings. *Journal of the Audio Engineering Society*, 18(5):490–496.

Maher, R. C., Lindemann, E., and Barish, J. (1996). Old and new techniques for artificial stereophonic image enhancement. In *Proceedings of the AES 101st international convention*, Los Angeles.

Malham, D. (1999). Homogeneous and nonhomogeneous surround sound systems. In *Proceedings of the AES UK conference: Second Century of Audio*, London.

Mansour, D. and Gray, A. H. (1982). Unconstrained frequency- domain adaptive filter. *IEEE Trans. Acoustics, Speech and Signal Processing*, 30:726–734.

Marin, O. S. M. and Perry, D. W. (1999). Neurological aspects of music perception and performance. In Deutsch, D., editor, *The psychology of music*, chapter 17. Academic Press.

Marr, D. and Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society*, 207:187–217.

Marshal, A. and Barron, M. (2001). Spatial responsiveness in concert halls and the origins of spatial impression. *Applied Acoustics*, 62:91–108.

Martens, W. L. (1999). The impact of decorrelated low-frequency reproduction on auditory spatial imagery: Are two subwoofers better than one? In *Proceedings of the AES 16th international conference on spatial sound reproduction*, pages 87–77, Rovaniemi, Finland.

Martens, W. L. (2001). Two-subwoofer reproduction enables increased variation in auditory spatial imagery. In *Proceedings of the International Workshop on Spatial Media*, Aizu-Wakamatsu, Japan.

Martens, W. L. (2004). Decoupled loudness and range control for a source located within a small virtual acoustic environment. In *Proceedings of the 2004 International Conference on Auditory Display*, Sydney, Australia.

Martens, W. L. (2006). Contextual effects in sensory evaluation of spatial audio: Integral factor or nuisance? In *Proceedings of Spatial audio and sensory evaluation techniques conference*, Guilford, UK.

Martens, W. L. and Woszczyk, W. (2003). Guidelines for enhancing the sense of presence in virtual acoustic environments. In *Proceedings of the 9th International Conference on Virtual Systems and Multimedia*, pages 306–313, Montreal, Canada.

Martin, G. (2001). *A Hybrid Model for Simulating Diffused First Reflections in Two-dimensional Synthetic Acoustic Environments.* PhD thesis, Department of sound recording, McGill University.

Martin, G., Woszczyk, W., Corey, J., and Quesnel, R. (1999). Sound source localization in a five-channel surround sound reproduction system. In *Proceedings of the AES 107th international convention*, New York.

Mason, R. (2002). *Elicitation and measurement of auditory spatial attributes in reproduced sound.* PhD thesis, Univerity of Surrey, England. School of Performing Arts.

Mason, R., Brookes, T., and Rumsey, F. (2004). Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli. *Journal of the Acoustical Society of America*, 117(3):1337–1350.

Mason, R., Brookes, T., and Rumsey, F. (2005). The effect of various source signal properties on measurements of the interaural crosscorrelation coeffcient. *Acoustical Science and Technology*, 26(2):102–113.

Mason, R., Ford, N., Rumsey, F., and de Bruyn, B. (2001). Verbal and nonverbal elicitation techniques in the subjective assessment of spatial sound reproduction. *Journal of the Audio Engineering Society*, 49(5):366–384.

Mason, R. and Rumsey, F. (1999). An investigation of microphone techniques for ambient sound in surround sound systems. In *Proceedings of the AES 106th international convention*, Munich, Germany.

McAdams, S. (1983). Spectral fusion and the creation of auditory images. In Clynes, M., editor, *Music, Mind, and Brain: The Neuropsychology of Music*, pages 279–298. New York: Plenum Press.

McGrath, R., Waldmann, T., and Fernström, M. (1999). Listening to rooms and objects. In *Proceedings of the AES 16th international conference on spatial sound reproduction*, Rovaniemi, Finland.

Merimaa, J. and Hess, W. (2004). Training of listeners for evaluation of spatial attributes of sound. In *Proceedings of the AES 117th international convention*, San Francisco.

Merimaa, J. and Pulkki, V. (2005). Spatial impulse response rendering I: Analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115–1127.

Mershon, D. H. (1997). Phenomenal geometry and the measurement of perceived auditory distance. In Gilkey, R. and Anderson, T. R., editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter 13, pages 257–274. Erlbaum, Mahwah, NJ.

Meyer, E. and Schodder, G. R. (1952). Über den Einfluss von Schallrückwürfen auf Richtungslohalisation und Lautastärke bei Sprache. *Nach. Akad. Wiss. Gottingen, Math. Phys. Klasse IIa*, 6:31–42.

Middlebrooks, J. C. (1992). Narrow-band sound localization related to external ear acoustics. *Journal of the Acoustical Society of America*, 92(5):2607–2624.

Miles, M. T. (1996). An optimum linear-matrix stereo imaging system. In *Proceedings of the AES 101st international convention*, Los Angeles.

Møller, H., Sørenson, M. F., Jenson, C. B., and Hammershøi, D. (1996). Binaural technique: Do we need individual recordings? *Journal of the Audio Engineering Society*, 44(6):451–469.

Moore, B. C. J. (1997). *An introduction to the psychology of hearing.* Academic Press, San Diego, Calif., 4th edition.

Morimoto, M. (1993). Relationship between auditory source width in various

sound fields and degree of interaural cross-correlation. *Applied Acoustics*, 38:291–301.

Morimoto, M. (2001). How can auditory spatial impression be generated and controlled? In *Proceedings of the International Workshop on Spatial Media*, Aizu-Wakamatsu, Japan.

Morimoto, M. (2002). The relation between spatial impression and the precedence effect. In *Proceedings of the International Conference on Auditory Display.*

Morimoto, M. and Asaoka, A. (2004). Multi-dimensional analysis of "reverberance". In *Proceedings of the 18th International Congress on Acoustics*, Kyoto, Japan.

Morse, P. M. and Ingard, K. U. (1968). *Theoretical Acoustics.* Princeton University Press.

Neely, S. T. and Allen, J. B. (1979). Invertibility of room impulse response. *Journal of the Acoustical Society of America*, 66:165–169.

Neher, T. (2004). *Towards A Spatial Ear Trainer.* PhD thesis, Univerity of Surrey, England. School of Performing Arts.

Neher, T., Rumsey, F., and Brookes, T. (2002). Training of listeners for the evaluation of spatial sound reproduction. In *Proceedings of the AES 112th international convention*, Munich, Germany.

Nelson, P. A., Kirkeby, O., Takeuchi, T., and Hamada, H. (1997). Sound fields for the production of virtual acoustic images. *Journal of Sound and Vibration*, 204(2):386–396.

Nelson, R. (1995). *Probability, Stochastic Processes, and Queueing Theory.* Springer-Verlag.

Nicol, R. and Emerit, E. (1998). Reproducing 3D sound for videoconferencing: a comparison between holophony and ambisonic. In *Proceedings of the Digital Audio Effects Workshop*, Barcelona, Spain.

Nielsen, S. H. (1993). Auditory distance perception in different rooms. *Journal of the Audio Engineering Society*, 41:755–770.

Nunally, J. and Bernstein, I. (1994). *Psychometric Theory.* McGraw-Hill, New York, 3rd edition.

Ochiai, K., Araseki, T., and Ogihara, T. (1977). Echo canceller with two echo path models. *IEEE Trans. Communications*, 25:589–595.

Ohgushi, K., Komiyama, S., Kurozumi, K., Morita, A., and Ujihara, J. (1987). Subjective evaluation of multichannel stereophony for HDTV. *IEEE Transactions on Broadcasting*, 33:197–202.

Okano, T. (2000). Image shift caused by strong lateral reflections, and its relation to inter-aural cross correlation. *Journal of the Acoustical Society of America*, 108(5):2219–2230.

Oldfield, S. R. and Parker, S. P. A. (1984). Acuity of sound localization: A topography of auditory space. I. Normal hearing conditions. *Perception*, 13:581–600.

Olive, S. E. (2003). Differences in performance and preference of trained versus untrained listeners in loudspeaker tests: a case study. *Journal of the Audio Engineering Society*, 51(9):806–825.

Olive, S. E. and Toole, F. E. (1989). The detection of reflections in typical rooms. *Journal of the Audio Engineering Society*, 37:539–553.

Oppenheim, A. V. and Shafer, R. W. (1999). *Discrete-Time Signal Processing.* Prentice-HalL.

Papoulis, A. and Pillai, S. (2002). *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, 4th edition.

Pelkowitz, L. (1981). Frequency domain analysis of wrap-around error in fast convolution algorithms. *IEEE Trans. Acoustics, Speech and Signal Processing*, 29:413–422.

Pelorson, X., Vian, J.-P., and Polack, J.-D. (1992). On the variability of room acoustical parameters: Reproducibility and statistical validity. *Applied Acoustics*, 37:175–198.

Plomp, R. (1976). Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound wave speech or noise. *Acustica*, 31:200–211.

Plomp, R. and Mimpen, A. M. (1981). Effect of the orientation of the speaker's head and the azimuth of a noise source on the speech-reception threshold for sentences. *Acustica*, 48:325–328.

Pohlmann, K. C. (2000). *Principles of digital audio.* McGraw-Hill, New York, forth edition.

Polack, J.-D. (1993). Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics. *Applied Acoustics*, 38:235–344.

Proakis, J. and Manolakis, D. (1996). *Digital signal processing: principles, algorithms and applications.* Macmillan, 3rd edition.

Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45:456–466.

Pulkki, V. (1999a). Coloration of amplitude-panned virtual sources. In *Proceedings of the AES 110th international convention*, Amsterdam. The Netherlands.

Pulkki, V. (1999b). Uniform spreading of amplitude panned virtual sources. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York.

Pulkki, V. and Hirvonen, T. (2005). Localization of virtual sources in multichannel audio reproduction. *IEEE Transactions on Speech and Audio Processing*, 13:105–119.

Pulkki, V. and Karjalainen, M. (2001). Localization of amplitude-panned virtual sources I: Stereophonic panning. *Journal of the Audio Engineering Society*, 49:739–752.

Queen, D. (1979). The effect of loudspeaker radiation patterns on stereo imaging and clarity. *Journal of the Audio Engineering Society*, 27:368–379.

Rasch, R. A. and Plomp, R. (1982). The listener and the acoustic environment. In Deutsch, D., editor, *The Psychology of Music*, pages 135–147. Academic Press, first edition.

Ratliff, P. A. (1974). Properties of hearing related to quadraphonic reproduction. Technical report, BBC Research Department.

Reichardt, W. and Lehmann, U. (1978). Raumeidruck als Oberbegriff von räumlichkeit und Halligkeit, Erläuterungen des Raumeindrucksmasses. *Acustica*, 40:277–290.

RIAA (2004). 2004 year end statistics by the Recording Industry Association of America.

Rosen, G. L. and Johnston, J. D. (2001). Automatic speaker directivity control for soundfield reconstruction. In *Proceedings of the AES 19th international conference*, Schloss Elmau, Germany.

Rumsey, F. (1999). Controlled subjective assessments of two-to-five-channel surround sound processing algorithms. *Journal of the Audio Engineering Society*, 47(7/8).

Rumsey, F. (2001). *Spatial Audio.* Focal press.

Rumsey, F. (2002). Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society*, 50(9):652–666.

Rumsey, F. (2006). Spatial audio and sensory evaluation techniques- context, history and aims. In *Proceedings of Spatial audio and sensory evaluation techniques conference*, Guilford, UK.

Rumsey, F., Zieliński, Kassier, R., and Bech, S. (2005a). On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *Journal of the Acoustical Society of America*, 118(2):968–976.

Rumsey, F., Zieliński, Kassier, R., and Bech, S. (2005b). Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences. *Journal of the Acoustical Society of America*, 117(6):3832–3840.

Scheiber, P. (1972). Quadrasonic sound system. US patent # 3,632,886.

Scheirer, E. D. (2000). *Music-Listening Systems.* PhD thesis, Massachusetts Institute of Technology.

Schroeder, M. (1965). New method of measuring reverberation. *Journal of the Acoustical Society of America*, 37:409–412.

Schroeder, M. R. (1970). Digital simulation of sound transmission in reverberant spaces. *Journal of the Acoustical Society of America*, 47(2):424–431.

Schroeder, M. R. (1987). Statistical parameters of the frequency response curves of large rooms. *Journal of the Audio Engineering Society*, 35(5):299–305.

Schroeder, M. R., Gottlob, D., and Siebrasse, K. (1974). Comparative study of european concert halls: Correlation of subjective preference with geometric and acoustic parameters. *Journal of the Acoustical Society of America*, 56(4):1195–1201.

Schroeder, M. R. and Kuttruff, H. (1962). On frequency response curves in rooms. *Journal of the Acoustical Society of America*, 34:76–80.

Segar, P. and Rumsey, F. (2001). Optimisation and subjective assessment of surround sound microphone arrays. In *Proceedings of the AES 110th international convention*, Amsterdam, Netherlands.

Shinn-Cunningham, B. G. (2006). The real reasons you should invest in a surround-sound system. *Journal of the Acoustical Society of America*, 119(5):3280.

Shinn-Cunningham, B. G., Durlach, N. I., and Held, R. M. (1998). Adapting to supernormal auditory localization cues I: Bias and resolution. *Journal of the Acoustical Society of America*, 103(6):3656–3666.

Shinn-Cunningham, B. G., Schickler, J., Kopco, N., and Litovsky, R. Y. (2001). Spatial unmasking of nearby speech sources in a simulated anechoic environment. *Journal of the Acoustical Society of America*, 110:118–1129.

Shynk, J. (1992). Frequency-domain and multirate adaptive filtering. *IEEE Signal Processing Magazine*, pages 15–36.

Slock, D. T. M. (1993). On the convergence behaviour of the LMS and

the normalized LMS algorithms. *IEEE Trans. on Signal Processing*, 41(9):2811–2825.

Snow, W. (1953). Basic principles of sterephonic sound. *Journal of the SMPTE*, 61:567–589.

Sommen, P. C. W., VanGerwen, P. J., Kotmans, H. J., and Janssen, A. J. E. M. (1987). Convergence analysis of a frequency-domain adaptive filter with exponential power averaging and generalized window function. *IEEE Trans. on Circuits and systems*, 34(7):788–798.

Sonke, J.-J. (2000). *Variable acoustics by wave field synthesis.* PhD thesis, TU Delft.

Soulodre, G. A., Lavoie, M. C., and Norcross, S. G. (2003). Objective measurements of listener envelopement in mulitchannel surround systems. *Journal of the Audio Engineering Society*, 51(9):826–840.

Spors, S., Kuntz, A., and Rabenstein, R. (2003). An approach to listening room compensation with wave field synthesis. In *Proceedings of the AES 24th International Conference on Multichannel Audio*, Banff, Canada.

SPSS (1995). Statistical package for the social sciences. "Training, building ANOVA models".

Steinke, G. (1996). Surround sound–the new phase: An overview. In *Proceedings of the AES 100th international convention*, Copenhagen, Denmark.

Tak, W. (1958). The 'Electronic Poem' Performed in the Philips Pavilion at the 1958 Brussels World Fair: The Sound Effects. *Philips Technical Review*, 20(2/3):43–44.

Takeuchia, T., Nelson, P. A., and Hamada, H. (2000). Robustness to head misalignment of virtual sound imaging systems. *Journal of the Acoustical Society of America*, 109(3):958–971.

Theile, G. (1991). On the naturalness of two-channel stereo sound. *Journal of the Audio Engineering Society*, 39(10):761–767.

Theile, G. (2000). Multichannel natural recording based on psychoacoustic principles. In *Proceedings of the AES 108th international convention*, Paris.

Theile, G. and Plenge, G. (1977). Localization of lateral phantom sources. *Journal of the Audio Engineering Society*, 25(4):196–200.

Thurlow, W. R. and Runge, P. S. (1967). Effects of induced head movements on localization of direction of sound sources. *Journal of the Acoustical Society of America*, 42:480–488.

Tohyama, M. and Suzuki, H. (1989). Interaural cross-correlation coefficients in stereo-reproduced sound fields. *Journal of the Acoustical Society of America*, 85:780–786.

Toole, F. E. (1982). Listening tests, turning opinion into fact. *Journal of the Audio Engineering Society*, 30:431–445.

Toole, F. E. (1983). Subjective measurements of loudspeakers - A comparison of stereo and mono listening. In *Proceedings of Audio Engineering Society 74th Convention*, New York.

Torger, A. (2001). BruteFIR. `http://www.ludd.luth.se/~torger/brutefir.html`.

Torger, A. and Farina, A. (2001). Real-time partitioned convolution for Ambiophonics surround sound. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York.

Torgerson, W. S. (1958). *Theory and Methods of Scaling.* Wiley, New York.

Trochimczyk, M. (2001). On the signification of spatial sound imagery in new music. *Computer Music Journal*, 25(4):39–56.

Tukey, J. W. (1977). *Exploratory data analysis.* Addison-Wesley, Reading, Mass.

Ueda, Y. and Ando, Y. (1997). Effects of air conditioning on sound propagation in a large space. *Journal of the Acoustical Society of America*, 102:2771–2775.

Usher, J., Cooperstock, J., and Woszczyk, W. (2004a). A multi-filter approach to acoustic echo cancelation for teleconferencing. In *Proceedings of the 147th Meeting of the Acoustical Society of America*, New York.

Usher, J., Martens, W. L., and Woszczyk, W. (2004b). The influence of the presence of multiple sources on auditory spatial imagery. In *Proceedings of the 18th International Congress on Acoustics*, Kyoto, Japan.

Usher, J. and Woszczyk, W. (2003). Design and testing of a graphical mapping tool for analyzing spatial audio scenes. In *Proceedings of the AES 24th International Conference on Multichannel Audio*, Banff, Canada.

Usher, J. and Woszczyk, W. (2004). Visualizing auditory spatial imagery of multi-channel audio. In *Proceedings of the AES 116th international convention*, Berlin, Germany.

Usher, J. and Woszczyk, W. (2005). Interaction of source and reverberance spatial imagery in multichannel loudspeaker audio. In *Proceedings of the AES 118th international convention*, Barcelona, Spain.

Usher, J. S. (2001). Computational auditory scene analysis of two channel audio material to predict image locations spectrally, as would be perceived

on a loudspeaker pair. Undergraduate honours project. School of Acoustics, University of Salford, England. http://www.music.mcgill.ca/~usher/papers/salford/casalist.pdf.

Vanderkooy, J. (1994). Aspects of MLS measuring systems. *Journal of the Audio Engineering Society*, 42(4):219–231.

Verheÿen, E. (1998). *Sound Reproduction by Wave Field Synthesis.* PhD thesis, TU Delft.

Vogel, P. and de Vries, D. (1994). Electroacoustic system response in a hall: a convolution of impulse sequences. *Journal of the Audio Engineering Society*, 42:684–690.

Wagener, B. (1971). Räumliche Verteilungen der Hörrichtungen in synthetischen Schallfeldern. *Acustica*, 25:203–219.

Waller, K. W. (1994). Multi dimensional sound circuit. US patent #5,333,201.

Warren, R. M. (1998). Pitch. Auditory postings (online). http://www.auditory.org/mhonarc/1998/msg00351.html.

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. In *IRE WESCON Convention Record*, pages 96–104.

Widrow, B. and McCool, J. M. (1976). Stationary and nonstationary learning characteristics of the LMS adaptive filter. In *Proceedings of the IEEE*, volume 64, pages 1151–1162.

Woram, J. (1970). Experiments in four channel recording techniques. In *Proceedings of the AES 39th international convention*, New York.

Woszczyk, W., Cooperstock, J., Roston, J., and Martens, W. L. (2005). Shake, rattle, and roll: Getting immersed in multisensory, interactive music via broadband networks. *Journal of the Audio Engineering Society*, 53(1):336–344.

Woszczyk, W. R. (1990). A new method for spatial enhancement in stereo and surround recording. In *Proceedings of the AES 89th international convention*, Los Angeles.

Woszczyk, W. R. (1993). Quality assessment of multichannel sound recordings. In *Proceedings of the AES 12th international conference*, Copenhagen, Denmark.

Yost, W. A. (1991). Auditory image perception and analysis: The basis for hearing. *Hearing Research*, 56:8–18.

Yost, W. A. (1997). The cocktail party problem: Forty years later. In Gilkey, R. H. and Anderson, T. R., editors, *Binaural and spatial hearing*, chapter 17, pages 329–347. Erlbaum, Mahwah, NJ.

Zacharov, N. (1998). Subjective appraisal of loudspeaker directivity for multichannel reproduction. *Journal of the Audio Engineering Society*, 46:288–303.

Zacharov, N. and Koivuniemi, K. (2001a). Audio descriptive analysis and mapping of spatial sound displays. In *Proceedings of the 2001 International Conference on Auditory Display*, Espoo, Finland.

Zacharov, N. and Koivuniemi, K. (2001b). Unravelling the perception of spatial sound reproduction: analysis and external preference mapping. In *Proceedings of the AES 111th international convention*, New York.

Zahorik, P. (1997). Scaling perceived distance of virtual sound sources. *Journal of the Acoustical Society of America*, 101(5):3105–3106.

Zahorik, P. (2002). Assessing auditory distance perception using virtual acoustics. *Journal of the Acoustical Society of America*, 111(4):1832–1846.

Zolzer, U. (1997). *Digital audio signal processing.* Wiley, Chichester.

Zwicker, E. and Fastl, H. (1999). *Psychoacoustics: Facts and models.* Springer, Berlin, 2nd edition.