Diploma Thesis

Audio augmented reality in telecommunication

Hannes Gamper

Signal Processing and Speech Communication Laboratory
Graz University of Technology

Advisor: Univ.-Prof. Dipl.-Ing. Dr. techn. Gernot Kubin
Co-Advisor: Adjunct professor D.Sc.(Tech.) Tapio Lokki

Helsinki, February 2010

Kurzfassung

Telekommunikationssysteme ermöglichen die Kommunikation und Interaktion mit entfernten Benutzern. Audiokommunikation in ihrer natürlichsten Form, dem Gespräch, ist binaural. Heutige Kommunikationssysteme verfügen oftmals nur über monauralen Ton, wodurch räumliche Information verloren geht. Dies führt zu einer Verschlechterung des Hörerlebnisses und der Sprachverständlichkeit. In dieser Arbeit wird die Implementierung eines binauralen Telekommunikationssystems mittels "Audio Augmented Reality" (AAR), also durch Ton erweiterte Realität, vorgestellt. AAR erweitert die auditive Wahrnehmung durch die Einbettung virtuellen Raumklangs. In einem Telekommunikationssystem erhöht die Verwendung von Raumklang die Sprachverständlichkeit und das Gefühl der Immersion. Als Anwendungsbeispiel für AAR dient eine Telekonferenz. Die Konferenz wird über binaurale Kopfhörer mit integrierten Mikrophonen von einem der Teilnehmer aufgezeichnet. Algorithmen zur Kompensation von Kopfbewegungen während der Aufnahme werden präsentiert. Diese stellen eine korrekte Wahrnehmung der Richtung der einzelnen Konferenzteilnehmer sicher. Zur Evaluierung des AAR Systems wurde eine Benutzerstudie durchgeführt. Durch die Aufbereitung der binauralen Aufnahme werden die Richtungen der virtuellen Sprecher fixiert. Dies führte zu einer signifikanten Verbesserung der Unterscheidbarkeit der Konferenzteilnehmer gegenüber einer unbearbeiteten Aufnahme. Durch Unterstützung der räumlichen Trennung binaural aufgezeichneter Klangquellen übertrifft das AAR System konventionelle Telekommunikationssysteme hinsichtlich der Unterscheidbarkeit der Sprecher.

Schlagworte: Audio augmented reality (AAR), Virtual auditory display, binaurale Telekommunikation, binaurales Voice-over-IP (VoIP), KAMARA Headset

Abstract

Telecommunication systems have evolved to allow users to communicate and interact over distance. Audio communication in its most natural form, the face-to-face conversation, is binaural. Current telecommunication systems often provide only monaural audio, stripping it of spatial cues and thus deteriorating listening comfort and speech intelligibility. In this work, the application of binaural audio in telecommunication through audio augmented reality (AAR) is presented. AAR aims at augmenting auditory perception by embedding spatialised virtual audio content. Used in a telecommunication system, AAR enhances intelligibility and the sense of presence of the user. As a sample use case of AAR, a teleconference scenario is devised. The conference is recorded through a headset with integrated microphones, worn by one of the conference participants. Algorithms are presented to compensate for head movements and restore the spatial cues that encode the perceived directions of the conferees. To analyse the performance of the AAR system, a user study was conducted. Processing the binaural recording with the proposed algorithms places the virtual speakers at fixed directions. This improved the ability of test subjects to segregate the speakers significantly compared to an unprocessed recording. The proposed AAR system outperforms conventional telecommunication systems in terms of the speaker segregation by supporting spatial separation of binaurally recorded speakers.

Keywords: Audio augmented reality (AAR), virtual auditory display, binaural telecommunication, binaural voice-over-IP (VoIP), KAMARA headset

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

....

date (signature)

Acknowledgements

This diploma thesis was written in the course of the KAMARA 2009 project, a cooperation of the Nokia Research Center Helsinki, the Department of Signal Processing and Acoustics and the Department of Media Technology of the Helsinki University of Technology, where the research for this thesis was conducted.

Thanks to D.Sc. Tapio Lokki for granting me the chance to work in this project and for his academic guidance and inspiration throughout my thesis work. Thanks also to the Virtual Acoustics team for the welcoming atmosphere at the laboratory, which made my stay in Finland so enjoyable.

Thanks to Prof. Matti Karjalainen and the people from the Acoustics Lab and the Nokia Research Center for the collaboration in the course of this project.

Thanks to Prof. Gernot Kubin for approving of my plans to conduct the research in Finland and for supervising my thesis.

Thanks to Prof. Mark Billinghurst for his assistance with the design and analysis of the user study.

Thanks to my family and friends for their support all through my studies. Special thanks to my parents for teaching me the importance of education.

Helsinki, February 2010 Hannes Gamper

Contents

List of Figures

List of Tables

List of abbreviations

AAR	audio augmented reality	1, 2, 4, 5, 12–21, 23, 26, 27, 29, 30, 45, 65–67
ANOVA	analysis of variance	52–54, 56, 58, 59
AR	augmented reality	1, 3–5, 17, 19, 20, 33, 67
BRIR	binaural room impulse response	25, 41, 42, 67
HRIR	head-related impulse response	8, 22, 25
HRTF	head-related transfer function	8–10, 15, 22–25, 38–43, 66
IACC	interaural cross-correlation	37
IHL	inside-the-head locatedness	14, 22, 23, 25, 42, 43, 60, 61, 63, 67
ILD	interaural level difference	7, 8, 10, 15, 21–23, 35–39, 42, 43, 51, 66, 67
ITD	interaural time difference	6–8, 11, 15, 21–23, 26, 34, 35, 37–40, 42, 43, 51, 66, 67
KAMARA	killer applications for mobile augmented reality audio	2, 27, 29, 32, 33, 38, 39, 41, 42, 45, 46, 48, 61, 63, 65–67
Pd	Pure Data	30, 32, 39
SHAKE	sensing hardware accessory for kinesthetic expression	17, 30–33, 38, 47
VAD	virtual auditory display	19, 20
VoIP	voice-over-IP	1, 2, 32, 33, 65
VR	virtual reality	3

Chapter 1

Introduction

1.1 Motivation

Removing the "tele" of telecommunication

Research in the area of augmented reality (AR) deals with the question of how to blur or remove the boundaries between the real physical world and informationally enhanced virtual environments [Azuma et al., 2001, Milgram et al., 1995]. The goal of AR is to "augment" the human sensory perception with virtual content. Whilst this is often limited to the sense of vision, the present work is an attempt to point out the potential of augmenting the auditory perception. Audio augmented reality (AAR) has applications in various areas, ranging from entertainment to military use cases (cf. section 2.6).

In this work, an AAR telecommunication system is proposed. Conventional telecommunication systems often provide the user only with a monaural audio stream, played back via a headset or a hand-held device. The term "monaural" refers to the fact that only one ear is necessary to interpret the auditory cues contained in the audio stream. However, face-to-face communication, which is considered the "gold standard" of communication [Nardi and Whittaker, 2002, Rohde et al., 1997], is inherently binaural. In a face-to-face conversation, a listener is able to segregate multiple talkers based on their position, a phenomenon called the "cocktail party effect" [Cherry, 1953]. The position information is encoded into the audio stream in the form of spatial cues. The most important cues are interaural differences, i.e. differences between the ear signals. Monaural audio as employed in conventional telecommunication systems, such as mobile phones and voice-over-IP (VoIP) softwares, does not support interaural cues and hence deteriorates the communication performance compared to face-to-face communication [Lindeman et al., 2009, Billinghurst et al., 2002].

AAR aims at overcoming these limitations by embedding "realistic" virtual audio into the auditory perception. This requires careful design of the ear signals, taking into account the properties of spatial hearing. In doing so, AAR takes full advantage of the capabilities of the human auditory system, which is particularly beneficial in a communication scenario. The "cocktail party" principle for instance holds also for a multi-party telecommunication scenario. Separating the speech signals of participants spatially, through the use of AAR, improves the listening comfort and intelligibility [Kapralos et al., 2008, Drullman and Bronkhorst, 2000]. In contrast to the sense of vision, auditory perception is not limited to a "field of view". Therefore, in AAR, virtual audio can be placed anywhere in the surroundings of a listener, regardless of the listener's orientation. The participants of a teleconference can thus be distributed all around the user. By registering the virtual speakers with the environment, an AAR user can turn towards a conferee the same way as in a face-to-face conversation.

A major challenge in telecommunication lies in the physical distance itself, which puts limits to the naturalness of interaction with a remote end. Communication over distance suffers

from a lack of "social presence", compared to face-to-face communication [Bazerman et al., 2000]. Through spatial audio, an AAR telecommunication system improves the sense of "presence" [Shilling and Cunningham, 2002, Lehnert and Blauert, 1991] and "immersion" [Kapralos et al., 2008]. AAR attempts to engage users in a similar fashion as they would be in a face-to-face conversation, thus resolving the limitations distance imposes on communication. It takes advantage of the benefits of binaural audio signals and incorporates them into a communication scenario. In this work, an AAR telecommunication system based on binaural recordings is presented. The recordings are obtained from the KAMARA headset [Härmä et al., 2004], a binaural headset with embedded microphones.

1.2 Scope of the thesis

The research for this work was conducted in the course of the KAMARA 2009 project. Previous work in the KAMARA project has mainly focussed on the perception of one's own environment through the KAMARA headset as a "pseudo-acoustic environment" [Härmä et al., 2004]. Little work has been done to study the implications of listening to someone else's environment. In a AAR telecommunication system, the user is provided with binaural audio from a remote end. In the test scenario, two users are connected via binaural voice-over-IP (VoIP). Both users are wearing a KAMARA headset. The signals of the microphones of the remote KAMARA headset are transmitted to the earphones of the local headset. Thus, the local user's auditory perception is augmented with a binaural recording of the environment of a remote user.

The aim of this work is to study telecommunication over such a binaural VoIP connection. Algorithms are developed to process a binaural recording from a remote end and embed it into the auditory perception of the local user. This serves as a proof-of-concept for employing AAR in a telecommunication scenario. A user study is conducted to analyse the ability of users to localise and segregate remote speakers with the proposed system.

1.3 Organisation of the thesis

Chapter 2 gives a brief introduction to AAR. It discusses basic principles of spatial hearing and enabling technologies for the creation of virtual audio content for AAR. The chapter closes with an overview of related AAR work. Chapter 3 describes the auralisation of virtual sound for AAR using headphones. Particular attention is given to the externalisation of virtual audio content. Chapter 4 introduces the experimental setup of an AAR telecommunication system. Chapter 5 presents the results of a user study conducted to analyse and evaluate the performance of the test system. Chapter 6 concludes the work with a discussion of the main results and an outlook on future research directions.

Chapter 2

Audio augmented reality

2.1 Introduction to augmented reality

> "**augmented reality** *n.* the use of technology which allows the perception of the physical world to be enhanced or modified by computer-generated stimuli perceived with the aid of special equipment; reality as perceived in this way." [The Oxford English Dictionary, 2006]

Since Daedalus in the Greek mythology crafted wings of wax and feathers to help him and his son Icarus escape their exile on Crete, mankind has strived for overcoming the limitations of the human physique through the invention of appropriate aids. The desire to fly being an ancient example, more recently science fiction authors picture super-human abilities related to the sensory apparatus, such as "x-ray" eyes and super-ears. Brooks, in his acceptance lecture of the Allen Newell Award, indeed imagines "intelligence amplifying systems" [Brooks, 1996] as future tools to push the boundaries of the human mind's capabilities.

In 1968 Sutherland published his work "A head-mounted three dimensional display" [Sutherland, 1968], where he presented an apparatus that allowed the user to perceive virtual 3-D objects embedded into the real surroundings. The objects were shown on a see-through display mounted in front of the user. By changing the perspective of the objects in accordance with the head movements, a "kinetic depth effect" was achieved, adding a virtual depth dimension to the information displayed. The virtual objects thus appeared to be "overlaid" onto the real world. This was to mark the beginnings of what is called augmented reality (AR) [Azuma et al., 2001]. As the name implies, AR aims at augmenting the sensory perception through additional (computer generated) stimuli and information [Rozier et al., 2000]. As an application of AR, Furness postulates the creation of virtual visual, auditory, and tactile worlds in military crew stations to "make optimum use of the spatial and psychomotor capabilities of the human" [Furness, 1986]. In reference to Brooks' lecture [Brooks, 1996], Azuma states that AR is in fact an example for "intelligence amplifying" tools, as it "enhances a user's perception of and interaction with the real world" [Azuma, 1997]. He describes AR as a variation of virtual reality (VR), with the following properties [Azuma et al., 2001]:

Augmented reality (AR)

- combines real and virtual objects in a real environment;
- runs interactively, and in real time; and
- registers (aligns) real and virtual objects with each other.

In contrast to VR, AR resembles an overlay onto, rather than a replacement of, the real world. Whereas in VR the user should ideally feel fully immersed into a virtual scenery, AR is intended

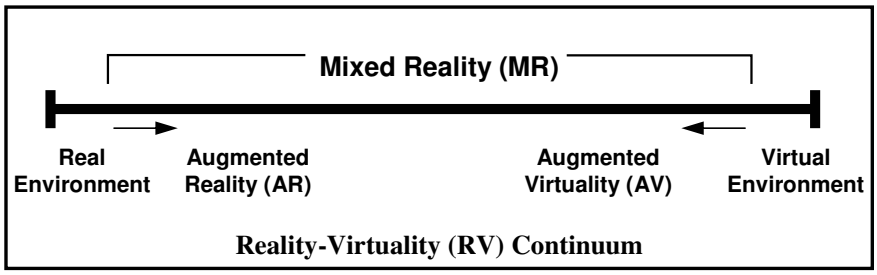


Figure 2.1: *Reality–virtuality continuum.* The scale shows the transition between the real and virtual worlds. Augmented reality combines elements of both ends. Adapted from Milgram [Milgram et al., 1995].

to assist, entertain, or guide the user inside the real world. It could in fact be interpreted as an interface between purely virtual digital objects and the physical reality [Pentenrieder et al., 2007], or an intermittent step on the "reality–virtuality continuum" [Milgram et al., 1995] (see fig. 2.1).

The possibility of embedding informative visual and auditory objects into the surroundings of the user gives rise to a wide range of (future) application areas: From industrial factory planning [Pentenrieder et al., 2007] and aircraft maintenance [Regenbrecht et al., 2005] to outdoor gaming [Avery et al., 2005, Piekarski and Thomas, 2002], battlefield navigation [Julier et al., 2000], and finally the aforementioned "x-ray Vision" [Azuma, 1997] – for doctors, by projecting 3-D datasets such as computed tomography (CT) onto the patient in real time.

In the development of user interfaces, prevalence has traditionally been given to the human vision over other senses [Cohen and Wenzel, 1995]. This manifests itself also in the research related to AR: Most of the application scenarios presented above consist of purely visual augmentation of the user's surroundings, at the expense of other sensory stimuli, and sound in particular. This imbalance seems unfortunate, given the potential of using auditory objects in an AR scenario. Sound is a key element for conveying information, attracting attention, creating ambience and emotion [Shilling and Cunningham, 2002]. George Lucas, a world famous movie director and producer, indeed stated that "Sound is 50 percent of the moviegoing experience." [Lucasfilm Ltd., 2004]. Enhancing the perception through virtual auditory objects leads to the definition of an audio-based alternative to visual AR: audio augmented reality (AAR).

2.2 Definition of audio augmented reality

> "Audio augmented reality [...] intends to superimpose virtual sounds to a physical space" [Warusfel and Eckel, 2004]

Whilst Azuma [Azuma et al., 2001] claims he does not limit AR to the sense of sight, his definition of it is clearly tailored to scenarios focussing on visual augmentation of reality. Cues provided to other senses, such as hearing and touch, are thus often either omitted completely [Sutherland, 1968, Pentenrieder et al., 2007, Piekarski and Thomas, 2002, Echtler et al., 2003, Feiner et al., 1997] or included only to support or enhance the visuals [Furness, 1986, Behringer et al., 1999]. In contrast to this, examples of audio-only or audio-centred AR are presented and discussed. In this audio augmented reality (AAR) sound is the only or major cue used for augmenting reality. It can be defined in reference to Azuma's definition of AR [Azuma et al., 2001] as follows:

audio augmented reality (AAR)

- combines real and virtual audio objects in a real auditory environment and
- runs interactively, and in real time.

An audio object is a perceptual entity identifiable by a listener as a single source of auditory events (e.g. a person talking, a musical instrument, but also abstract sounds not originating from a physical source such as synthesised alarms), and thus separable from other audio objects. This concept is comparable to the "spatial audio objects" in spatial audio object coding (SAOC) [Herre and Disch, 2007]. Following the definition of AR [Azuma et al., 2001], real objects are those that evoke sensory events (either visual or auditory) due to their physical presence in the user's surroundings. Virtual objects, on the other hand, evoke sensory events as caused by a physical source, in absence thereof. In the acoustic domain, real audio objects are thus those coinciding with a physical sound source. A virtual audio object, however, causes an auditory event whose position does not coincide with the position of the sound source (a more detailed description of auditory events is given by Blauert [Blauert, 1996]).

Registering real and virtual objects with each other is not included in the definition of AAR, as it is a concept not directly translatable from the visual to the acoustic domain. Spatial alignment is less critical for sounds than for visual objects, due to the properties of the human hearing. The minimum audible angle, i.e. the smallest sound source displacement detectable to the human ear, reaches its lower limit of 1° in front of the listener, at about 500 Hz [Mills, 1958]. In comparison, the human eye is able to detect differences of less than 1 minute of arc [Behringer et al., 1999], corresponding to a sixtyfold spatial resolution. Blauert refers to this phenomenon as the "localisation blur" of the human ear [Blauert, 1996] (see section 2.3.4). This implies that audio objects are in general less defined than visual objects, and have a somewhat "blurred" position and extent. The accuracy of their registration with other audio objects is therefore of less importance than in the visual domain. Also, a possible "misalignment error", e.g. two overlapping audio objects, is less disturbing than an error in the overlap of visual objects, as the latter void the physical plausibility of the virtual scene. This occlusion problem, whilst being a major issue in the registration of visual objects [Zhu and Pan, 2008], does not apply as such to the acoustic domain. Sound sources overlapping in space are both physically feasible (e.g. two talkers played through a single loudspeaker) and acoustically separable (the brain performs an "auditory scene analysis" to segregate different sound sources [Shilling and Cunningham, 2002]).

The alignment of audio objects in an AAR scenario is not only less critical than in the visual domain, it can indeed be entirely neglected for some applications (e.g. the "diary in the sky" [Walker et al., 2001]). Lokki defines these nonregistered audio objects as "freely floating acoustic events" [Lokki et al., 2004], as they are neither tied to nor defined spatially relative to other elements of the auditory scene. Their only point of reference is the user's head. "Localized acoustic events" [Lokki et al., 2004], on the other hand, are audio objects that are overlaid onto physical objects in the user's surroundings. The design and creation of these virtual acoustic events are based on the principle of human spatial hearing.

2.3 Spatial hearing

The human hearing experience consists of a conglomerate of auditory events that are distinct in time and space [Blauert, 1996]. Hearing, i.e. the perception of an auditory event that occurs at a certain time, at a certain place, with certain attributes, is thus inherently spatial [Blauert, 1996]. In the following sections the human ear's ability to perform "localisation", i.e. the act of relating attributes of the sound present at the ears to the location of an auditory event [Blauert, 1996], is examined.

The focus lies on determining which sound attributes are related to the azimuth of an auditory event (in particular its direction on the horizontal plane), which are related to the elevation (in particular its direction on the median plane), and which determine the perception of distance and whether it is located in front or in the back of the listener (cf. fig. 2.2). For the following considerations the positions of the sound source and the auditory event are assumed to coincide.

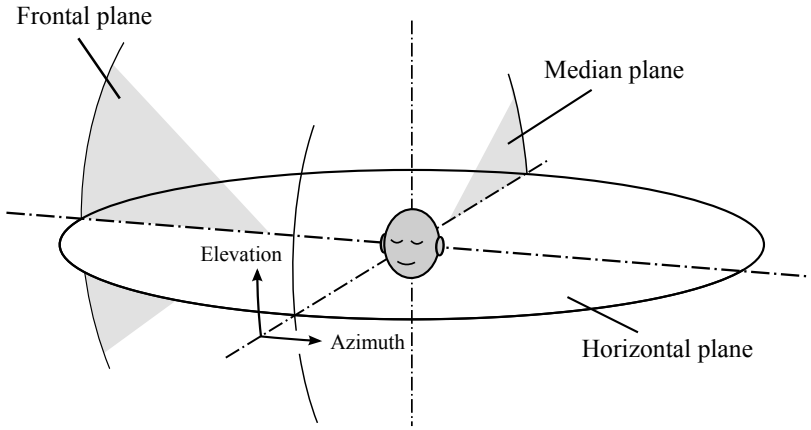


Figure 2.2: *Median, frontal and horizontal plane.* After Lentz and others [Lentz et al., 2006].

This means that the sound attributes described hereafter enable and favour the perception of a sound source at its true spatial location. (For details on noncoincident perception see section 2.4). It is assumed that the listener and the sound source are situated in an anechoic environment, where no reflections from surrounding objects, walls, etc. occur. Some implications of reflections on the spatial perception are mentioned in section 2.3.9.

2.3.1 Lateral localisation and binaural spatial cues

Over a hundred years ago, Lord Rayleigh studied the human localisation by presenting pure tone stimuli with different frequencies to test subjects using tuning forks. His assumption was that differences in the ear signals, called "interaural" cues [Blauert, 1996], could be interpreted by the human ear for localisation. He found that for frequencies above 500 Hz the human auditory system can evaluate level differences between the ears to determine the lateral position of a sound source. For lower frequencies he stated that the lateral position can be inferred from a difference in the ongoing phase. In his "Duplex Theory" he describes this link between level and phase differences as the major cue for human localisation [Macpherson and Middlebrooks, 2002].

Rayleigh's findings can be explained through basic principles of physics. Sound propagates in air in spherical longitudinal waves, if the sound source is small compared to the wavelength [Rossing and Fletcher, 2004]. For a sound source positioned to the left or the right of the median plane, the propagation paths from the source to the ears differ in length. The wave front therefore first reaches the ipsilateral ear (i.e. the ear to the same side of the median plane as the sound source), and then the contralateral ear (i.e. the ear to the opposite side of the median plane as the sound source), with a delay proportional to the path difference. This delay is called interaural time difference (ITD) and represents a major cue for the auditory system to determine the lateral position of a sound source. For pure tones, the ITD can be derived from phase differences. Above 1.5–1.6 kHz, however, no lateral displacement is noticeable, due to the relation between head size and wavelength [Blauert, 1996]. In this frequency region, the wavelength approaches 20 cm, which roughly corresponds to the distance between the ears. Thus, for frequencies above this limit, phase differences are ambiguous. For complex high-frequency sounds, however, timing information can be extracted from the sound envelope [Macpherson and Middlebrooks, 2002].

Representing the ears by two points in the horizontal plane, all source positions that result in the same path length difference to the ears and thus in the same ITD lie on a hyperbola [Blauert, 1996]. In the far field (i.e. for large source distances), the hyperbola can be approximated by its asymptotes, assuming planar wave fronts [Rossing and Fletcher, 2004]. In three-dimensional

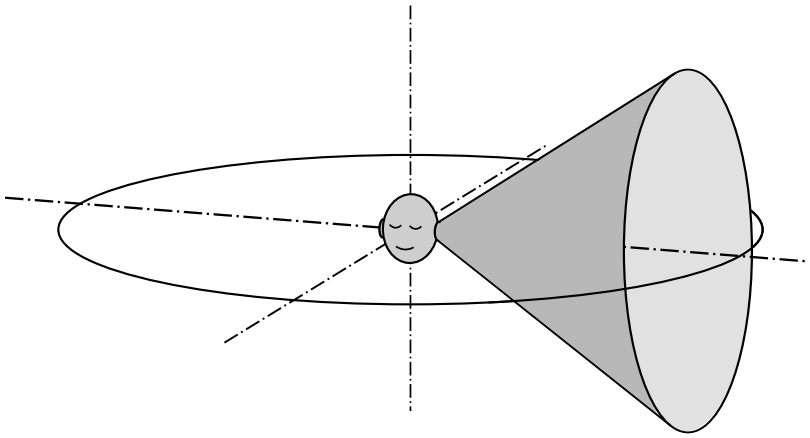


Figure 2.3: *Cone of confusion.* Representing the ears by two points, all sources lying on the shell of a cone yield the same ITD.

space, an elliptical hyperboloid, or its approximation, the shell of a cone, describes the locus of all source positions with the same ITD (cf. fig. 2.3). The actual source position on this cone cannot be determined from the ITD alone, which is why it is called the "cone of confusion" [Blauert, 1996]. Duda and others, however, note that this assumption holds only for a spherical head model [Duda et al., 1999] and that the ITD values can change remarkably around this cone of confusion (see section 3.1).

As described above, the wave front passes the head on its way from the source to the contralateral ear. For small wavelengths compared to the size of the head, acoustic shadowing occurs, resulting in a lower sound pressure level at the contralateral than at the ipsilateral ear. This is referred to as interaural level difference (ILD). The ILD diminishes below 1.5 kHz, due to head diffraction for wavelengths larger than the head size [Rocchesso, 2002]. For narrowband signals, the same ILD may occur for different directions of sound incidence [Blauert, 1996]. Therefore, the ITD and ILD cues are sufficient to determine the azimuth of a source, but cannot resolve the cone of confusion, and in particular the vertical or front–back localisation [Macpherson and Middlebrooks, 2002], which is the focus of the following sections.

2.3.2 Vertical localisation and monaural spatial cues

For a sound source lying anywhere on a cone of confusion, the interaural time and level differences remain approximately constant. A special case of such a cone of confusion is the median plane, where both ITD and ILD are close to zero and thus no binaural cues are present. Yet for sufficiently long or repeated broadband signals the auditory system is able to determine the elevation of a sound source even in the median plane [Blauert, 1996]. As the ear signals in this case, assuming a symmetrical head, are identical, in principle only one ear is necessary to interpret them. This leads to the assumption that directional hearing in the median plane is based on monaural cues. In fact these monaural cues allow for the perception of source elevation on any cone of confusion [Pulkki, 2001].

Localisation cues in general can be subdivided into temporal and spectral cues (cf. table 2.1). Wightman and Kistler argue that the auditory system is not able to interpret monaural temporal cues and that they are therefore irrelevant for human sound localisation [Wightman and Kistler, 1997]. The authors argue, that the perception of elevation is instead mainly based on monaural spectral cues. These cues take the form of direction-dependent peaks and valleys in the spectrum of an incoming sound, caused by the filtering behaviour of the pinna.

Blauert [Blauert, 1996] experimented with narrow-band noise emitted from various source

	Temporal	Spectral
Monaural	Monaural phase	Overall level Monaural spectral cues
Binaural	interaural time difference (ITD)	interaural level difference (ILD) Binaural spectral differences

Table 2.1: *Localisation cues.* The ILD is taken as the wide-band level difference. Adapted from Wightman [Wightman and Kistler, 1997].

positions in the median plane. He states that the perceived direction is entirely dependent on the centre frequency of the noise signal. This implies that spectral peaks are the predominant cues for the perception of elevation. Bloom [Bloom, 1977] concluded that a narrow notch with elevation dependent centre frequency in the spectrum of a signal could create the impression of source elevation.

Though it has been shown by these studies how a source can be made to appear to be elevated by applying elevation dependent spectral peaks and/or notches, their contribution to the perception of elevation is not yet fully understood [Wightman and Kistler, 1997]. Recent studies claim that the combination of peaks and notches in the signal spectrum above 5 kHz caused by the filtering behaviour of the concha is responsible for the perception of sound source elevation [Iida, 2008]. Based on these assumptions, Saxena and Ng [Saxena and Ng, 2009] succeeded in localising sound sources using only a single microphone (and thus truly monaural signals), by evaluating the distortions introduced to the sounds by an artificial pinna. In the lower frequency range, torso reflections can introduce elevation-dependent notches, that may serve as elevation cues [Zotkin et al., 2004].

For the human auditory system to be able to recognise these spectral changes, especially in absence of binaural cues, it must have prior knowledge about the source spectrum. Familiarity with the source signal may therefore be one prerequisite for directional hearing in the median plane [Blauert, 1996], spectral content of the signal above 5 kHz another [Wightman and Kistler, 1997], though it has been shown by Algazi and others that monaural spectral features exist also below 3 kHz [Algazi et al., 2001b]. The third and most important requirement however is the knowledge of the pattern of these direction-dependent spectral changes, which is discussed in the next section.

2.3.3 Head-related transfer function

Sound travelling from a source to the eardrums of a listener undergoes reflections and shadowing caused by the pinnae, the head and the torso, that impose a characteristic shape onto its spectrum. As stated in the previous section, the human auditory system evaluates changes in the spectrum of an incoming signal to derive its direction [Wightman and Kistler, 1997]. To investigate upon this human sound localisation process, Wightman and Kistler [Wightman and Kistler, 1989] and Middlebrooks and others [Middlebrooks et al., 1989] conducted careful measurements of the filtering behaviour of the human head and torso by inserting probe microphones into the ears of human test subjects. Herewith the linear distortions of a test signal on its way to the eardrums was measured. This filtering behaviour is described by the head-related transfer function (HRTF). It is defined as the relation of the sound pressure at a point in or in front of the human ear canal to the sound pressure at the centre of the head in absence of the head [Riederer, 1998]. The HRTFs and their time-domain analogue, the head-related impulse responses (HRIRs), proved to be highly direction-dependent. Stern and others define this as the

beginnings of the "modern era" of research in the area of human sound localisation [Stern et al., 2006].

Blauert states that the ear input signals, i.e. the sound signals in the ear canals, are the primary input for spatial hearing [Blauert, 1996]. Sound transmitted to the inner ear via the temporal bone through bone conduction [Blauert, 1996] (cf. section 2.5 for an example application) is of secondary importance. It can therefore be concluded that the acoustic cues responsible for human sound localisation are contained in the ear input signals as a result of the direction-dependent HRTFs. Measuring and analysing them sheds light on the process of human sound localisation and is of ongoing research interest.

Acquisition of head-related transfer functions

Various studies have aimed at attaining a comprehensive set of HRTFs to cover a whole range of angles of sound incidence and describe their filtering characteristics [Gardner and Martin, 1995, Møller et al., 1995, Algazi et al., 2001c, Warusfel and Eckel, 2004, Kim et al., 2005]. The measurement procedure in those studies generally follows a similar outline. The test subject is seated in an anechoic environment and exposed to a test signal originating from a defined direction. The signal is recorded at both ears via microphones placed at the subject's ear entrances (referred to as "blocked meatus" measurements, which are generally preferred over measurements at or close to the ear drum [Hammershøi and Møller, 1996]). The direction of sound incidence is systematically changed in azimuth and/or elevation, yielding a spherical or hemispherical measurement grid.

As HRTFs vary from person to person, measurements have been made using a number of human test subjects to identify and analyse individual HRTF differences [Møller et al., 1995, Algazi et al., 2001c, Warusfel and Eckel, 2004]. Measurements involving human test subjects, besides being highly individual, are however prone to measurement errors. The measurements are very sensitive to the placement of the test subject and the microphone, as well as movements of the subject. Riederer addresses these issues in his study on the repeatability of HRTF measurements [Riederer, 1998]. To improve controllability and repeatability and to obtain an HRTF dataset for a "mean" anthropometry, measurements were often performed using a dummy head [Bovbjerg et al., 2000, Algazi et al., 2001c, Kim et al., 2005], such as the KEMAR head and torso manikin [Burkhard and Sachs, 1975, Gardner and Martin, 1995].

As a result of these studies, and to support further research in the area, some HRTF databases are freely available online [Ircam & AKG Acoustics, 2002, CIPIC/IDAV Interface Laboratory, 2004]. For methods to customise and model HRTFs see section 3.3.2.

2.3.4 Localisation blur

Blauert introduces the term "localisation blur" to describe the accuracy of the human spatial hearing [Blauert, 1996]. It is defined as the minimum displacement of a sound source perceivable by 50 percent of experimental subjects. The localisation blur is lowest in front of the listener, where lateral displacements of the sound source of about 1° are detectable. This sets the upper limit for the spatial resolution of the human auditory system. Away from the forward direction in the horizontal plane the localisation blur in general increases, reaching maxima to the sides of the listener. Its value also depends on spectral content and duration of the test signal.

For vertical displacement the localisation blur is in general higher than for horizontal displacement. The minimum blur in the forward direction is about 4° for white noise, 9° for speech of a familiar person and 17° for speech of an unfamiliar person. The value increases for sources above or behind the listener.

Localisation blur also occurs in distance perception. It describes the accuracy by which listeners are able to judge the distance of a sound source.

2.3.5 Perception of distance

The distance of an auditory event is defined relative to the midpoint of the axis connecting the ears [Blauert, 1996]. A distance of zero thus refers to an auditory event being located inside the head. This is referred to by Blauert as "inside-the-head locatedness" and considered a fault of binaural reproduction systems (cf. section 3.3).

The perception of distance is strongly dependent on familiarity with the signal. Normal speech can be located quite well by human listeners. The localisation performance degrades for unusual types of speech such as whispering [Blauert, 1996]. For unfamiliar sounds more than 3 m away distance is mainly dependent on the loudness of the signal: Louder sounds appear to be closer and vice versa.

Efforts have been made to identify distance cues in HRTFs [Qu et al., 2008, Huopaniemi and Riederer, 1998, Otani et al., 2009]. For relatively distant sources (more than 3 m), little or no dependence of the HRTFs on the source distance has been reported [Zotkin et al., 2004]. For sources in the proximity of the listener (distances below 2 m), however, changes in the ILD can be observed. Huopaniemi points out an ILD boost for sources close to either side of the listener [Huopaniemi and Riederer, 1998]. Similar results are reported by Otani and Hirahara [Otani et al., 2009].

For sources further away than about 1 m, reverberation can provide cues for distance perception [Shinn-Cunningham et al., 2005, Larsen et al., 2008]. Blauert states that the distance localisation blur, i.e. the minimum change of source distance perceivable under reverberant conditions (2–3 percent) is considerably lower than the localisation blur in an anechoic environment [Blauert, 1996]. Listeners can determine source distance more reliably in a reverberant environment, by assessing the direct-to-reverberant energy ratio [Larsen et al., 2008]: The energy of the direct sound field of a sound source decays with increasing distance, while the energy of the reverberant sound field remains approximately constant. As a listener moves closer to a source in a reverberant space, the direct signal energy and therefore the direct-to-reverberant ratio increases. The human auditory system is able to interpret this ratio to estimate the source distance.

2.3.6 Motional cues

HRTFs bear information about the position of a sound source. If the listener moves the head, the relative source position and therefore the monaural and interaural cues change in a particular way. The pattern of these changes is a "motional cue", interpretable by the auditory system. In the frequency domain these dynamic cues manifest themselves as a change of the spectral shape of the ear input signals. Satarzadeh and others point out that in many studies notches in static HRTFs are identified as major elevation cues [Satarzadeh et al., 2007]. Background noise and the unknown source spectrum however make the detection of these notches rather difficult. Head motion on the other hand reveals these notches more clearly, as the spectrum and thus the notches themselves change in a determined way due to the movement. According to Satarzadeh and others, studying psychoacoustic features and localisation in dynamic settings is a complex and relatively unexplored area of research.

Blauert states that people who are deaf in one ear can facilitate the localisation of a sound source by obtaining these motional cues through head movements [Blauert, 1996]. Such reflexive and/or voluntary head movements can improve localisation also for people with normal hearing. Turning the head towards the source decreases the localisation blur, and thus allows for a sharper determination of the source position.

Perhaps the most important motional cue serves for resolving the problem of front–back reversals in source localisation. As stated earlier, interaural cues may be ambiguous (see section 2.3.1). This often leads to a misjudgement of the source position: When presented with

binaural signals or a binaural recording, listeners often mistake sources in front to be in the rear and vice versa [Wenzel et al., 1993]. This confusion can be resolved through head movements. Assuming a source in the median plane, in theory no interaural cues are present. It cannot be determined from those cues whether the source is in front or in the back. However, if the listener turns the head to the left or to the right, a change in the interaural cues will occur. As one ear moves closer to the source and the other moves further away, the path length difference and thus the ITD changes. This change is detectable by the auditory system. Its sign depends upon the source position, and therefore unambiguously determines whether the source is in front or in the back. As Blauert points out, this motional cue is not purely auditory [Blauert, 1996]. Besides the change in the ear input signals the listener must recognise the direction of the head movement, through the senses of vision and balance, and the position of the neck muscles, which is why Blauert calls it a heterosensory cue.

2.3.7 Visual capture and multi-modal phenomena

The perception of the world is based on the input of a variety of senses. Whilst in the analysis of spatial hearing mostly auditory cues are considered, other senses influence the perception of sound. There is a strong interaction of auditory and visual cues. Malcolm Slaney published a "Critique of pure audition" [Slaney, 1998], where he explains crossmodal phenomena in the auditory and visual domain, i.e. how vision can affect audition and vice versa. Vision being one of the strongest human senses, it can dominate the perception of other senses, a phenomenon called "visual capture" [Yost, 1993]. Its impact in the acoustic domain lies in the fact that information gathered from visual cues may override conflicting information from auditory cues. The effect is particularly strong if temporal variations of the visible object are synchronous to sound fluctuations. A ventriloquist for example makes use of this effect. By moving the lips of a puppet synchronous to the speech the sound is perceived as emanating from the mouth of the puppet, even though the actual sound source is the mouth of the ventriloquist. Blauert reports an experiment by Klemm, where two hammers were placed to the left and right in front of a test subject [Blauert, 1996]. Two microphones, one next to each hammer, captured the sounds emanating from the hammers. These sounds were played back to the subject over headphones, inverting left and right channels. The inversion was perceived by subjects with eyes closed. When watching the hammers, however, subjects heard the hammer blows on the same side as they saw them. Vision had overridden the auditory perception.

Besides vision, Blauert also mentions the interaction of the senses of balance and touch on the perception of sound [Blauert, 1996]. They are however not considered in this work.

2.3.8 The cocktail party problem

So far only the cues arising from a single source have been considered. In a real listening situation, however, the listener is often surrounded by multiple competing sound sources. A typical example is a room filled with people engaged in a conversation. Referring to this situation, Cherry coined the name "Cocktail Party Problem" [Cherry, 1953]:

> "How do we recognize what one person is saying when others are speaking at the same time (the 'cocktail party problem')?"

Over the past decades researchers have given various explanations for the fact that a human listener can focus on a particular speaker in a group of speakers. Blauert states that by concentrating on one source, its signal is less masked by noise from other sources [Blauert, 1996]. The listener is thus able to identify and separate one source from competing others. Yost lists seven physical attributes that enable sound source determination [Yost, 1997], the most important being spectral separation, temporal separation and spatial separation of the source and the

competing sound sources. Cherry points out that factors like voice attributes (accent, pitch, speed, dynamics, gender), visual cues (lip movement, gestures) and context (conversation topic, syntax) might contribute to this ability. Other factors such as reverberation may deteriorate the intelligibility [Bronkhorst, 2000].

The above authors agree on the fact that the direction of a source seems to be a major cue for resolving the Cocktail Party Problem. One conclusion that can be drawn from these findings is that spatial separation of a sound source in the presence of competing sources can in general improve the intelligibility of that source. This also holds for virtual sources presented to the listener for example over headphones. To achieve spatial separation of virtual sounds, they have to be enhanced with distinct spatial attributes. This process called "auralisation" is discussed in section 2.4.

2.3.9 The precedence effect

In a room, the direct signal of a sound source is followed by multiple reflections from various directions. Each reflection constitutes a delayed copy of the direct signal, as though it would emanate from a mirror source. The auditory system resolves these conflicting cues and perceives a single source with fixed position. The direct signal undergoes the shortest path, and hence reaches the ears before all reflections. Therefore, the auditory system determines the direction of the first wavefront, and derives the source position from that direction. This "precedence effect" allows for sound source localisation in reverberant spaces [Rocchesso, 2002].

2.4 Auralisation in audio augmented reality

In the previous sections it was established why humans perceive an auditory event at the position of a sound source. As Blauert points out, however, it is particularly interesting to create auditory events at positions where no sound source is present:

> "The telecommunications engineer, of course, is especially interested in just those cases in which the positions of the sound source and the auditory event do not coincide." [Blauert, 1996]

This is a fundamental aspect in AAR, where auditory events need to be created at arbitrary positions. The following sections explain how this can be achieved using a fixed number of sound sources at fixed locations, by applying the principles of human spatial hearing.

To create a virtual sound source, ear input signals have to be presented to the listener that a real source from the same position would cause. This is referred to as auralisation. Kleiner and others define it as follows:

> "Auralization is the process of rendering audible, by physical or mathematical modeling, the sound field of a source in space, in such a way as to simulate the binaural listening experience at a given position in the modeled space." [Kleiner et al., 1993]

Auralisation in AAR is subject to further requirements. The acoustical augmentation of reality is based on virtual sounds being overlaid onto physical space. From the definition of AAR (see section 2.2), the criteria that these virtual sounds have to meet may be inferred:

They need to

- be combined with non-AAR sounds,
- be positioned arbitrarily and
- be interactive.

First of all, as AAR aims at augmenting rather than replacing reality, the virtual sounds have to coexist with real sounds. Thus some electro-acoustical transducers to reproduce virtual sounds and mix them with the real auditory scene are necessary for AAR (cf. section 2.5).

As established earlier, those virtual sounds and their corresponding auditory events may come from arbitrary locations, whilst the sound sources producing them are at fixed positions. Thus virtual sounds in AAR have to be positioned independently of the locations of the available sources. This implies that the sound stemming from those sources has to be altered in a way that it bears the localisation cues of an auditory event at a certain arbitrary position when it reaches the AAR user's eardrums (cf. section 3.1).

A last criterion is interactiveness. Vallino states that

> "The goal of augmented reality systems is to combine the interactive real world with an interactive computer-generated world in such a way that they appear as one environment." [Vallino, 1998]

For a user to perceive Augmented Reality as being natural and immersive, it has to react to user interaction in a similar way the nonaugmented world does. A basic way of interacting with an AAR scenario is to turn the head, thus changing the directions of real and virtual sounds relative to it. A change of the direction of incidence at the ears however comes with changes of the localisation cues contained in both real and virtual sounds. To account for these changes, the processing involved in creating the virtual sounds may need to be adjusted according to the head movement. This calls for the necessity to track the AAR user's head (cf. section 2.5.4).

2.5 Enabling technologies

Virtual sounds in AAR can emanate from any location in the physical space surrounding the listener. An AAR system therefore requires a sound setup capable of reproducing spatial sounds. Spatial sound technology goes beyond normal stereo and surround sound reproduction, which is tied to the horizontal plane, by allowing for virtual sounds at arbitrary positions [Kapralos et al., 2008]. As stated in the previous section, this can be achieved by creating the binaural listening experience that real sources at the same arbitrary positions would cause. 3-D Auralisation thus poses the challenge to the AAR transducer setup of providing the listener with very precisely tailored binaural ear input signals.

2.5.1 3-D audio using loudspeakers

Conventional stereo loudspeaker systems aim at exactly reproducing defined loudspeaker signals. This is not the main goal of AAR loudspeaker setups. AAR systems have to be carefully designed to produce certain sound pressures at the ears of the listeners. Cooper and Bauck use the term "transaural stereo" [Cooper and Bauck, 1989] for a stereo system that considers the ear input signals as the end of the reproduction chain. This involves certain preprocessing of the signals reproduced by the loudspeakers. Gardner names two main steps to reproduce 3-D audio using a pair of loudspeakers [Gardner, 1998]: First, the localisation cues representing a source in space at a certain position have to be encoded into the ear input signals. This process is called "binaural synthesis". To ensure exactly these signals arrive at the listener's ears, in a second step the transmission paths from each loudspeaker to the listener have to be inverted. In particular the crosstalk of the left speaker to the right ear and vice versa has to be accounted for in a process called "crosstalk cancellation".

The invention of the first crosstalk cancellation algorithms dates back to the early '60s. In 1966 Atal and Schroeder patented their "Apparent Sound Source Translator" [Atal and Schroeder, 1966]. The basic principles of the analogue "Atal-Schroeder crosstalk canceller"

still hold for modern digital implementations ([Rao et al., 2006], [Kim et al., 2007b], [Huang and Hsieh, 2007]).

In binaural reproduction, the ear input signals at both ears of the listener have to be controlled. This requires two separate audio channels, one for each ear. In a two-channel loudspeaker system however, the signals of each loudspeaker reach both ears of the listener, reducing the channel separation. The objective of the crosstalk canceller is to separate the left and right channel by eliminating the cross-paths from the left speaker to the right ear and from the right speaker to the left ear. In the frequency domain the ear input signals e_L and e_R can be described in matrix form [Parodi, 2008]

$$\underbrace{\begin{bmatrix} e_L(z) \\ e_R(z) \end{bmatrix}}_{\mathbf{e}(z)} = \underbrace{\begin{bmatrix} H_{L,d}(z) & H_{R,c}(z) \\ H_{L,c}(z) & H_{R,d}(z) \end{bmatrix}}_{\mathbf{H}(z)} \underbrace{\begin{bmatrix} l_L(z) \\ l_R(z) \end{bmatrix}}_{\mathbf{l}(z)}, \tag{2.1}$$

where L and R subscripts denote left and right, c and d subscripts denote cross and direct paths, $H_{i,j}(z)$ are the loudspeaker-to-ear transfer functions and $l_i(z)$ are the loudspeaker signals. The loudspeaker signals are a result of filtering the desired ear input signals $\mathbf{d}(z)$ with the crosstalk canceller matrix $\mathbf{C}(z)$

$$\mathbf{l}(z) = \mathbf{C}(z)\mathbf{d}(z). \tag{2.2}$$

Thus the ear input signals can be rewritten as

$$\mathbf{e}(z) = \mathbf{H}(z)\mathbf{C}(z)\mathbf{d}(z). \tag{2.3}$$

So for the ear input signals $\mathbf{e}(z)$ to equal the desired ear input signals $\mathbf{d}(z)$, the loudspeaker-to-ear transfer function matrix $\mathbf{H}(z)$ has to be inverted by the crosstalk canceller matrix $\mathbf{C}(z)$

$$\mathbf{e}(z) \stackrel{!}{=} \mathbf{d}(z) \longrightarrow \mathbf{C}(z) = \frac{1}{\mathbf{H}(z)}. \tag{2.4}$$

The ear input signals can therefore be controlled by pre-filtering the loudspeaker signals with the inverse of the transfer function matrix $\mathbf{H}(z)$. Problems associated with the calculation of this inverse in the case of singularities in the transfer functions $H_{i,j}(z)$ [Parodi, 2008] are not addressed here.

Reproducing 3-D audio over loudspeakers provides the advantage that the user does not need to wear any equipment. Furthermore loudspeaker systems suffer less from problems associated for instance with headphone reproduction [Zölzer et al., 2002]: listening fatigue and internalisation problems (i.e. an auralised sound source is perceived inside the head, also referred to as inside-the-head locatedness (IHL) [Blauert, 1996], see section 3.3). On the downside, however, the crosstalk cancellation performance is highly sensitive to the user's position and head orientation. The binaural listening experience is limited to a small listening area known as the "sweet spot". If no adaptation is used, the auralisation is degraded if the listener moves away from the centre of the sweet spot by as little as 10 cm or turns the head by more than 10° [Gardner, 1998]. To tackle these problems Gardner suggests the use of a head tracking system to steer the sweet spot to the listener's position and account for head rotations.

The biggest disadvantage of a loudspeaker-based system is the lack of mobility. Furthermore, it is not suitable for multiple simultaneous users. This makes it a rather unattractive option for AAR implementations. Therefore, only user-worn setups are discussed further. An example of a user-worn loudspeaker system is Nortel's "Soundbeam Neckset" [Sawhney, 1998]. Using two directional loudspeakers sitting on the shoulders of the user binaural signals are delivered to the ears. However, the system failed to gain public success.

2.5.2 Bone conduction

Sound transmission to the inner ear via the temporal bone is often considered secondary or negligible [Blauert, 1996]. Griesinger indeed stated that

> "People perceive sound distance and direction only through cues present in the sound pressures at the two eardrums. There are no magic bone conduction or body conduction effects." [Griesinger, 1990]

This assumption is certainly questionable, given for example the advances in the research of bone-anchored hearing aids, which deliver sound to hearing-impaired patients through vibrations transmitted to the skull. In terms of its performance regarding speech intelligibility, the bone-anchored hearing aid is comparable to air conduction hearing aids [Snik et al., 1995]. A recent study on bilaterally fitted bone-anchored hearing aids (in analogy to binaural headsets) by Priwin and others points out a "significant improvement in sound localisation" over the unilateral models [Priwin et al., 2004]. This clearly suggests that bone conductive devices have the potential to transmit spatial audio.

Perhaps the main advantage of using bone conduction for sound transmission is the possibility to place the transducers for instance behind or otherwise close to the ears, thus leaving the ear canals unoccluded. For users with normal hearing, especially those with visual impairments, it may be crucial to retain an unaltered and undiminished auditory perception of the real world. Driven by this assumption, Walker and Lindsay tested the performance of "bonephones", i.e. bone-conduction headphones, in a navigation scenario [Walker and Lindsay, 2005]. Users succeeded in reaching route waypoints by localising spatialised beacon sounds presented to them through vibrations induced to the skull by the bonephones. Walker and Lindsay stated that the usage of bone-related transfer functions (BRTFs, in analogy to HRTFs), might further improve the localisation performance.

Lindeman and others use a bone-conducting headset to enable what they call "Hear-Through Augmented Reality" [Lindeman et al., 2007]. Computer generated audio transmitted via bone conduction augments the real acoustic environment perceived through the unoccluded ear canals, thus the term "hear-through". The user is presented with high-fidelity real world sounds and virtual sounds, mixed together in the cochlea. The sounds transmitted over the headset to the user are private, as they cannot be heard by others. This allows for multiple simultaneous AAR users in the same space through individualised computer generated sounds. In an empirical study, Lindeman and others tested the ability of users of a bone-conducting headset to localise sound sources and classify source movement [Lindeman et al., 2008]. For moving sounds, the bone-conducting headset outperformed standard headphones. In terms of the localisation performance, however, Lindeman and others report a disadvantage over headphones and speaker-based systems. The authors conclude that their use of pre-recorded sounds as stimuli for the headset could be partially responsible for this result. Instead, virtual audio presented over the bone-conducting devices should be processed using bone-related transfer functions to improve the localisation performance. MacDonald and others showed that a user is able to localise sound events using a bone-conductive headset with almost the same accuracy as with standard headphones [MacDonald et al., 2006]. This was achieved by processing the virtual sounds with the individually measured HRTFs of each test participant. The test subjects were thus able to extract localisation cues from the signals despite the crosstalk of the channels, as vibrations from one side of the skull reach the cochleae of both ears. MacDonald and others concluded that the transcranial delay and attenuation, imposed on the sound when travelling from the transducer to the far ear, serve as localisation cues analogously to the ITD and ILD cues. Similar results in terms of the localisation performance were obtained by Stanley and Walker [Stanley and Walker, 2006].

Despite their potential, the usage of bone-conductive transducers for the reproduction of spatial audio is a relatively young area of research. Lindeman and others state that improvements on the frequency range and response of these devices are still necessary to allow for more complex and realistic sounds [Lindeman et al., 2008]. Röber found the perception of selected sound samples to be comparable to normal headphones [Röber, 2009]. Nevertheless the test subjects reported the performance of normal headphones to be higher when listening to speech, music and acoustics.

2.5.3 Headphones

As stated earlier, it is desirable for an AAR setup to be mobile. Thus the acoustic transducers necessary to reproduce virtual sounds need to be mobile too. An obvious choice when thinking of a user-worn audio playback device is a pair of standard headphones. In the context of AAR, however, an additional criterion besides mobility has to be met: the coexistence of real and virtual audio content. An AAR user must be presented with both real world and computer generated sounds simultaneously. As standard headphones occlude the ear canals, they are primarily intended only to reproduce audio, not to preserve the perception of real world audio through them at the same time.

The awareness of one's acoustic surroundings is not only an issue in AAR, but also in everyday life situations. Tappan suggests the usage of "nearphones" [Tappan, 1964], i.e. headphones that do not seal off the ears, to tackle the problem of acoustic insulation of the user. Whilst these "open-back" headphones provide the advantage of allowing a user to perceive environmental sounds through them [Nageno, 2001], they are not intended to leave them as unaltered as possible, which is a requirement for an AAR reproduction system [Röber, 2009, Härmä et al., 2004].

One possibility to achieve precise control over the binaural signals delivered to the user's ears whilst retaining a realistic perception of the acoustic environment is via a technique that Lindeman calls "Mic-Through Augmented Reality" [Lindeman et al., 2008]: Microphones placed near the ears capture the real world sounds. These are mixed with virtual sounds and played back to the user as AAR over a set of standard headphones. Härmä and others describe such a system for the application in "wearable augmented reality audio" (WARA) scenarios [Härmä et al., 2004]. For a detailed description of the system see section 4.1.1.

There are many advantages of using headphones for auralisation. Headphones provide excellent channel separation, which is a key issue in binaural reproduction. Therefore no crosstalk cancellation is necessary [Rao et al., 2006]. Due to their portability, the binaural listening experience using headphones is not limited to a certain place or area. There is no "sweet spot" like in loudspeaker reproduction [Kim and Choi, 2005]. The transmission paths from the transducers to the ears are invertible, ensuring precise control over the ear input signals. In fact Shilling and Shinn-Cunningham state that

> "Spatialized audio using headphones is the only audio technique that is truly 'virtual' since it reproduces azimuth, elevation, and distance and offers the sound engineer the greatest amount of control over the auditory experience of the listener." [Shilling and Cunningham, 2002]

Chapter 3 describes how spatialised audio for AAR is generated over headphones.

2.5.4 Head tracking

The perception of virtual auditory objects in AAR relies on carefully designed and controlled ear input signals. In the case of a real sound source, these ear input signals are dependent on the location of the source with respect to the listener. If orientation and/or position of the

listener change, the position of the source relative to the listener changes accordingly, which causes a change of the ear input signals. For virtual sound sources to preserve the illusion of being overlaid onto the acoustic environment of the listener, a behaviour resembling that of a real source has to be ensured. This means that the ear input signals generated to render a virtual source have to be adjusted if the listener moves the head.

To adapt ear input signals to movements of the listener, the head of the listener has to be tracked. There are various head tracking devices available, with quite substantial differences in terms of technology (mechanical, acoustical, optical tracking), price (from €30 do-it-yourself versions to around €4000 for a professional set of infrared cameras) and performance (various levels of accuracy and update rates). Peltola gives an overview of different devices [Peltola, 2009].

For AAR applications user-worn devices have the important advantage of being mobile. For many applications it is furthermore sufficient to track only the head orientation, not its position. Inertial sensors track orientation changes quite reliably by measuring acceleration and rotation of the device. They are an attractive option being relatively compact, lightweight and wireless. The inertial sensor used in this work for head tracking is the SHAKE (sensing hardware accessory for kinesthetic expression) device [Williamson et al., 2007], described in section 4.1.2.

2.6 Applications of audio augmented reality

Even though AAR has not been the focus of AR research, it's potential has been shown in numerous publications. A brief overview of various use cases and applications related to this work is given in the following sections.

2.6.1 Telecommunication

The perhaps earliest attempts to augment the auditory perception with spatial audio date back to World War I, where the "Pseudophone" apparatus was used to detect enemy aircraft [Wenzel, 1992]. Shinn-Cunningham defines this as the first teleoperator system [Shinn-Cunningham et al., 1997]: The orientation of artificial ears is coupled to the head orientation of the user, presenting the user with binaural signals of the remote environment the artificial ears reside in. The users perceive the remote acoustic environment as though they were physically there. This is called "telepresence". Lehnert and Blauert define it as "a state of mind in which [the user] perceives to exist and act in a different environment than the actual real one" [Lehnert and Blauert, 1991]. Katz and others describe an AR system, which allows a user to remotely drive or supervise an autonomous vehicle [Katz et al., 2007]. The system allows users to perceive and control the remote environment, as if they were present. To monitor the global environment of the vehicle, binaural audio is transmitted to the user.

The "Telehead II" is a remote controlled robot with a dummy head and microphones placed inside the ear entrances [Toshima et al., 2004]. The dummy head is synchronised with the head movement of a remote listener. By listening through the microphones of the robot, the listener can experience the remote sound environment.

The same principle can be applied to telecommunication. Telepresence allows a user to perceive a conversation with another individual as though both participants where in the same physical location. Jain calls this perception of being in a different physical environment "real reality", as opposed to "virtual reality" [Jain, 2000]. The user can seamlessly engage in social interaction in a remote environment. The use of spatial audio enhances the sense of "immersion" [Kapralos et al., 2008].

Besides an improved sense of presence, the use of binaural audio in telecommunication has several other advantages. In everyday listening situations, the human hearing is able to segregate multiple simultaneous sound sources, a phenomenon referred to as the "cocktail party

problem" [Cherry, 1953]. In a conversation with multiple participants, it is thus advantageous to emulate a "cocktail party" situation, by separating the speech sources spatially. Drullman and Bronkhorst report a significant effect of spatial separation of multiple talkers using 3-D audio on communication performance [Drullman and Bronkhorst, 2000].

Whilst spatial separation usually is achieved by placing sources at various azimuth angles, Brungart and Simpson show that intelligibility of competing sources can be enhanced also when sources are placed at different distances in the near field [Brungart and Simpson, 2001]. These findings give rise to a variety of applications in the area of AAR telecommunication.

Previous work on audio augmented reality in telecommunication

Dalenbäck and others describe a teleconference system, where participants are seated around a virtual table [Dalenbäck et al., 1996]. A virtual conference room serves as the meeting place for distant conferees. Its room acoustic properties can be defined to yield a pleasant sound. The seating order defines the position at which each participant is rendered. The distinct and consistent direction of each participant improves the ability to segregate speakers. If head tracking is employed, the perceived positions of other participants remain fixed even in the presence of head movements.

Hindus and others present results of a field study using a telecommunication system called "Thunderwire" [Hindus et al., 1996]. The system relies completely on audio as the communication medium, no visual feedback cues are used. All users of "Thunderwire" are interconnected through a high quality "audio media space". This is a shared virtual acoustic environment, generated as a mixture of the real acoustic environments of all participants in the conversation. The system provides the possibility of group communication and conveying ambient sounds. The reality of each user is thus augmented with a seamless acoustic interface to remote environments and other users.

The "acoustic opening concept" simulates a physical window connecting two rooms [Beracoechea et al., 2008]. The basic idea is to make two virtually connected walls of the rooms acoustically transparent. This is achieved by capturing the sound field in one room and reproducing it in the other. AAR is used to overcome the physical distance and connect both ends of the conversation in a shared augmented reality space.

Lindeman and others report on the use of Second Life, an internet-based application granting access to a virtual 3-D world, to hold the program committee meeting for the IEEE Virtual Reality conference 2009 [Lindeman et al., 2009]. Avatars representing members of the committee were seated in a virtual conference space. The graphical representation was to aid the text- and audio-based communication. Spatialised audio was used to segregate participants of the meeting. The authors conclude that this virtual meeting is a feasible alternative to a face-to-face meeting.

2.6.2 Navigation

One of the most active research areas related to AAR is its use in navigation scenarios. The basic idea is to equip the user with a location-aware device that provides information relevant to the current position, or guidance to reach another position. Bederson presents a prototype for an AAR museum tour guide [Bederson, 1995]. The system basically consists of a portable audio player, a microprocessor and an infrared receiver. When approaching a museum piece to be described, an infrared transmitter mounted above the piece transmits a code to the device, and the microprocessor starts the audio sample describing the piece. This early prototype illustrates one key element of an AAR system: The perception of the real world is enhanced with auditory stimuli that provide relevant information in the given context.

Eckel introduces "LISTEN", a project dealing with the study and development of audio-augmented environments [Eckel, 2001a,Eckel, 2001b]. In the course of the project, Warusfel and

Eckel introduce a platform for exploring a virtual environment overlaid onto the real environment through position-tracked headphones [Warusfel and Eckel, 2004]. The system should allow for "spatial interaction" of the user with virtual content, by triggering sound events through "spatial behaviour". As part of the LISTEN project, Zimmermann and Lorenz present an AAR museum guide [Zimmermann and Lorenz, 2008]. Aim of the system is to provide a personalised audio-augmented environment, tailored to the context of the user. Virtual sound sources are presented over a pair of position-tracked headphones. The virtual sound scape responds to the position and head orientation of the user. An intelligent personalisation process adapts the sound scape according to the user's visit history of the exhibition. The system shows how machine learning may be employed for intelligent AAR.

Lokki and Gröhn report results of a navigation study in an immersive virtual environment [Lokki and Gröhn, 2005]. The authors state that test subjects could navigate through a complex 3-D model without visual stimuli, guided only by auditory cues. A similar study was conducted by Sundareswaran and others [Sundareswaran et al., 2003]. The test system was a prototype for mobile security applications. The authors conducted an AR experiment, in which users where asked to navigate to virtual auditory entities, guided only by acoustic cues. The results of the experiment suggest the potential of audio-only navigation systems.

A navigation system specifically designed for visually impaired individuals is described by Loomis and others [Loomis et al., 1998]. The system provides information about the environment through which blinds user are travelling, allowing them to explore familiar and unfamiliar environments. This system is an example of enriching auditory perception to supplement or substitute the perception of other senses, such as vision.

Dalenbäck and others point out the potential of using auralisation of virtual spaces in architectural acoustics, enabling users to experience simulated rooms and buildings by interacting with the system for example through head movements [Dalenbäck et al., 1996]. Using position and orientation of the user as an input to the system is a basic interface paradigm of many AAR applications. Zotkin and others present algorithms for the creation of virtual acoustic environments that respond to dynamic user interaction [Zotkin et al., 2004]. A head tracking system measures position and orientation of the user. This allows the authors to overlay a virtual sound source onto a physical object, thus augmenting its perception with virtual auditory content.

2.6.3 Virtual auditory displays

Auditory displays are used to display information to a user through the auditory system [Shinn-Cunningham et al., 1997]. Employing the properties of the human spatial hearing allows the creation of virtual auditory displays (VADs) [Shilling and Cunningham, 2002]. A major benefit of using VADs is that they do not constrain the user to turn towards the display. As Shinn-Cunningham points out, many tasks of human operators require responding to spatial information [Shinn-Cunningham, 1998]. This information could be presented via VADs, to supplement the often overloaded human vision. Various applications of VADs have been studied.

"Nomadic radio" is a system granting users access to information and communication services through an AAR interface [Sawhney and Schmandt, 2000]. The delivery of information is filtered based on the content of the message and the context of the user. This is a fundamental difference to AAR navigation systems discussed above, where the overlay of virtual content is tied to the physical location of the user.

A similar concept to the "Nomadic Radio" is the basis for "Audio Aura", a system to provide users with serendipitous information [Mynatt et al., 1998]. Interaction with the computer is made "implicit", by interpreting physical actions in the real world to trigger the delivery of background information through audio. The authors describe the information as "useful but not required". As users do not have to rely on it, the audio information overlaid onto the auditory perception does not have to be invasive, i.e. it can remain in the background. This

nonintrusive augmentation provides the possibility of truly seamless informational enhancement of the perception of the real world.

An AR system for device diagnostics and maintenance is introduced by Behringer and others [Behringer et al., 1999]. 3-D audio cues are used to indicate objects outside the field of view of the user. This employs an advantage of the auditory over the visual perception: It can process sensory cues from all directions, regardless of the orientation of the listener.

In an experiment to study the perception of self-motion, Väljamäe and others used auditory cues to create the illusion of circular vection in a virtual environment [Väljamäe et al., 2009]. This implies that VADs can be used to alter the way interaction with the physical world is perceived.

"Diary in the sky" is a mobile AAR calendar application [Walker et al., 2001]. Instead of putting calendar event entries on a visual display, they are presented to the user as spatialised audio. This provides the possibility of mapping event parameters to sound attributes. The event time could for instance be mapped to the azimuth of the auditory event, making effective use of the ear's omni-directionality.

2.6.4 Entertainment

VADs are a potential alternative to costly multi-loudspeaker home entertainment systems [Shinn-Cunningham, 1998]. Social networking is another application area for AAR. Rozier and others describe "Hear&There", an AAR system allowing users to leave "audio imprints" at outdoor locations [Rozier et al., 2000]. The idea of audio-augmenting a space is taken from audio guides offered in museums. In "Hear&There", the audio information is created by other users of the system and linked to a specific location.

A different form of audio tagging is presented with the "AudioMemo" application [Peltola, 2009]. It provides the possibility to store binaural recordings along with the recording position and orientation of the user, who can afterwards browse through the recordings and take an acoustic walk through the recorded path.

Lyons and others present "Guided by Voices", an AAR game [Lyons et al., 2000]. Players interact with the game by walking around the real world, to collect virtual objects and meet virtual game characters. The game is set in an audio-only environment. The authors argue that despite the absence of visual feedback, setting the game as an overlay to the real world makes it immersive. In multi-player mode many users occupy the same physical location, and thus engage in active social interaction, unlike PC video games. The game blurs the boundaries between the physical world, the virtual game environment and the imagination of the player.

Chapter 3

Headphones reproduction

3.1 Auralisation using headphones

In the previous chapter the advantages of headphones reproduction in AAR over loudspeaker-based setups are discussed. The present section describes the signal processing involved in the creation of AAR content for headphones. It can be subdivided into three steps:

- Generation of spatialised audio ("binaural synthesis"),
- equalisation (using a "binaural reproduction filter" [Kim and Choi, 2005]) and
- mixing of real and virtual audio content.

Whilst the binaural synthesis is, in theory, independent of the transducer technology, the equalisation and mixing of the audio material as described here is specific to headphone reproduction.

3.2 Binaural synthesis

The "raw material" for generating virtual auditory events for AAR are either synthesised sounds or recordings of real sounds. The process of "spatialising" these sounds is referred to as "binaural synthesis" [Chanda et al., 2006], and defined by Jot and others as:

> "Binaural synthesis is a process by which, from a primary monophonic recording of a source signal, a three-dimensional sound image can be reproduced on headphones." [Jot et al., 1995]

The goal of binaural synthesis is to enhance an input signal with the localisation cues of a virtual source in space. If the resulting binaural signals are presented to listeners as ear input signals, they (ideally) perceives the input signal as emanating from the position of the virtual source. Binaural synthesis could thus be described as the processing toolbox for auralisation.

The first step towards creating binaural signals from a monaural sound is to provide interaural cues, i.e. ITD and ILD. The level and delay of one channel with respect to the other is adjusted according to the desired azimuth and elevation of the virtual source. Though these cues can be applied separately, using both improves the spatial impression [Zölzer et al., 2002]. By approximating the two ears as points in free space, the path difference Δs to each point is given by a simple law [Blauert, 1996]

$$\Delta s = d \sin \varphi, \tag{3.1}$$

with d =21 cm (i.e. the distance of the two points), and φ being the angle of incidence of the plane wave. In this "sine law", derived by Hornbostel and Wertheimer in 1920, the parameter

d does not correspond to the actual distance between the ears (cf. fig. 4.3). Also the shadowing effect of the head is ignored. Whilst this formula is applicable for simple source panning in the horizontal plane, for binaural synthesis a more comprehensive approach is needed.

To generate convincing interaural cues, the influence of the head on the wave propagation has to be modelled. For this purpose, the head can be approximated by a rigid sphere of similar dimensions than the head [Blauert, 1996]. Calculating the sound field on the surface of this sphere when exposed to a sound source allows to estimate the time and intensity differences between two points on the sphere representing the ears. These differences are frequency dependent [Rocchesso, 2002]. The ILD for sound incidence from a certain direction increases with frequency, as shorter wavelengths are stronger attenuated by the head (see section 2.3.1). A reciprocal effect holds for the ITD. For low frequencies the ITD is higher than for high frequencies, due to the increased path length caused by diffraction of low frequency components around the head [Rocchesso, 2002].

It has been argued that the effect of this diffraction is less apparent in ITDs measured on test subjects, and that ITD variations across the frequency range are perceptually irrelevant [Wightman and Kistler, 1997]. Duda and others however point out ITD variations between different test subjects [Duda et al., 1999]. The simple spherical model does not account for ITD changes around a cone of confusion (see section 2.3.1). Duda and others therefore argue for the use of an adaptable ellipsoidal head model to calculate individual and more accurate ITD values. An overview of different methods to determine ITD values for binaural synthesis is presented by Minnaar and others [Minnaar et al., 2000].

Besides ITD values, accurate ILD cues are necessary for spatialisation. Simple geometric head models provide only a very rough estimation of ILD values, as they neglect the influence of the pinnae. By filtering the sound reaching the ear canals, the pinna encodes direction- and distance-dependent spatial cues into the ear input signals. These cues are of major importance for human sound localisation [Blauert, 1996]. Therefore,for spatialisation it is necessary to imitate the pinnae's filtering behaviour. The pinna affects the incoming sound in many ways, through reflections, shadowing, diffraction and resonance [Blauert, 1996]. An overview of various approaches to model the localisation cues generated by the pinnae is presented by Satarzadeh and others [Satarzadeh et al., 2007]. Such models can become quite complex, and in combination with models for the head shadowing and the influence of the torso rather cumbersome.

Another approach for generating authentic spatial cues is to rely on measurement data. As described in section 2.3.3, HRTFs contain the localisation cues of a sound source at a given azimuth and elevation angle. Applying appropriate HRTFs for left and right ear to a monaural signal encodes exactly those cues into the signal, thus evoking the impression that the sound is emanating from the desired direction. HRTFs, or their time-domain equivalent, the HRIRs, corresponding to the desired virtual source direction can be applied to the signal in the form of two finite impulse response filters. Convolving the input signal with the HRIRs of the left and right ear directly results in a binaural signal enhanced with the localisation cues of the source recorded during the HRIR measurements.

3.3 Externalisation

The convolution of a monaural signal with a pair of HRIRs should provide sufficient localisation cues to correctly identify the position of a virtual sound source. Yet the complexity of the human hearing and its sensitivity to nuances in the HRTFs poses a challenge to this approach. Therefore, the reproduction of signals over headphones may suffer from problems like front–back reversals and IHL.

3.3.1 Inside-the-head locatedness and front–back reversals

A common problem in binaural reproduction is the internalisation of sounds, or inside-the-head locatedness (IHL), as Blauert calls it [Blauert, 1996]. Blauert defines IHL in terms of the perceived distance of a virtual source [Blauert, 1996]. If the perceived distance is smaller than the radius of the head, the source is perceived inside the head. This is an undesirable effect, especially in the context of AAR, where virtual sources are overlaid onto the surroundings of the user. IHL is often associated with headphone reproduction of binaural signals. As Rocchesso points out, however, also loudspeaker reproduction can cause IHL [Rocchesso, 2002]. He states that:

> "It seems that human subjects tend to internalize the perceived objects when the total stimulation, as coming from all sensorial modes, cannot be produced by natural situations involving distant sources" [Zölzer et al., 2002].

The opposite of an internalised source is an externalised source. In reference to Blauert's definition of IHL (see above), externalisation could thus be defined in terms of the perceived distance. As Moore and others remark, however, distance is not a reliable externalisation measure due to the variability of distance estimates [Moore et al., 2007]. Instead, the authors propose to use a definition by Hartmann and Wittenberg to define and measure externalisation [Hartmann and Wittenberg, 1996]: A virtual source is externalised and localised if it is indistinguishable from a real-world source. To check whether this holds, Härmä and others suggest the use of a modified "Turing test" [Härmä et al., 2003] (cf. section 4.1.1). A major goal of AAR systems is to create virtual sources that are correctly localised in the surrounding space and thus externalised.

Another problem often observed in binaural reproduction is a front–back reversal in the perceived position of a virtual sound source [Zölzer et al., 2002]. Wenzel and others state that the most common confusion is that of a source in the front hemisphere of the listener being judged as residing in the rear [Wenzel et al., 1993]. They conclude that a possible explanation for this lies in the fact that ITD and ILD values are roughly constant around a cone of confusion, and thus ambiguous. The authors further assume that in absence of a visual stimulus supporting the virtual auditory event, this ambiguity is solved by judging the source as being in the rear, i.e. outside the field of view. When listening to real sources, the cues to disambiguate the cones of confusion are contained in the HRTFs, and any impairment of them increases the rate of front–back reversals [Wenzel et al., 1993].

3.3.2 Individual(-ised) head-related transfer functions

HRTFs are highly individual. In binaural synthesis, for best localisation performance, the HRTFs of the listener herself should be applied. Using the HRTFs of another test subject generally deteriorates the spatial perception, and may lead to the aforementioned problems of IHL and front–back reversals. Møller and others found the use of nonindividual binaural recordings (i.e. recordings made at the ears of another test subject) to cause a deterioration of the localisation performance in the horizontal plane, as well as front–back reversals [Møller et al., 1996]. With the use of individual recordings, however, performance was comparable to the real listening situation. Wenzel and others showed that the rate of front–back reversals for virtual sources processed with nonindividual HRTFs may be the quadruple of the free-field rate, using real sound sources, whereas sources processed with the test subjects' own HRTFs resulted in doubling of the rates [Wenzel et al., 1993]. In addition to front–back reversals, up–down confusion was observed.

These findings clearly suggest that using individual HRTFs enhances binaural synthesis. In most cases, however, measuring an individual set of HRTFs is not feasible, due to the time and equipment necessary for the measurement. To overcome the problems involved with using

a generic HRTF set of a "random" person, HRTFs from dummy heads representing a "mean" anthropometry, are often employed instead in binaural synthesis [Cohen et al., 1993, Kim et al., 2005, Hirahara et al., 2007, Beracoechea et al., 2008] (see section 2.3.3). However, this method may still lead to a considerable deterioration of the localisation performance, as Møller and others have shown [Møller et al., 1999]. In their study, the localisation performance of human subjects exposed to a real sound field and to the sound field as recorded by a dummy head was compared for eight different dummy heads. Comparing the results to an earlier study involving only human subjects [Møller et al., 1996], Møller and others found that the localisation performance of subjects listening to a dummy head recording was equal or worse than with a recording made on another human test subject. This indicates that in terms of the localisation performance, the use of a dummy head provides no advantage over the use of a random human subject. Møller and others point out, however, that the localisation performance improved when an appropriate human subject was chosen instead of a random subject. Recent research on the development of artificial heads is concerned with improving dummy heads to better match the human anthropometry. An example of these efforts is the development of the head and torso simulator VALDEMAR [Christensen et al., 2000]. Fels and Vorländer propose the adaptation of dummy heads to approximate the filtering characteristics of children [Fels and Vorländer, 2004]. Fastl suggests the introduction of a standardised dummy head shell for all artificial heads, and gained support for his idea from both dummy head manufacturers and users [Fastl, 2004].

It remains to be questioned whether a "golden model" for artificial heads will be found, given the discrepancies between human physiques. A compromise between feasibility and accuracy of an HRTF dataset for a listener is the individualisation of HRTFs. In this process, anthropometric measures of the listener are considered to adapt an HRTF model or a generic HRTF dataset accordingly. Personalised HRTFs allow for an undistorted perception and good localisation of virtual sources [Zotkin et al., 2004]. Genuit patented a method to derive a physical transfer function model from 34 anthropometric measures of a human test subject's head, shoulders and pinnae [Genuit, 1987]. The influence of each body part is modelled by tuning filters, resonators, adders and time-delay elements accordingly. Based on Genuit's work, Sottek and Genuit describe a physical model adaptable to individual variations in the HRTFs by subdividing the body into very simplistic geometric models [Sottek and Genuit, 1999]. The models for head, shoulders, pinnae and cavum conchae are adjustable in height, width and position according to the corresponding anthropometric measures of the test subject to be modelled. A similar approach was taken by Algazi and others, who propose the use of anthropometric measures to derive composition rules to combine models of the contribution of the head, torso and pinnae to a complete HRTF model [Algazi et al., 2001a]. Kim and others present a time-domain based modelling approach [Kim et al., 2007a]. The authors state that their model would be suitable also for parametric individualisation. Satarzadeh and others show how the HRTF for isolated pinnae, called pinna-related transfer function (PRTF), can be modelled and parametrised using pinna measures of a human test subject [Satarzadeh et al., 2007]. The authors point out, however, that the suggested method is not applicable to every pinna in general, and that further research is necessary to be able to derive a complete model from anthropometric measures.

As stated by Møller and others, improved localisation performance can be attained by choosing the best match from a set of measured HRTFs [Møller et al., 1999]. As shown by Zotkin and others, the process of choosing a match can be automatised by evaluating an image of the listener's pinna [Zotkin et al., 2003]. From the image, anthropometric measures are estimated and the best-matching HRTF is chosen from a dataset based on these measures. The authors further explain how a generic HRTF can be individualised by combining the measured HRTF with a personalised head-and-torso model. The head-and-torso model can be adjusted using three body measurements: the torso radius, the head radius and the neck length. The authors suggest the use of such a head-and-torso model in research and software development for improved localisation and subjective quality. The personalisation however occasionally results in

a deterioration of the performance, which indicates that the proper choice of a matching HRTF from a dataset is still an ongoing research problem [Satarzadeh et al., 2007].

3.3.3 Reflections and reverberation

Individualisation of HRTFs improves the localisation performance and ensures an undistorted perception of a virtual source, but may still lead to IHL. One potential defect lies in the nature of HRTFs. HRTFs capture the filtering behaviour introduced by the listener to a sound field. The influence of the environment and/or room is not included in them, as they are anechoic by definition. When sound propagates in a real acoustic environment, it is affected by surfaces and objects in a similar way it is affected by the presence of a listener (see section 2.3.3). Zotkin and others state that processing a monaural anechoic sound with anechoic HRTFs results in a virtual source that may appear inside or very close to the listener's head [Zotkin et al., 2004]. It is known that reverberation added to a virtual sound may improve the perceived externalisation [Begault, 1992, Zölzer et al., 2002, Shinn-Cunningham et al., 2005].

This however implies a considerable increase in terms of the complexity and the computing resources required for rendering virtual audio. To include the room influence in the creation of virtual sounds, the virtual room and all acoustically relevant surfaces and their influence on the simulated sound field of the virtual source have to be modelled. The model is used to predict the direction, amplitude and spectral shape of reflections reaching the listener's ears. Two common methods two simulate the sound field in a virtual room are ray tracing and mirror imaging, a brief introduction to which is given for example by Blauert [Blauert, 1996], Savioja [Savioja et al., 1999] or Rocchesso [Rocchesso, 2002]. This yields the simulated impulse responses of the virtual room, for sound travelling from the virtual source to the ears of the listener.

The second step in adding environmental cues to the virtual sound is to filter each reflection with an appropriate HRTF, to correctly reproduce its direction of arrival to the listener. By processing the virtual indirect sound field in this way a "spatial reverberation" [Begault, 1992] is obtained. The result is a pair of impulse responses (one for each ear), the binaural room impulse response (BRIR). It comprising the influence of the room as well as the filtering behaviour of the listener. The BRIR can be applied to a virtual input signal the same way as an anechoic HRIR: convolution encodes the spatial cues contained in the impulse response into the input signal. The fact that the influence of a virtual room is contained in the BRIR, however, puts some limitations as to its applicability. The BRIR obtained in the described way is only valid for a certain position of both source and listener in the modelled room. Sophisticated binaural room simulations also take factors like source directivity into account [Blauert, 1996]. Thus, a comprehensive set of measured BRIRs is not feasible, although measurements have been made for partially static scenarios, for example static listener orientation with variable source and listener positions [Shinn-Cunningham et al., 2005], static source and listener positions with variable head orientation [Lindau et al., 2008], and static listener position and orientation with variable source positions [Rychtáriková et al., 2009]. Therefore, to allow for more flexibility, binaural synthesis has to rely on a modelling approach to obtain a comprehensive set of BRIRs.

3.3.4 Head movements

Motional cues are an essential part of the real world listening experience. They occur when the direction of a sound source changes with respect to the orientation of the head of a listener (cf. section 2.3.6). In normal headphone reproduction, no motional cues are present. The directions of virtual sources relative to the orientation of the head remain fixed. If the listener turns the head, the virtual scene rotates accordingly. Zotkin and others argue that this prevents externalisation: A virtual source immune to head movements is instinctively placed at the origin of the moving coordinate system, i.e. inside the head [Zotkin et al., 2004].

By accounting for head movements and thus creating motional cues, in combination with correct ITDs and reverberation, Zotkin and others report "very good externalisation" of sound rendered through headphones. Also Rocchesso states that these dynamic cues improve externalisation in binaural synthesis via headphones [Rocchesso, 2002]. Furthermore, similar to real world listening, head movements improve the localisation performance in binaural synthesis [Minnaar et al., 2001] and resolve the problem of front–back reversals.

For a headphone-based AAR system to respond to head movements and create dynamic cues, the position and orientation of the user's head has to be tracked. The dynamic cues are generated by updating the virtual scene and audio according to the head tracking data.

3.3.5 Visual and other cues

Auditory perception is to some extent multisensory. The perception of an acoustic entity can be influenced by nonauditory cues stemming for instance from the sense of vision.

In binaural synthesis, the absence of such visual cues makes the perception of a virtual sound source in front of the listener difficult [Wenzel et al., 1993]. The human brain generally assumes an invisible source to be in the back or places it inside the head. Conflicting cues are likely to be resolved in favour of the visual domain, which is related to the phenomenon of "visual capture", discussed in section 2.3.7. Consider a virtual sound source rendered at a certain position in space. If visual cues are present indicating a different position of the source, the sound is likely perceived to be emanating from that position rather than the position indicated by the auditory cues.

If coherent with auditory cues, visual cues offer a way to enhance and reinforce auditory perception. In fact the visual capture can be used to create a more realistic perception of a virtual auditory space [Yost, 1993]. Moore and others point out that visual cues could facilitate the creation of externalised virtual sound sources even if deficient auditory cues were used [Moore et al., 2007]. Zotkin and others created externalised virtual sources by providing the listeners of their experiment with a visual point of reference in the form of a small physical cube [Zotkin et al., 2004]. The sound source was rendered at the position of this cube, and listeners were successfully made to believe that the sound was emanating from the cube. The authors however did not report whether the illusion of the externalised source could still be achieved in absence of this physical object.

3.4 Equalisation

By enhancing binaural signals with the features discussed above, externalisation can be achieved or improved. This however implies that the generated signals and all encoded cues reach the ear canals as unaltered as possible. Ideally, the ear input signals are equal to the binaural signals generated in the binaural synthesis. Auralisation in AAR heavily depends on precise control of the ear input signals. When playing back binaural signals, the influence of the reproduction chain on the ear input signals has to be eliminated. This requires a flat frequency response of the transducers and the playback device. A flat frequency response of the reproduction chain can be ensured by applying a correction filter to the binaural signals, thus inverting the transmission paths of the binaural signals to the ear entrances. Zahorik and others show the importance of accurate compensation for the impulse responses of headphones in binaural synthesis [Zahorik et al., 1995]. By using an appropriate equalisation filter, the authors were able to create virtual sound sources that are indistinguishable from real sound sources. Shortening the impulse response of the equalisation filter through windowing resulted in a deterioration of the performance, and increased discriminability of the virtual sources from the real sources. Kim and Choi point out, that the equalisation filter response varies among individuals [Kim and Choi, 2005]. Therefore, to achieve externalisation of virtual sound sources, individual equalisation

filters should be applied to the binaural signals. It should be pointed out that obtaining the individual transmission paths of binaural signals to the ear drums is a difficult process [Griesinger, 1990]. To measure the pressure at the ear drum, a probe microphone has to be inserted into the ear canals, a task that should preferably be assigned to an otologist.

A technique proposed by Hiipakka and others overcomes these problems by estimating the pressure at the ear drum from a measurement at the canal entrance [Hiipakka et al., 2009]. The authors use a pressure–velocity probe to measure the sound pressure and volume velocity at the canal entrance. By combining the two measurements, the pressure at the eardrum and the transmission path to the ear drum is computed. From the transmission path, individual equalisation filters can be derived. They ensure unaltered perception of binaural signals and thus improve the performance of the binaural synthesis.

3.5 Mixing

After carefully designing and synthesising the binaural signals and the reproduction system, the virtual sounds have to be overlaid onto the real acoustic environment. The perception of both real and virtual environment results in AAR. Therefore, ear input signals have to be generated that contain both virtual and real world sounds. The sections above focus on ways to achieve and preserve correct localisation cues in binaural signals when delivered to the ear entrances. In AAR, however, unaltered perception of the real world is as important as the perception of the virtual world. This means that the impact of the AAR system on the perception of the real world has to be eliminated. The transducer system used to reproduce virtual sounds should thus be acoustically transparent. If loudspeakers or bone-conducting devices are used for reproduction, the ear canals of the listener remain unobstructed. This guarantees unaltered auditory perception of the real world. In the case of headphones usage, the ear canals of the listener are blocked. Sounds from outside reaching the ear entrances are affected by the presence of the headphones. In this work, a special headset is used, the KAMARA headset. It is designed to minimise the influence of the transducer system on the perception of the real world whilst retaining precise control over the ear input signals (cf. section 4.1.1).

If the localisation cues of both the real world sounds and the virtual sounds are preserved, overlaying the virtual sounds onto the real acoustic environment is achieved by simply adding or mixing the real and virtual signals. The listener, presented with a mix of unaltered real world sounds and binaurally synthesised virtual sounds, perceives the virtual sound sources as embedded into the real acoustic environment. The virtual sources thus appear to be emanating from defined locations in the real world environment, which is the basis of AAR.

Chapter 4

Experimental setup

4.1 Hardware and software platform

4.1.1 KAMARA headset

When using a headset as the transducer system for AAR, acoustical transparency of the headset has to be ensured. This can be achieved by capturing the real-world sounds at the ears and playing them back through the earphones. Mixing these captured real-world sounds with virtual sounds is the basic working principle of "mic-through augmented reality" [Lindeman et al., 2008], which refers to the fact that the real world is perceived through microphones.

The KAMARA headset, introduced by Härmä and others [Härmä et al., 2004], is an implementation of mic-through augmented reality (see fig. 4.1). It is used in this work as the acoustical transducers for AAR. The KAMARA headset consists of a pair of insert-earphones with integrated miniature microphones. Sounds captured by the microphones can be played back directly to the earphones plugged into the user's ear canals. The usage of insert-earphones provides the advantage of leaving the pinnae of the listener uncovered. Capturing the real-world sound close to the ear entrance preserves the filtering behaviour of the pinnae. Pinna cues are important for the localisation of real-world sounds. Inserting the earphones into the ear canal minimises effects of the transmission paths from the earphone to the ear drum. This simplifies the task of equalisation, as the equalisation filters do not have to account for pinna reflections.

Equalisation of the KAMARA headset

As the perception of the real world environment through the KAMARA headset is slightly altered, Härmä and others define it as a "pseudo-acoustic environment". It is augmented by overlaying virtual sounds onto the pseudo-acoustic environment in an "augmentation mixer". The problem of colouration of the pseudo-acoustic environment by the reproduction chain can be addressed by integrating appropriate equalisation filters into the augmentation mixer. The equalisation filters used in this work are described by Riikonen and others [Riikonen et al., 2008].

Mixing pseudo-acoustic environment and virtual sounds

The augmentation mixer mixes real (i.e. pseudo-acoustic) sounds recorded with the microphones and virtual sounds stemming from an external input source. This external source consists for instance of a computer generating the virtual audio content to be overlaid onto the pseudo-acoustic environment. The mixing is achieved by applying appropriate gains and summing the signals. To analyse the performance of the augmentation process, Härmä and others performed an adapted version of a *Turing test* [Härmä et al., 2003]: The test determined whether listeners were able to distinguish between sounds from the pseudo-acoustic and the virtual environment.

Figure 4.1: *KAMARA headset.* Microphones embedded into the headset record audio signals close to the ear canal entrances. Through the microphones a "pseudo-acoustic environment" is perceived [Härmä et al., 2003].

The authors report that test subjects could correctly distinguish pseudo-acoustic and virtual sound in 68 percent of the test cases. For speech signals the rate was even lower, approaching chance level. This indicates that carefully designed virtual content integrated into a pseudo-acoustic environment makes virtual sources nearly indistinguishable from real ones.

4.1.2 SHAKE head tracking device

To track the head of the listener, the SHAKE device is used [Williamson et al., 2007]. The SHAKE is equipped with triple axis accelerometers, gyroscopes and magnetometers. From the sensor values, the 3-D orientation of the device is determined. Small size, Bluetooth connectivity and an internal battery make the device portable and thus suitable for AAR applications. The SHAKE is an inertial sensor, thus no external references, except earth's gravity and polarity, are necessary for tracking [IEEE, 2001]. The sensors of the SHAKE device are positioned such as to provide sensor data in all three dimensions of space relative to the axes of the device (see figure 4.2).

The gyroscopes measure rotation about, the accelerometers acceleration and the magnetometers magnetic force along each axis. Besides the raw sensor data, the SHAKE SK6 used in this work provides a heading angle calculated internally. The heading angle, which resembles a compass heading derived mainly from the magnetometer values, proved to be too noisy and unreliable in the proximity of metallic devices for tracking the head orientation. Therefore, a C library was written for Pure Data (Pd) that calculates the 3-D orientation from the raw sensor values.

Calculating the 3-D orientation of the SHAKE device

The SHAKE orientation is given as the orientation of each axis in a global Cartesian coordinate system. The orientation of the global coordinate system is defined as the orientation of the SHAKE device during initialisation (i.e. before any motion), with the z-axis parallel to the (estimated) gravity vector and the x-axis pointing forward. The orientation of the device at any time is given relative to this initial orientation. It can be described by a 3x3 orientation matrix A consisting of the vectors representing the SHAKE axes

$$A = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{bmatrix}, \tag{4.1}$$

with x_i, y_i and z_i indicating the i-th element of the axes x, y and z (cf. fig. 4.2). During initialisation, the orientation matrix is reset to the identity matrix. From the gyroscope data, the relative rotation of the device about each axis can be derived at each sampling instant.

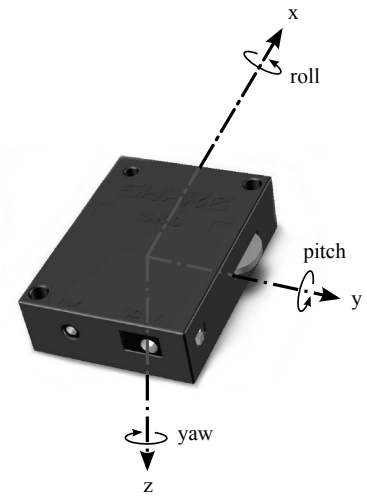


Figure 4.2: *SHAKE device.* Gyroscopes, accelerometers and magnetometers measure rotation (i.e. pitch, roll and yaw) about, and acceleration and magnetic force along the three axes, x, y and z, of the device.

Multiplying the rotation g measured about one axis with the sampling period T of the gyroscope yields an estimate for the rotation angle α about this axis. As an example, a rotation about the x-axis of the SHAKE device is calculated (k is the current sampling instant)

$$\alpha_x[k] = T_x \cdot g_x[k]. \tag{4.2}$$

Once the rotation is determined, the orientation matrix $A[k]$ is updated accordingly. Following the above example, the SHAKE axes have to be rotated about the x-axis of the SHAKE by the angle α_x. In matrix notation, this is given by

$$A[k] = Q_x \cdot A[k-1], \tag{4.3}$$

where Q_x is the rotation matrix describing the desired rotation of A about the x-axis. The rotation matrix Q_x is calculated by deriving the quaternions describing the rotation [Kuipers, 2002]. Given a (normalised) rotation axis

$$r = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \tag{4.4}$$

and a rotation angle α about this axis, the quaternions are calculated as

$$\begin{aligned} q_0 &= \cos(\alpha/2) \\ q_1 &= a \cdot \sin(\alpha/2) \\ q_2 &= b \cdot \sin(\alpha/2) \\ q_3 &= c \cdot \sin(\alpha/2). \end{aligned} \tag{4.5}$$

From these quaternions the rotation matrix Q is derived as

$$Q = \begin{bmatrix} 1 - 2 \cdot (q_2^2 + q_3^2) & 2 \cdot (q_1 q_2 - q_0 q_3) & 2 \cdot (q_1 q_3 + q_0 q_2) \\ 2 \cdot (q_2 q_1 + q_0 q_3) & 1 - 2 \cdot (q_1^2 + q_3^2) & 2 \cdot (q_2 q_3 - q_0 q_1) \\ 2 \cdot (q_3 q_1 - q_0 q_2) & 2 \cdot (q_3 q_2 + q_0 q_1) & 1 - 2 \cdot (q_1^2 + q_2^2) \end{bmatrix}. \tag{4.6}$$

The orientation of the SHAKE is updated at every sampling instant of the gyroscopes by deriving the rotation matrix Q for each axis and multiplying it with the orientation matrix A. This

provides a quite accurate way to track relative head motion. Integrating the noisy sensor values to estimate the absolute orientation, however, inevitably leads to an orientation drift. The earth's polarity and gravity are used as external references to compensate for the drift and continuously recalibrate the calculated orientation. To remove sensor bias, a moving average is subtracted from the raw gyroscope and accelerometer values.

4.1.3 Pure Data programming environment

Pure Data (Pd) is a graphical programming environment [Puckette, 1996]. It serves as the main data and control interface in the experimental implementation. Internal functions of Pd are used in combination with external libraries written in C/C++ by members of the Pd community and myself to perform the following tasks:

- audio in-/output,
- communication with and data acquisition from the SHAKE device and
- audio and data processing and logging.

The main reasons for the choice of Pd in this work lie in the possibility to perform real-time signal processing with relatively little software overhead, and the large pool of easily accessible communication ports and protocols. Direct control over the audio hardware of the computer running Pd allows real-time audio input to and output from the system via the computer's audio I/O. This minimises the processing delay. Communication with the SHAKE device is achieved by establishing a Bluetooth connection to a virtual comport. The data retrieved over the serial port connection is processed in the C external described in the previous section to obtain the 3-D orientation of the SHAKE device, which controls the audio processing. Data logging facilitates debugging and analysis of the system.

4.2 Implementation

4.2.1 Introduction to the KAMARA 2009 project

The present work is part of the KAMARA (killer applications for mobile augmented reality audio) 2009 project, a cooperation between the Nokia Research Center Helsinki, the Laboratory of Acoustics and Audio Signal Processing and the Department of Media Technology of the Helsinki University of Technology. The goal of the project is to study various aspects and issues concerning the usage of the KAMARA headset in a telecommunication scenario. The study is subdivided into two tasks. Task I concentrates on the analysis of the binaural audio recorded by a KAMARA user. Task II, i.e. the present work, deals with the problems involved in presenting this binaural audio to a another KAMARA user over a telecommunication system. A usage scenario is presented to demonstrate the results of both tasks.

4.2.2 Usage scenario

The usage scenario assumes one-way telecommunication between a remote and the local end. Data and audio from the remote end are transmitted to the local end via VoIP or similar technology. A KAMARA headset user on the remote end (hereafter referred to as the "remote user") participates in a meeting with multiple participants, all located in the same room as the user (the "remote room"). The microphones of the KAMARA headset record the audio of the meeting. The head orientation of the remote user is tracked with the SHAKE device. Task I focusses on the analysis of this remote end audio. Part of this analysis is to determine which talker is speaking when, and from which direction.

At the local end, a KAMARA user (the "local user") is presented with the audio recorded at the remote end of the simulated teleconference. The present work is concerned with task II of the KAMARA 2009 project. The playback of the binaural recording from the remote end to the local KAMARA user is studied. The goal is to determine the audio processing necessary to enhance listening comfort and speech intelligibility. For this the information gathered in task I about the remote end talker turns and directions is necessary. The demonstration implementation merges the results of tasks I and II.

In the described telecommunication scenario, multiple remote talkers are presented to the local user over a VoIP connection. This gives rise to the "cocktail party problem", described in section 2.3.8: The listener has to segregate various talkers and sound sources to be able to follow the conversation. Spatial cues play an important role in this segregation task. By capturing the sound at the remote end with the KAMARA headset, and playing it back to the local listener with a similar headset, these cues are preserved. The local user is thus able to segregate various remote talkers and sound sources based on the spatial cues contained in the binaural VoIP audio.

Ideally, both the remote and the local user keep their heads still during the conversation. In this case the perceived direction of each source at the local end corresponds to the actual direction with respect to the remote user. The local user can rely on spatial cues to map sounds to sources and hence segregate them. If the remote user rotates the head, however, the relative direction and therefore the interaural cues of each source change accordingly. As a result, whenever the remote user rotates the head, the remote sources are perceived at the local end as changing their positions. Moving sources might deteriorate both the listening comfort and the speaker segregation at the local end, due to the lack of reliable interaural cues. One aim of this work is therefore to preserve these cues and the benefits of using binaural audio in telecommunication. This "de-panning" process is described in section 4.2.3.

Whilst head rotation at the remote end creates the illusion of moving sources, head rotation at the local end is perceived as each source having a fixed direction relative to the local user. Thus, if the local user changes the head orientation, the audio scene rotates accordingly. In a telecommunication scenario, however, it might be desirable to register the sound with the environment of the user, as in AR. This allows the user for example to turn the head to look at a remote talker, which is a natural behaviour in face-to-face communication. The second aim of this work is therefore to register sounds from binaural VoIP with the environment of the local user. This "panning" process is described in section 4.2.4.

As an additional feature, the local user could be given the possibility to define the perceived position of the remote sound sources. A very simple approach to position virtual sound sources using finger snaps or claps is presented in section 4.2.6.

4.2.3 De-panning of binaural audio

Head movements during a binaural recording via the KAMARA headset alter the interaural cues of the recorded sound source. When listening to the recording, the source appears to be rotating around the listener. "De-panning" is the process of compensating for the head movement. The resulting de-panned recording contains interaural cues resembling a recording scenario without head movement. A nonmoving source will thus be perceived as being nonmoving even in the presence of head movement during the recording (cf. fig. 4.6b). In the previously described usage scenario, this implies that remote participants of a teleconference, recorded via a KAMARA headset worn by one of the participants, will always be perceived by the local participant at their position relative to the remote participant, irrespective of the head orientation.

Two measures need to be known for the de-panning process: The head orientation of the remote user and the position of the sources. In the demonstration implementation, the head orientation is tracked with the SHAKE device. If the position of the sound sources is not fixed or known, it can be determined from the recorded audio. Speaker direction estimation is covered

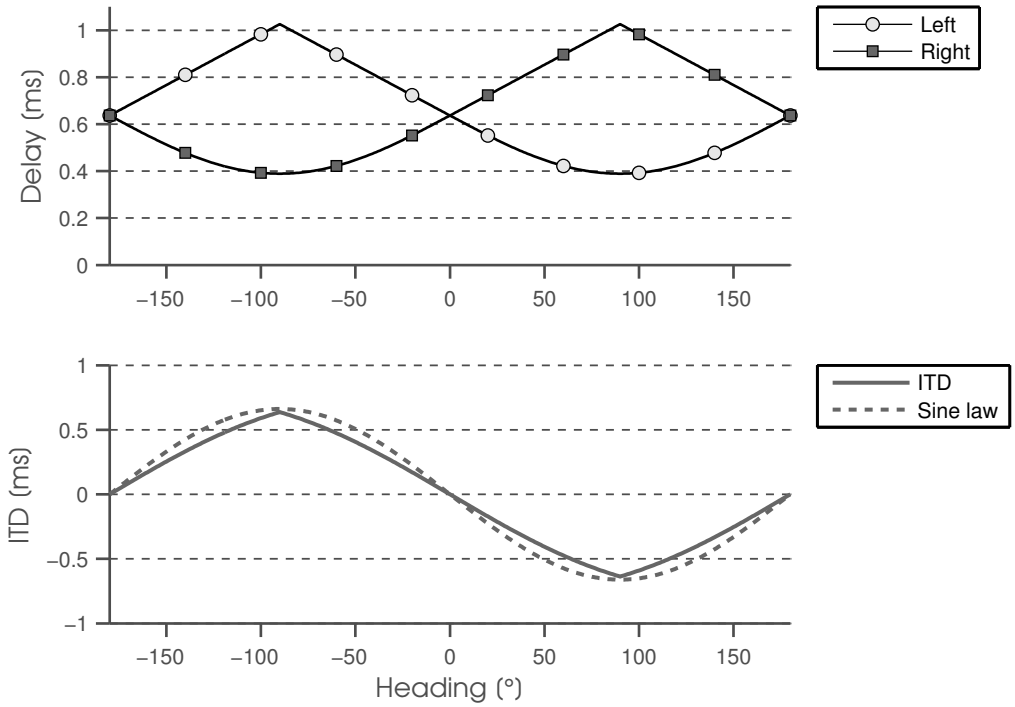


Figure 4.3: *ITD correction.* An angle-dependent delay ($TD_{correction}$) is applied to both channels to obtain the desired ITD. The graph above shows the delay values to compensate the ITD of a binaural recording for head orientation angles from −180° to 180°. The ITD resulting from delaying both channels is given in the graph below. It follows approximately the "sine law" (dashed line, cf. eq. 3.1).

in task I of the Kamara2009 project, and is not part of this work.

The aim of the de-panning process is to remove the alterations of the interaural cues introduced by head movement. These alterations occur both in the time domain and in the spectral domain. The following sections propose methods to remove or minimise these alterations. Limitations of the proposed methods are discussed in section 4.3.

ITD correction

The most important alteration of interaural cues caused by head movement during a binaural recording is a change in the time of arrival of the signal at both ears. This results in an altered ITD. If, for simplicity, the ITD is assumed to be frequency-independent (see Wightman and Kistler [Wightman and Kistler, 1997]), it can be represented by a simple delay of the signal at one ear with respect to the other. The head movement affects this delay. Thus, by delaying the binaural signals appropriately in the de-panning process, this alteration of the ITD can be removed

$$ITD_{desired} = ITD_{current} + ITD_{correction}, \tag{4.7}$$

where $ITD_{desired}$ is the ITD of the sound source without head movement, $ITD_{current}$ is the ITD after the head movement and $ITD_{correction}$ is a correction delay to compensate for the head movement. The correction delay results from applying an appropriate delay $TD_{correction}$ to each channel

$$ITD_{correction} = TD_{correction,left} - TD_{correction,right}. \tag{4.8}$$

The resulting azimuth-dependent $ITD_{correction}$ is shown in fig. 4.3, below. The calculation of the right channel delay is given as

$$\begin{aligned} TD_{current} &= TD(\alpha_{source} - \alpha_{heading}) \\ TD_{desired} &= TD(\alpha_{source}) \\ TD_{correction} &= TD_{desired} - TD_{current} + TD\left(\frac{\pi}{2}\right), \end{aligned} \tag{4.9}$$

where α_{source} is the source azimuth angle and $\alpha_{heading}$ is the azimuth angle of the head orientation. The calculation of the left channel delay is analogous, with angles multiplied by -1. The last term in eq. 4.9 is a constant positive offset to ensure $TD_{correction} > 0$, such that only positive delays are applied to each channel. $TD(\alpha)$ is the frequency-independent delay as a function of the angle of incidence [Rocchesso, 2002]:

$$TD(\alpha) = \begin{cases} \frac{f_s}{\omega_0} \cdot [1 - \cos(\alpha)] & \text{if } |\alpha| < \frac{\pi}{2}, \\ \frac{f_s}{\omega_0} \cdot \left[|\alpha| - \frac{\pi}{2} + 1\right] & \text{else.} \end{cases} \tag{4.10}$$

with

$$\omega_0 = \frac{c}{r}, \tag{4.11}$$

where r denotes the head radius (i.e. half the distance between the two ear entrances) and c the speed of sound. By delaying each signal with an appropriate $TD_{correction}$, depending on the head azimuth, the influence of head rotation on the ITD can be eliminated. Fig. 4.3 shows the $TD_{correction}$ values dependent on the head azimuth.

ILD correction

Head rotation affects the effect of head shadowing on the ear input signals and thus changes the ILD. To minimise this alteration, an approach analogous to the previously described ITD correction is taken

$$ILD_{desired}[dB] = ILD_{current}[dB] + ILD_{correction}[dB], \tag{4.12}$$

where $ILD_{desired}$ is the ILD of the sound source without head movement, $ILD_{current}$ is the ILD after the head movement and $ILD_{correction}$ is a gain factor to compensate for the head movement. The correction gain factor results from applying an appropriate gain $LD_{correction}$ to each channel

$$ILD_{correction}[dB] = LD_{correction,left}[dB] - LD_{correction,right}[dB]. \tag{4.13}$$

The calculation of the right channel gain factor is given by

$$\begin{aligned} LD_{current} &= LD(\alpha_{source} - \alpha_{heading}) \\ LD_{desired} &= LD(\alpha_{source}) \\ LD_{correction}[dB] &= LD_{desired}[dB] - LD_{current}[dB], \end{aligned} \tag{4.14}$$

with α_{source} and $\alpha_{heading}$ denoting the source and head azimuth angle respectively. The calculation of the left channel gain factor is analogous, with angles multiplied by -1. The calculation of the LD factors is based on a simple 1-pole/1-zero head shadow model proposed by Rocchesso [Rocchesso, 2002]

$$H_{hs}(z, \alpha) = \frac{(\omega_0 + \rho(\alpha) \cdot f_s) + (\omega_0 - \rho(\alpha) \cdot f_s)z^{-1}}{(\omega_0 + f_s) + (\omega_0 - f_s)z^{-1}}, \tag{4.15}$$

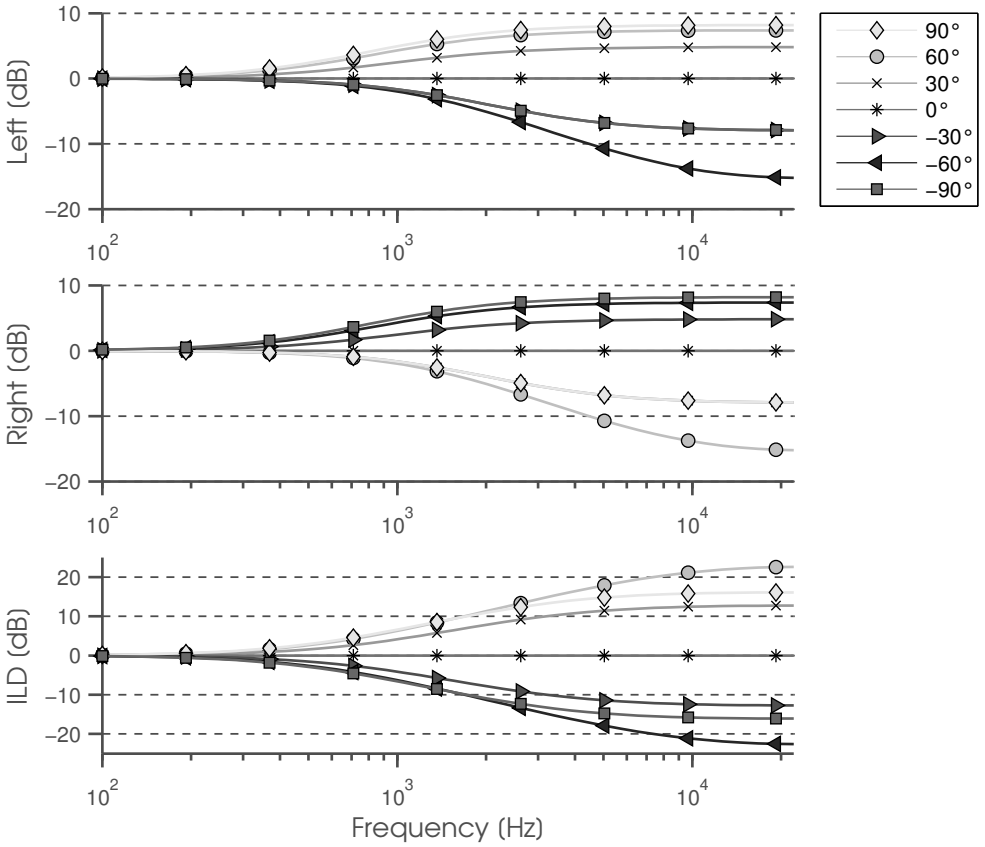


Figure 4.4: *ILD correction.* The zero-pole-gain filter responses are shown for a head orientation from $-90°$ to $90°$, for left and right channel. The bottom graph displays the resulting ILD correction. The strongest ILD correction is reached for an azimuth of $\pm 60°$.

where H_{hs} is the transfer function modelling the head shadowing effect. It basically describes a shelving filter with a gain ρ at the Nyquist limit dependent on the azimuth α; ω_0 is defined in eq. 4.11, f_s denotes the sampling rate, and ρ is given by

$$\rho(\alpha) = 1.05 + 0.95 \cos(\frac{6}{5}\alpha). \tag{4.16}$$

LD is defined as the gain of the shelving filter at the Nyquist limit

$$LD(\alpha) = \rho(\alpha). \tag{4.17}$$

The transfer function has an azimuth-dependent zero q_{hs} and a fixed pole p_{hs}

$$p_{hs} = \frac{1 - \frac{\omega_0}{f_s}}{1 + \frac{\omega_0}{f_s}}. \tag{4.18}$$

From this head shadow model, a simple zero-pole-gain shelving filter is derived to achieve the gain correction $LD_{correction}$, by filtering both channels with the following transfer function

$$H_{LD}(z, \alpha) = k \cdot \frac{1 - q(\alpha)z^{-1}}{1 - pz^{-1}}, \tag{4.19}$$

where k is the filter gain, p is the pole and q is the zero of the filter. The pole of the filter is fixed (as defined in eq. 4.18)

$$p = p_{hs}. \tag{4.20}$$

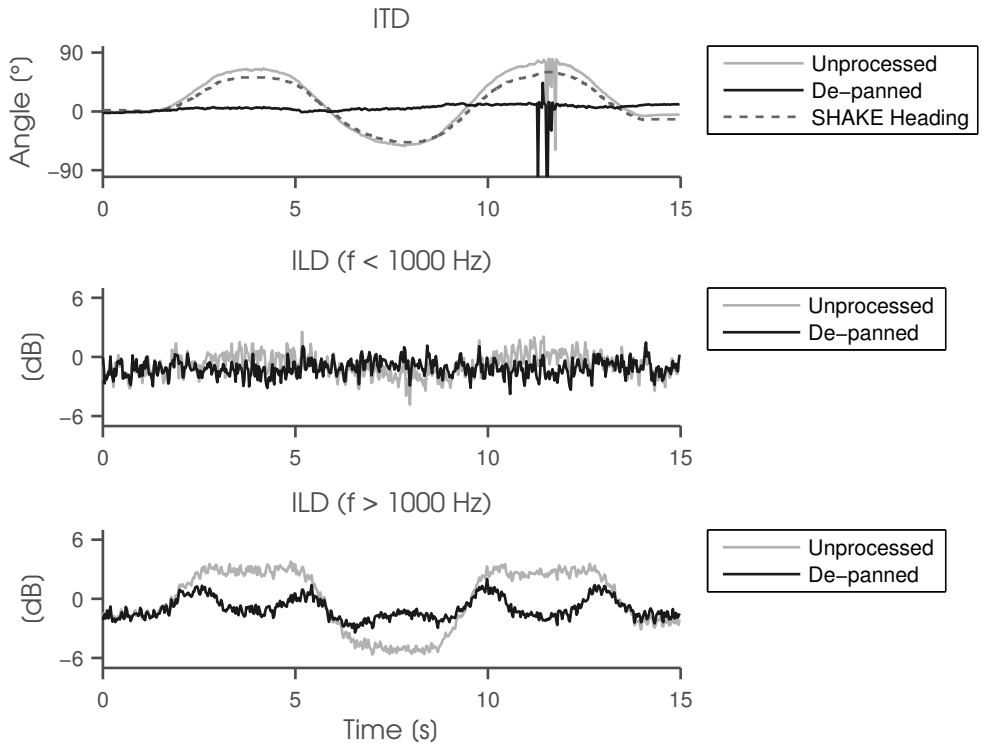


Figure 4.5: *Example of ITD and ILD correction.* Head movement (dashed line) causes ITD and ILD variation. The ITD is calculated as the maximum of the IACC. Around 12 s into the recording, the head orientation approaches 60°. Head shadowing lowers the signal-to-noise ratio of the direct path, causing strong early reflections to distort the IACC. The ILD values are calculated below and above 1000 Hz.

The gain k and zero q of the filter are described by two criteria: At low frequencies, the impact of head shadowing is negligible, therefore no gain correction is applied. The filter has a DC gain of unity (cf. eq. 4.21). At high frequencies, the impact of the head rotation on the head shadowing effect and thus the ILD is minimised by applying the gain factor $LD_{correction}$. The gain at the Nyquist limit equals $LD_{correction}$ (cf. eq. 4.22).

$$H_{LD}(z,\alpha)|_{z=1} = k \cdot \frac{1-q(\alpha)}{1-p} \stackrel{!}{=} 1 \tag{4.21}$$

$$H_{LD}(z,\alpha)|_{z=-1} = k \cdot \frac{1+q(\alpha)}{1+p} \stackrel{!}{=} LD_{correction}(\alpha). \tag{4.22}$$

Solving eq. 4.21 and eq. 4.22 for q and k yields

$$q(\alpha) = \frac{\phi - 1}{\phi + 1} \tag{4.23}$$

for the filter zero q with

$$\phi = LD_{correction}(\alpha)\frac{1+p}{1-p} \tag{4.24}$$

and

$$k = \frac{1-p}{1-q(\alpha)} \tag{4.25}$$

for the filter gain k. By applying a zero-pole-gain filter H_{LD} with appropriate parameters to both channels, the impact of the head rotation on the ILD of the recorded binaural signals is lowered. Fig. 4.4 shows the filter response of the zero-pole-gain filter H_{LD} for various head orientations.

The effect of the de-panning algorithm applied to a binaural recording is shown in fig. 4.5. The input signal is white noise, played back from a loudspeaker in a small office environment and recorded with the KAMARA headset. During the recording, the head orientation changed from about 60° to −60° and back. The head was tracked with the SHAKE device. The resulting ITD change is converted to an angle offset θ of the recorded source with the following approximation (after Raspaud and Evangelista [Raspaud and Evangelista, 2008])

$$\theta = g^{-1}(\omega_0 \cdot ITD) \tag{4.26}$$

with

$$g^{-1}(x) = \frac{x}{2} + \frac{x^3}{96} + \frac{x^5}{1280}, \tag{4.27}$$

where θ is the azimuth offset, ω_0 as defined in eq. 4.11, and ITD is the ITD value. The angle offset calculated from this value is close to the head orientation given by the SHAKE head tracking device (cf. fig. 4.5, dashed line). The offset is corrected by the de-panning algorithm using the head orientation information from the SHAKE device.

The ILD change due to head shadowing is negligible for frequencies below 1000 Hz (cf. fig. 4.5, middle graph). Above 1000 Hz, the de-panning algorithm compensates for the head shadowing effect (cf. fig. 4.5, bottom graph). The measured ILD values are well below the theoretical values, due to the reverberation in the small office environment (cf. section 4.2.5).

Other effects of head rotation

The impact of head rotation on a binaural recording is rather complex. Correcting ITD and ILD only, despite being the dominant spatial cues, cannot account for all the alterations introduced when the recording head is rotated. The simple head shadowing model described in eq. 4.15 is a very rough approximation of the azimuth-dependent spectral shape of a binaural recording – the HRTF. The model does not take into consideration pinna and shoulder reflections. These manifest themselves as azimuth-dependent peaks and notches in the HRTF, which the brain of a human listener is trained to recognise and map to the corresponding direction. De-panning binaural audio by correcting only ITD and ILD thus inevitably leads to contradictory spatial cues in the de-panned audio. It is shown by Wightman and Kistler, however, that listeners determine the position of virtual sound sources mainly relying on the ITD cue, even in the presence of conflicting other cues that indicate an opposite direction [Wightman and Kistler, 1997]. Therefore, correcting the ITD in the de-panning process yields the desired perceived direction also for a contradictory other cues, such as the spectral shape.

Whilst the described dominance of the ITD is apparent in a static scenario, with fixed recording head, continuous head rotation is rather problematic. When the head is rotated, the spectral shape of the binaural recording changes continuously, introducing motional cues (see section 2.3.6). In a static scenario, the peaks and notches caused by pinna and shoulder reflections may not be perceptible, especially if the spectrum of the sound source is unknown. Head movements, however, reveal them, as their position and shape changes according to a particular pattern the auditory system is trained to recognise. Superimposing these motional cues onto the de-panned ITD and ILD cues creates a rather unpleasant listening experience: Whilst the static ITD and ILD cues indicate a static source, the motional cues reveal the head rotation, indicating a moving source. This unnatural listening situation deteriorates the listening comfort and the externalisation of the binaural recording and should be avoided when using the current setup.

4.2.4 Panning of binaural audio

Once the head rotation at the remote end in the telecommunication scenario is compensated for via the de-panning process, the binaural recording can be presented to the local participant

via the KAMARA headset. The local participant perceives the remote participants of the teleconference at fixed locations, independent of the head orientation of the remote user wearing the recording headset. As described earlier, it might be desirable to register the binaural sounds with the environment of the local participant, to allow the user for instance to turn towards the remote speakers. This "panning" process is analogous to the de-panning process described in the previous section. The head of the local participant has to be tracked and the interaural cues of the binaural recording need to be adjusted according to the desired perceived direction of the sound source (cf. fig. 4.6c). Again, this is achieved by tuning ITD and ILD.

In fact the de-panning and the panning process can be merged. Instead of de-panning the recording to the original position (to compensate for head rotation of the remote user) and then panning it to desired position (given by the head orientation of the local user), the recording can directly be panned to the desired position, by combining the head orientations of the remote and the local user

$$\alpha_{desired} = \alpha_{current} - \alpha_{remote} + \alpha_{local}, \tag{4.28}$$

where $\alpha_{desired}$ is the desired azimuth of the recorded source at the local end, $\alpha_{current}$ is the perceived azimuth after rotation of the recording head, α_{remote} is the azimuth of the recording head and α_{local} is the azimuth of the head of the local user.

Merging de-panning and panning to a single process provides the advantage of eliminating redundant calculations and reducing the computational complexity. Low latency is vital in an interactive telecommunication scenario. Processing the binaural audio in a single step has another major benefit: In a communication scenario it is natural for participants to turn towards the speaker. Therefore, the head orientations of both the remote and the local user are assumed to be similar, if the speaker is registered with the local user's environment. In this case, little or no processing is applied to the binaural recording (cf. fig. 4.6d), as the actual source position, relative to the remote user, and the desired source position, defined by the head orientation of the local user, are similar or identical. Without processing of the binaural recording the spatial cues including the HRTF are left unaltered. This minimises the negative effect of the processing on the listening comfort and the externalisation of the binaural recording.

4.2.5 Implementation in Pure Data

The audio processing is implemented in the programming environment [Puckette, 1996]. A computer running Pure Data (Pd) is fed with the binaural KAMARA recording and the head tracking information of the remote and the local KAMARA user. If a sound source is recorded at the remote end, its azimuth is transmitted to the local end. From this angle and the head orientations of the remote and local user the control signals for the algorithms are calculated.

The various processing blocks are depicted in fig. 4.7. The zero-pole-gain shelving filter, defined by eq. 4.19, is applied to correct the ILD of the recorded source and pan it to the desired position. A wetness factor σ is used to control the amount of correction

$$y = \sigma x_f + (1 - \sigma)x, \tag{4.29}$$

where x is the input signal (i.e. one channel of the binaural recording), x_f is the input signal filtered with the zero-pole-gain filter and y is the output signal of the filter block. The wetness factor σ is chosen dependent on the room reverberation. In a reverberant space, the actual ILD differs from the theoretical ILD given by eq. 4.15. Reflections from various directions balance out the energy reaching both ears, and thus lower the ILD. The wetness factor accounts for this, by lowering the effect of the zero-pole-gain filter.

Next, the ITD is corrected by delaying both channels appropriately. This is achieved using Pure Data's *vd∼* object, which implements a delay line with 4-point interpolation, allowing for fractional delays.

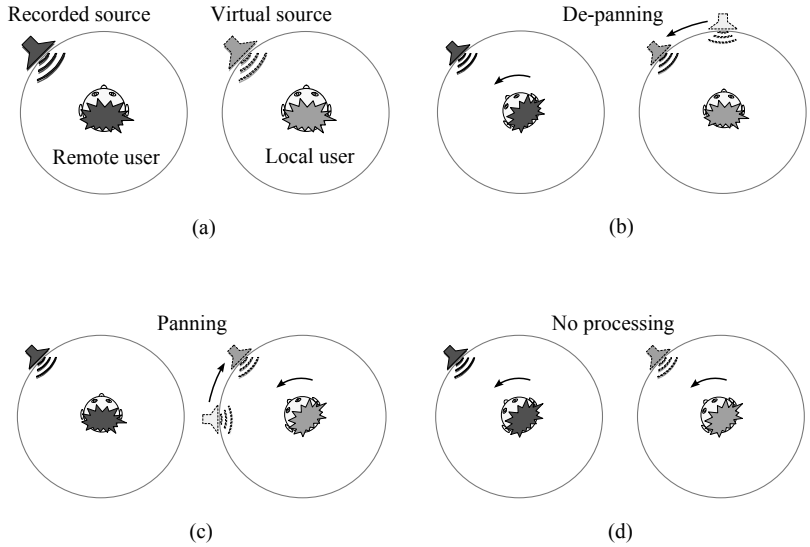


Figure 4.6: *De-panning and panning.* (a) The source recorded at the remote end is perceived at the local end as a virtual source at the same direction. (b) De-panning is applied to compensate for head movements of the remote user. (c) Panning compensates for head movements of the local user, to register the virtual sources with the environment. (d) No processing is necessary if the head orientation of both users is the same, e.g. if both are facing the source.

The mixing stage

After the ITD correction, a crossover filter consisting of two second-order high- and lowpass butterworth filters is applied to both channels. The signal is split into a high and a low frequency channel with a crossover frequency of 1000 Hz. The signal from the ear which is closer to the recorded source (the ipsilateral ear) is mixed into the signal of the other (contralateral) ear. The amount of mixing depends on the desired source direction. It increases if the source is panned to the front of the user. The underlying assumption is that the interaural differences vanish when the user turns towards the source. Mixing the high-frequency channels lowers the interaural differences above the crossover frequency.

The low frequency channel is left unaltered, as the head shadowing has little effect on it. This preserves the decorrelation of the left and right channel, which is an important factor for the externalisation of binaural sound [Rocchesso, 2002]. After mixing, the high and low frequency channels are combined to a single channel again, to yield left and right output signals.

The swapping stage

The last block determines how the recording channels on the remote end are mapped to the playback channels on the local end:

- *Direct mapping:* the left output channel of the crossover filter is played back to the left ear of the local user, the right channel is played back to the right ear.
- *Swapped mapping:* the playback channels are swapped, hence the left channel is played back to the right ear, and the right channel is played back to the left ear.

Direct mapping is applied if the ipsilateral ear is the same during recording and playback, i.e. the source is recorded to the same side as it is desired to be perceived. Swapped mapping is applied if the ipsilateral ear changes from recording to playback, i.e. the source is recorded to the opposite side as it is desired to be perceived. As Gardner and Martin state, HRTFs are

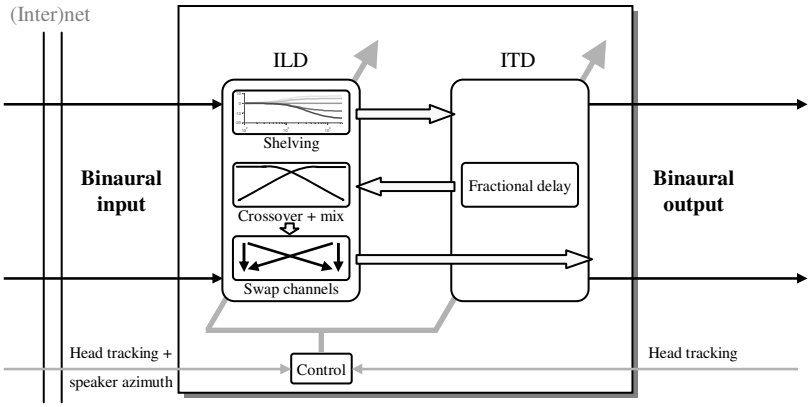


Figure 4.7: *Processing of binaural audio.* The binaural input from a remote KAMARA user is processed to compensate for head movements and registered with the environment of the local user.

symmetric, assuming a symmetric head [Gardner and Martin, 1995]. The response measured at the left ear at azimuth α is the same as at the right ear at azimuth $-\alpha$. This implies that also the ear input signals are identical in this case. Thus, if a source is recorded at azimuth α but desired to be perceived at azimuth $-\alpha$, instead of panning the source from α to $-\alpha$, swapped mapping is applied and no panning is necessary. If the desired azimuth is $-\alpha+\epsilon$, after swapped mapping is applied, only the offset ϵ has to be compensated for through panning. This avoids the need to convert the contralateral ear in a recording to the ipsilateral ear in the playback, which would require reversing the head shadowing effect to restore the damped high frequency content in the recorded signal, a problematic issue especially in situations with low signal-to-noise ratio.

4.2.6 Virtual sound source positioning using finger snaps

Binaural signals result from the filtering behaviour imposed by the head, shoulders, torso and the room on a sound sample. A binaural room impulse response (BRIR) is the time-domain representation of this filtering behaviour. It is the impulse response measured at the ear entrances of a listener or dummy head for a certain source position inside the room. Convolving a monaural sound sample with this BRIR yields the same binaural signals as though the sample was played back from the source position used in the impulse response measurement. When listening to the resulting binaural signals, the virtual sound is perceived as emanating from this position in the recorded room.

Applying the BRIR to a monaural speech sample results in the perception of the speaker being spatialised, as in a binaural recording. a straightforward way to obtain a BRIR is to record an impulse with the KAMARA headset inside the room. By using a finger snap or clap as the excitation signal, a BRIR can be obtained on the fly. Though the BRIR is coloured with the spectrum of the snap or clap, it contains the filtering behaviour of the KAMARA user's own head and pinnae and listening space. This might increase the perceived realism of virtual speech sources, as their spatial attributes match the actual surroundings of the listener, as well as the listener's own HRTF. An overview of the system is shown in fig. 4.8; a detailed description can be found elsewhere [Gamper and Lokki, 2009].

The position of the snap or clap determines the perceived position of the virtual source. Thus, participants of a teleconference can be positioned around a KAMARA user by simply clapping or snapping at the desired positions. This could be used as an alternative to or in combination with the binaural recording of a remote meeting with the KAMARA headset. A problem not addressed so far in this work is the perception of the remote KAMARA user's own voice. As it contains little or no interaural cues, being recorded at the centre of the KAMARA

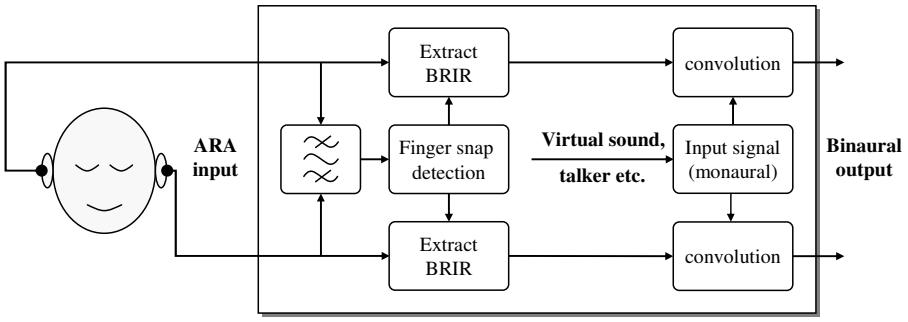


Figure 4.8: *Instant BRIR acquisition.* If a finger snap is detected in the signal of the remote KAMARA user, a BRIR is extracted from each microphone channel and convolved with the input signal, i.e. a monaural speech signal of a virtual remote teleconference participant. Convolving each speaker with a separate snap, the participants can be spatially separated [Gamper and Lokki, 2009].

microphones, it may suffer from poor externalisation and IHL. This problem could be tackled by applying a BRIR, obtained on the fly at the local end.

4.3 Limitations

Certain criteria have to be met for the proposed algorithms to be applicable. The de-panning and panning algorithm presented above does not support multiple simultaneous sources. The algorithm is designed to pan only one source at a time. In the case of a telecommunication scenario, the algorithm fails if the speakers do not talk in turns. This poses some limitations on the usability of the system under certain circumstances.

A problem related to a situation with multiple simultaneous talkers are strong room reflections in a reverberant space. Each reflection reaches the ears and thus the recording microphones as a delayed version of the original signal filtered with the transfer function of the reflection path and an HRTF, before being recorded. As the algorithms are designed to compensate sound from just one source direction, the directions of these reflections are not taken into account in the processing of the binaural recording. Head rotation affects these reflections, as their direction relative to the head orientation changes and they are filtered with different HRTFs depending on the head orientation. This introduces additional motional cues to the binaural recording. These cues may be contradictive to the adjusted ITD and ILD cues, and thus deteriorate performance.

The performance of the algorithm heavily relies also on accurate head tracking and determination of speaker positions. Errors or noise in either of the measurements deteriorate the performance and listening comfort. The update rate of the head tracking device puts limits to the responsiveness of the system. In case of fast head movements, the audio scene lags behind. The algorithms currently only compensate for head movement in the horizontal plane. The elevation of both the local and the remote user's head is neglected. The system does not account for sources placed above or below the horizontal plane, or for a recording head being bent up or down.

A common problem with binaural audio is IHL. Generating externalised virtual sound sources is difficult, especially in absence of visual cues. Several factors in the presented implementation have a negative impact on the externalisation. First of all, HRTFs are highly individual. Thus, binaural audio recorded on one listener and played back to another deteriorates the spatial listening experience and thus the externalisation. The human brain is not trained to listen through someone else's ears, and, as Rocchesso states, tends to internalise unnatural sounding audio events [Rocchesso, 2002]. The de-panning and panning process does not take into account

the fine structure of the HRTFs of either the remote or the local user. This introduces further artefacts which degrade the externalisation. Some spatial cues are neglected, such as room reflections and motional cues. If these cues conflict with the ITD and ILD, the externalisation suffers. Mixing the high frequency channels of the binaural recording lowers the decorrelation - and favours IHL.

A possible approach to tackle the aforementioned problems is to determine the fine structure of the HRTFs of the user. Inverse filtering using a standard HRTF dataset to extract the non-spatialised input sound from the binaural recording, a technique used in robot audition [Keyrouz et al., 2007], could be considered as an alternative approach to the proposed method and is left to future research. The same holds for the processing of multiple simultaneous speakers and room reflections.

Chapter 5

Evaluation

5.1 User study

To evaluate the performance of the proposed AAR telecommunication system under controlled conditions, a formal user study was conducted. The study was designed to prove or falsify the following hypotheses:

> **Hypothesis I:** *Listeners can localise speakers in binaural audio recorded on a human subject other than themselves. The localisation performance does not deteriorate when de-panning is applied to the binaural recording.*
>
> **Hypothesis II:** *Panning (i.e. registering the binaural audio with the environment of the listener) improves the localisation performance.*
>
> **Hypothesis III:** *Interaural cues improve the ability of listeners to segregate multiple speakers.*
>
> **Hypothesis IV:** *Turning towards a speaker improves the ability to segregate that speaker from other speakers.*

The following sections describe the test setup and procedure. Results are presented and discussed at the end of this chapter.

5.2 Method

To test the presented algorithm, a telecommunication scenario similar to the one described in section 4.2.2 was simulated: The test subjects were presented with a binaural recording of a remote conference. The conference and its participants were recorded via a KAMARA headset. The binaural audio was processed, i.e. de-panned and panned, and played back to the test subject over a pair of headphones. The test subject had to perform various tasks related to the hypotheses introduced above.

5.2.1 Audio material

To ensure repeatability of the test and to avoid technical problems, the remote conference was recorded before the actual user study. As the location of the simulated conference, a lecture hall with a reverberation time of 0.3–0.5 seconds was chosen. The floor plan of the hall is square with an area of $95\,\mathrm{m}^2$. Approximately 1.5 m above the ground at a radius of 5–6 m 12 Loudspeakers are arranged in the horizontal plane at intervals of about 30°. A subset of these loudspeakers was

used to play back recordings of male speech from various directions (cf. figs. 5.1 and 5.3). The recordings are taken from the Bang & Olufsen CD *Music for Archimedes* [Bang & Olufsen, 1992] and from the TIMIT database [Garofolo et al., 1993]. The simulated conference was recorded with a KAMARA headset worn by a user sitting in the centre of the hall.

5.2.2 Test procedure

The test subjects were seated in front of a computer, either in an office, a studio or a home environment. The binaural recording of the simulated remote conference was processed on the computer and played back to the test subjects via Sennheiser HD-590 headphones. These full size headphones were chosen for playback instead of the KAMARA headset due to their superior quality. In-ear phones such as the ones used with the KAMARA headset are difficult to fit to the ears of a user and their performance is sensitive to the placement, which in turn might affect the test results.

After a short introduction to the test, each test subject was given a questionnaire with instructions for various tasks. Every subject had to accomplish all tasks and was thus tested in all conditions (*within–subjects* design). To minimise learning effects, the test subjects were divided into groups. The order of the tested conditions was randomised among groups.

A total of 13 test subjects participated to the study. 5 of the test subjects were students of the Department of Media Technology of the Helsinki University of Technology. Having vast experience in using and assessing spatial audio, they were classified as "professional listeners". The other 8 subjects had little or no experience with spatial audio, and were thus classified as "naïve listeners". The inexperienced subjects were given a short introduction to spatial audio and the working principle of the head-tracking device and the audio panning before the test. The test consists of two main tasks, described in the following sections.

5.2.3 Task I – speaker localisation

The first task tests the ability of test subjects to localise a speaker in a binaural recording. The recording consists of ten repetitions of a male speech sample from the "Music for Archimedes" CD [Bang & Olufsen, 1992]. The sample duration is about 11 seconds, with 1 second of silence between each repetition. Two different conditions are tested in Task I: *static* and *de-panned*.

For the *static* condition, the binaural recording was made using five loudspeakers: three in front (at 30°, 0° and −30° azimuth), one to the right (at −90°), and one in the back (at 150°). The anechoic speech sample was played from each loudspeaker, in random order (cf. fig. 5.1). Each direction occurred twice, yielding a total of ten repetitions of the speech sample. The recording was made without head movements.

The *de-panned* condition assumes a situation were the remote participant recording the conference is turning towards the currently active speaker, a natural behaviour in an actual communication scenario. This results in all speakers being recorded in front of the user. Therefore, little or no interaural cues are present in the recording to allow a listener to segregate the speakers. Neglecting the influence of room reflections, if the same speech sample is played from various directions, turning towards the active speaker will indeed result in a series of almost identical binaural recordings. Thus, to simulate this scenario, just one loudspeaker in front of the KAMARA user, at 0° azimuth, was used for the recording, with the KAMARA user facing the loudspeaker. The speech sample was the same as in the *static* condition. The recorded sample was then de-panned to encode the interaural cues of the same azimuth angles as used in the *static* condition (i.e. 150°, 30°, 0°, −30° and −90°). The listener should thus perceive the speakers as emanating from these directions, even though they were recorded from just one direction. Again, the order of the directions was randomised, with each direction occurring twice. The recording setups for Task I are shown in fig. 5.1.

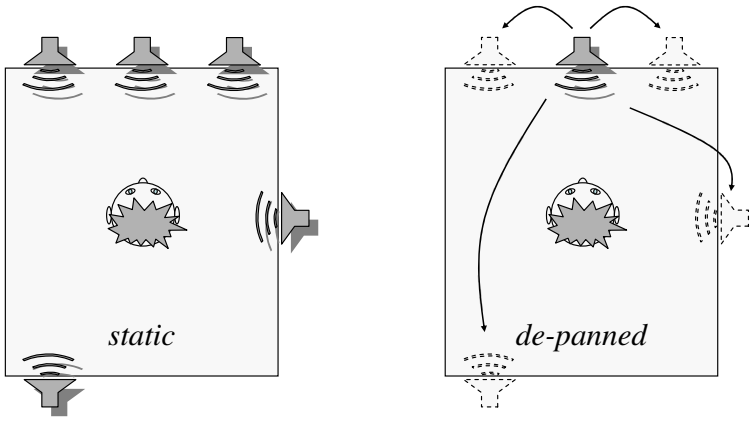


Figure 5.1: *Recording conditions for speaker localisation task.* For the *static* recording, the speech sample is played back from one of five different loudspeakers. For the *de-panned* recording, only one loudspeaker is used. The spatial separation of the speech signals is obtained through de-panning.

	Subtask I - *without panning*	Subtask II - *with panning*
conditions	*static* *de-panned*	*static* *de-panned*

Table 5.1: *Order of recording conditions.* Task I is subdivided into two subtasks, *without panning* and *with panning*. Each subtask is tested in two conditions, *static* and *de-panned*.

Task I is subdivided into two subtasks. In each subtask both conditions, i.e. *static* and *de-panned*, are tested. In the first subtask, the binaural recording is not registered with the environment of the test subject; no panning is performed. In the second subtask, the head of the test subject is tracked, and the binaural recording is panned accordingly to register it with the environment. The test subject is thus able to turn towards the active speaker.

Subtask I – without panning

In the first subtask of Task I, the test subjects were asked to specify the direction of the speakers in the binaural recording. The subjects had to choose from twelve potential directions, corresponding to the loudspeaker positions in the recording hall, as shown in fig. 5.2. Only five out of twelve directions were actually used in the recording, with each direction occurring twice (see fig. 5.1). The task was performed in two conditions, *static* and *de-panned*, as described earlier. The subjects were not allowed to train or repeat the task. Learning effects were expected to occur, favouring the condition tested second. To minimise this effect the order of the conditions was randomised among subjects.

The hypothesis of this task (hypothesis I) is that there is no significant difference in the localisation performance of test subjects between a recording made with loudspeakers at different positions (*static* condition) and a recording made with just one loudspeaker and processed to yield the impression of speakers emanating from various directions (*de-panned* condition).

Subtask II – with panning

In the second subtask, the head of the test subject was tracked with the SHAKE device. With the head orientation of the test subject, the binaural recording of the simulated remote conference was panned and registered with the environment. The test subjects were asked to turn towards

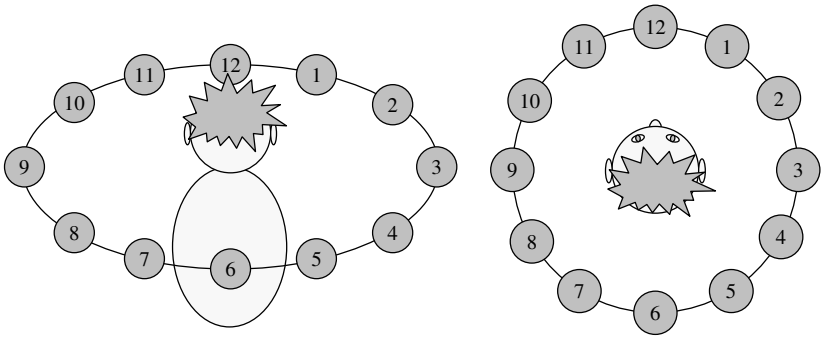


Figure 5.2: *Localisation questionnaire.* In subtask I, test subjects of the user study had to map the perceived speaker position to one of 12 given directions.

the speaker they hear. When a test subject confirmed to have reached the desired direction, the head orientation was logged. Again, this was tested in the *static* and *de-panned* condition, in random order to minimise learning effects.

The hypothesis of this task (hypothesis II) is that test subjects can localise speakers more accurately by turning towards them than guessing their direction. Performance in the *de-panned* condition is expected to be slightly better: In this case, once the test subject faces the virtual speaker, the original binaural recording is delivered nearly unprocessed, as the speaker is recorded in front of the KAMARA user. This is supposed to yield better localisation accuracy and externalisation than the *static* case, where the virtual speakers have to be panned to be perceived as being in front.

5.2.4 Task II – speaker segregation

Task II of the user study examines the ability of test subjects to segregate speakers of a remote conference with multiple participants. In the simulated conference, four male speakers are positioned around the remote KAMARA user. The task of the test subject is to listen to the conference and identify one speaker among the four. The speech samples for this task are taken from the TIMIT database [Garofolo et al., 1993]. Eight male speakers of the database were chosen, and combined to two groups of four speakers. Ineach tested condition, every speaker utters five words, for a total of twenty words per condition. The speakers talk in turns, in random order. In addition, one complete sentence is recorded from each of the eight speakers. The sentence is the same for all speakers, and about 2 seconds long. Three different conditions are tested in this task: *static*, *moving* and *de-panned*.

In the *static* condition, each speaker is assigned a different loudspeaker in the recording hall. The speech samples are played back in random order from these loudspeakers. Thus each speaker is recorded at a different but fixed azimuth: speaker 1 at 60°, speaker 2 at 30°, speaker 3 at 0° and speaker 4 at −30°. This simulates a situation where the conference participants are seated around a table with the KAMARA user.

In the *moving* condition, a situation is simulated where the remote KAMARA user turns towards the currently active speaker. All speakers are recorded with one loudspeaker in front of the KAMARA user. The KAMARA user faces the loudspeaker throughout the recording, similar to the recording for the *de-panned* condition in Task I. All four speakers are recorded at 0° azimuth, in random order, each uttering five words.

The last condition, *de-panned*, assumes the same situation as the *moving* condition, i.e. that the remote KAMARA user turns towards the active speaker. Again, all speech samples are recorded from one loudspeaker in front of the KAMARA user at 0° azimuth. The binaural recording is then de-panned to encode interaural cues into each recorded speaker. As a result,

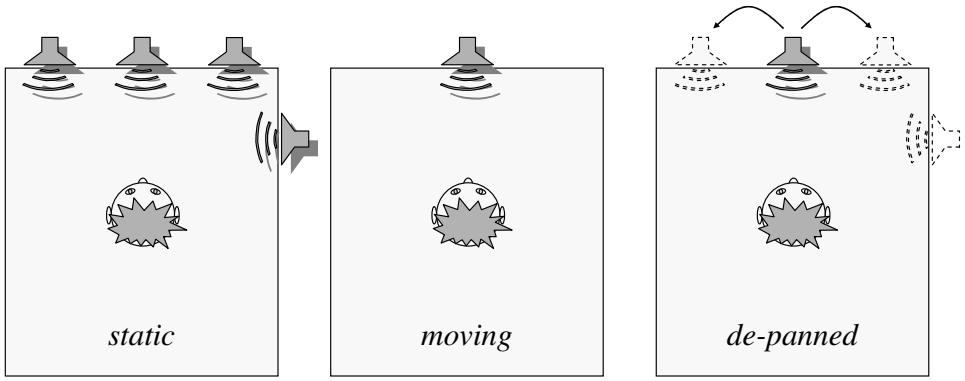


Figure 5.3: *Recording conditions for speaker segregation task.* For the *static* recording, a separate loudspeaker is used for each speaker. The *moving* and *de-panned* recordings are obtained from just one loudspeaker. De-panning is applied to separate the speakers on the *de-panned* recording spatially, thus simulating the speaker positions used in the *static* recording.

the perceived speaker azimuths are the same as in the *static* condition, where each speaker was actually recorded from a different loudspeaker position. The same group of four speakers is used as in the *static* condition, with a different set of words. The loudspeaker setups for each recording condition are depicted in fig. 5.3.

Task II is again subdivided into two subtasks. In subtask I, all three conditions, i.e. *static*, *moving* and *de-panned*, are tested. No panning is performed, therefore the virtual speakers are not registered with the environment of the test subject. In subtask II, head tracking is used to pan the binaural recording and register the remote talkers with the environment. This allows the test subject to turn towards the speakers.

In each subtask, the test subject is presented with a list of twenty words. The words are listed in order of appearance. At the beginning of each test round, the four speakers recorded for the test condition introduce themselves by saying one sentence each. After this introductory round, the speakers utter five words each in random turns. The test subject is asked to remember the first speaker to be heard in the introductory round and mark the words uttered by that speaker. In doing so, the test subject has to segregate the four speakers. Task II tests the segregation performance in the previously described conditions.

To investigate upon the impact of learning effects on the performance, both subtasks are repeated three times. As only eight speakers in total are used for Task II, it is assumed that test subjects become acquainted with the different voices, which might improve the segregation performance from the first round to the last. To counterbalance the order in which the conditions are presented, the order is defined by a *Latin square* [Rapanos, 2008] (see table 5.2). The test subjects are divided into three different groups. Each group starts with a different row of the Latin square, to minimise learning effects.

Subtask I – without panning

Three conditions are tested in subtask I: *static*, *moving* and *de-panned*. A set of four male speakers is presented to the test subject in each condition. The speakers are introduced with a short sentence. The test subjects were instructed to remember the first speaker they hear. From a list of words the test subjects had to mark the words uttered by this speaker. The subtask was repeated three times, with changing order of the conditions (see table 5.2).

The hypothesis of this task (hypothesis III) is that speaker segregation performance is considerably worse in the *moving* condition, where all speakers are perceived at 0° azimuth, and

	Subtask I - *without panning*			Subtask II - *with panning*	
Round 1:	*static*	*moving*	*de-panned*	*static*	*de-panned*
Round 2:	*moving*	*de-panned*	*static*	*de-panned*	*static*
Round 3:	*de-panned*	*static*	*moving*	*static*	*de-panned*

Table 5.2: *Latin square ordering of recording conditions.* The rows of the matrix define the order the conditions appear in each round. In subtask I, each condition appears only once in each round and at each position (3 x 3 Latin square). In subtask II one row is repeated, as there are only two conditions in three rounds.

thus cannot be segregated based on interaural cues. This would result in higher error rates of the words being marked. It is further assumed that performance in the *de-panned* and the *static* condition is equal. In the *de-panned* condition, interaural cues were modified to match the *static* condition.

Subtask II – with panning

In the second part of task II, the conditions *static* and *de-panned* where tested. Test subjects were again instructed to remember the first speaker of the introductory round and mark the words uttered by this speaker in a list of twenty words. Via head tracking and panning the speakers were registered with the environment. Test subjects were advised to turn towards the speaker in question already in the introductory round and keep the head still afterwards. This way, every time the speaker talks, he is perceived in front, whilst all other speakers are perceived to either side of the test subject. The subtask was again repeated three times, with changing order of the conditions (see table 5.2). The condition *moving* was not tested, as panning of binaural audio is only reasonable in combination with de-panning: The head movement of the listener cannot be compensated through panning without also compensating for the head movement during the recording through de-panning.

It is assumed that facing the active speaker enhances segregation of that speaker from the others (hypothesis IV). This should manifest itself in lower error rates of the words marked compared to subtask I. Performance in the *de-panned* condition is expected to be better than in the *static* condition. In the *de-panned* condition, once the test subject faces the speaker in question, the binaural recording is delivered nearly unprocessed, hence processing artefacts are minimised.

5.3 Results

5.3.1 Objective and subjective measures

From the user study, both objective and subjective measures were obtained. The objective measures are the angle mismatch, the number of front–back reversals and the time needed to turn towards a speaker in the speaker localisation task, and the error rates in the speaker segregation task. The following sections explain how these measures were obtained.

As a subjective measure, test subjects were asked to judge the perceived difficulty of each subtask. The difficulty was marked on a balanced seven-step Likert scale [Gardner and Martin, 2007]. The seven-step scale was chosen to provide test subjects with enough choices to quantify differences in the perceived difficulty between the subtasks. To ensure the intervals between steps were perceived to be equidistant, exact verbal opposites were used on both ends of the scale, ranging from *not difficult* to *difficult*, with *medium* marking the centre point (see fig. 5.4).

How difficult was this task?	☐	☐	☐	☐	☐	☐	☐
	(not difficult)			(medium)			(difficult)

Figure 5.4: *Likert scale.* After each subtask, test subjects were asked to mark the perceived difficulty on a seven-step Likert scale.

The results of analysing both objective and subjective measures in each task are given in the following sections.

5.3.2 Task I – speaker localisation

Angle mismatch

One objective measure to determine performance in the speaker localisation task is the angle mismatch between the choice β of the test subject and the actual recording angle α. In the second part of Task I, where test subjects had to turn towards the speaker, the mismatch is calculated as the offset between the playback angle α and the head orientation β of the test subject. In case of optimal performance the mismatch would be 0° in both situations.

In each subtask, 10 observations were made of each test subject, yielding a total of 40 observations per subject, in 4 different conditions: *static* and *de-panned*, both without head tracking and panning (subtask I) and with head tracking and panning (subtask II). For every observation, the angle mismatch between the actual direction α of the speaker and the choice β of the test subject is calculated. This mismatch is compensated for front–back reversals. A front–back reversal occurs, when the test subject perceives the source as being in front when in fact it is in the back, and vice versa. The error due to the reversal is removed from the angle mismatch, as it would severely distort the measurement results [Wenzel et al., 1993]. If a test subject for instance perceives a source at 30° in front, when it was recorded at 150° in the back, the total offset without compensation would amount to 120°, which is not a value representative of the actual error caused by misinterpreting the interaural cues. Instead, the mismatch is calculated as if no front–back reversal had occurred. After this compensation, the angle mismatch in the given example amounts to only 30°. The reason for applying this compensation lies in the ambiguity of interaural cues. From interaural cues such as ILD and ITD alone, it cannot be determined whether a source is in the back or in front. Instead, the cues only define a cone of confusion, on which the source lies (see section 2.3.1). In the example above, 30° corresponds to the difference between the perceived direction and the angle of aperture of the cone of confusion on which the source lies. This difference is caused by a misjudgement of the interaural cues. Therefore, this approach allows to separate the influence of misjudging the interaural cues from the influence of front–back reversals on the localisation performance. The angle mismatch discussed hereafter refers to the mismatch after compensation for front–back reversals, which shows a normal distribution about 0°. The front–back reversals are analysed separately.

The localisation performance in each subtask is determined by the mean absolute angle mismatch Φ

$$\Phi = \frac{1}{N}\sum_{i=1}^{n} |\alpha_i - \beta_i|, \tag{5.1}$$

with $N = 10$ (i.e. the number of directions to be determined in each condition), the actual source direction α, and the user choice β. In subtask I (without panning) this performance measure is tested against hypothesis I, i.e. that the *de-panned* recording yields the same localisation performance as the *static* recording. In subtask II (with panning), the *de-panned* recording is expected to yield equal or better results than the *static* recording.

Front–back reversals

Another objective measure for localisation performance, besides the angle mismatch, is the number of front–back reversals. They are a common problem with binaural audio in absence of visual cues. Motional cues provided by head tracking and panning decrease the number of front–back reversals and thus improve localisation performance. Therefore, the number of front–back reversals is expected to be smaller in subtask II, where head tracking and panning is enabled, than in subtask I (hypothesis II).

Time needed to turn towards speaker

In subtask II, test subjects were asked to turn towards the active speaker and confirm when they perceived the direction of the speaker and their head orientation to match, i.e. when the speaker was perceived to be right in front. When the test subjects confirmed, their head orientation was logged. For every observation, the time needed to turn the head to the desired direction was measured, yielding a total of 10 measures per tested condition. Better performance is expected in the *de-panned* condition, where almost unprocessed audio is delivered to the test subject once the correct head orientation is reached. This should simplify the decision whether the source is perceived as being right in front and thus shorten the time needed to lock into the final head orientation.

Perceived difficulty

After each tested condition, test subjects were asked to mark the perceived difficulty on a Likert scale (cf. fig. 5.4). This serves as a subjective measure for each condition. As the perceived difficulty is expected to be highly individual, and the Likert scale is to be interpreted as an ordinal rather than an interval scale [Gardner and Martin, 2007], the results are expected to serve merely as a ranking of the tested conditions in terms of their relative perceived difficulty, rather than an absolute measure of the perceived difficulty.

Statistical analysis

The study is designed as a *within–subjects* test, i.e. each subject is tested in all conditions. To compare performance in the two conditions in each subtask, a paired two-way analysis is performed on the absolute values of the angle mismatches. By taking the absolute value, the analysis data is heavily skewed to the right. Some authors suggest to apply a nonparametric analysis in this case, as it does not require the data to be sampled from normally distributed populations [Zalis et al., 2005]. On the other hand, for large samples, parametric analyses are robust also when the data is sampled from a nongaussian population [Motulsky, 1995]. Boxplots of the absolute angle mismatches from both subtasks including the mean absolute angle mismatches are shown in fig. 5.5.

Results from both a (parametric) two-way analysis of variance (ANOVA) and a (nonparametric) Friedman analysis [Hill and Lewicki, 2006] are presented. The two-way ANOVA is performed to compare the mean absolute angle mismatch of the conditions in each subtask and to see whether there is a significant difference. Two factors are examined by the analysis: the test condition, i.e. *static* and *de-panned*, as factor I, and the test subject as factor II. The analysis indicates whether the null hypothesis, i.e. that all samples from one factor are drawn from the same population, may be rejected. As the amount of previous experience with spatial audio varied considerably among subjects, an influence of factor II on the results might imply an impact of previous experience on the localisation performance. In each condition, ten observations are made per subject. The analysis is thus performed for repeated measures. For a p-value $p < 0.05$ the null hypothesis is rejected, and the result is considered statistically significant. A

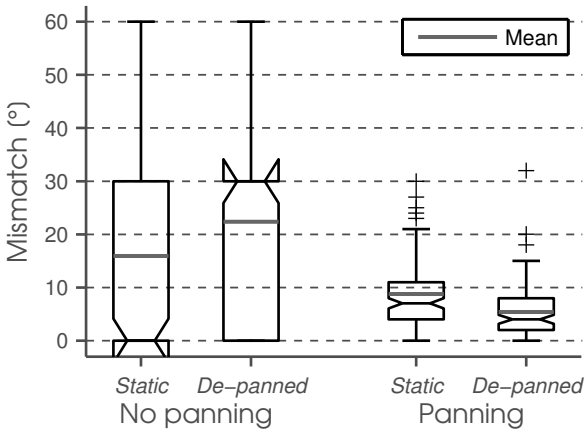


Figure 5.5: *Absolute angle mismatch.* Significant differences are found between the *static* and *de-panned* condition in subtask I without panning and in subtask II with panning. Both the *static* and *de-panned* condition without panning yield a significantly larger mean absolute angle mismatch than the *static* and *de-panned* condition with panning.

statistically significant difference can occur between test conditions, i.e. the mean absolute angle mismatch is significantly smaller in one condition than in the other, and between test subjects, meaning the performance varies significantly among subjects.

The Friedman analysis is the nonparametric counterpart of the parametric ANOVA. It is calculated on the ranks of the actual data, which makes it more robust against outliers. The analysis is performed for repeated measures, as there are ten observations per subject in each condition. The Friedman analysis indicates whether there is a significant effect due to factor I on the median of a sample. Factor I is the test condition. The null hypothesis states that there is no effect due to the factor I. If it can be rejected at the five percent significance level, the analysis indicates that factor I, the test condition, causes a significant difference between the sample medians. This means that the median of the absolute error is significantly larger or smaller in one condition than in the other, which may be interpreted as an indication for the dependence of the performance on the test condition. The analysis results are summarised in table 5.3.

For the main results, the test statistic is given along with the p-value derived from the statistic. For the ANOVA, the F-statistic is presented with the degrees of freedom df_1 and df_2 determining the F cumulative distribution function, as $F(df_1, df_2)$. For the Friedman analysis, the χ^2-statistic is presented with the degrees of freedom df_1 determining the χ^2 cumulative distribution function, as $\chi^2(df_1)$.

The box in box plots indicates the interquartile range from lower to upper quartile, with a line at the median value. Whiskers extend to the most extreme data values within 1.5 times the interquartile range. Non-overlapping notches indicate differences of the medians at the five percent significance level.

Applying the two-way ANOVA to the data of subtask I reveals that the mean absolute angle mismatch without panning is significantly smaller with the *static* recording (15.9°) than with the *de-panned* recording (22.4°), $F(1, 12) = 6.57$, $p_{Cond} = 0.0110$. The Friedman analysis yields an analogous result: The median of the absolute angle mismatch is significantly smaller with the *static* recording (0°) than with the *de-panned* recording (30°), $\chi^2(1) = 6.13$, $p_{Cond} = 0.0133$. Therefore, it can be concluded that the de-panning has a negative effect on the localisation performance. Hypothesis I, i.e. that the localisation performance is not affected by the de-panning, is thus falsified. No significant difference between subjects is found (ANOVA: $F(1, 12) = 1.02$, $p_{Subj} = 0.4315$, Friedman: $\chi^2(12) = 12.97$, $p_{Subj} = 0.3714$).

In subtask II the order is reversed: The mean absolute angle mismatch is significantly smaller

	No panning				Panning			
	static (deg)	*de-pan.* (deg)	p_{Cond}	p_{Subj}	*static* (deg)	*de-pan.* (deg)	p_{Cond}	p_{Subj}
Mean	15.9	22.4	0.0110	0.4315	8.8	5.4	0.0002	0.0952
Median	0	30	0.0133	0.3714	7	4	0.0000	0.1972

Table 5.3: *p-Values speaker localisation.* In subtasks I and II, a significant effect of the recording condition, i.e. *static* or *de-panned*, on the localisation performance is found. No significant difference between subjects is found in either subtask.

	p_{Cond}	p_{Pan}	p_{Int}
Two-way ANOVA	0.2547	0.0000	0.0003

Table 5.4: *Effect of recording condition and panning on localisation performance.* Whilst there is a significant effect of the panning, there is no significant effect of the recording condition. There is significant interaction between both effects.

in the *de-panned* condition (5.4°) than in the *static* condition (8.8°), $F(1, 12) = 14.42$, $p_{Cond} = 0.0002$). The Friedman analysis indicates a significantly smaller median with the *de-panned* recording (4.0°) than with the *static* recording (7.0°), $\chi^2(1) = 16.83$, $p_{Cond} = 0.0000$. Again, no significant difference between subjects is found (ANOVA: $F(1, 12) = 1.59$, $p_{Subj} = 0.0952$, Friedman: $\chi^2(12) = 15.87$, $p_{Subj} = 0.1972$).

The comparison of subtask I and II reveals a significant difference of the localisation performance between all four tested conditions, i.e. *static* and *de-panned* with and without panning (ANOVA: $F(3, 12) = 32.11$, $p_{Cond} = 0.0000$, Friedman: $\chi^2(3) = 14.14$, $p_{Cond} = 0.0027$). To determine which means are significantly different, a multiple comparison post test with Tukey-Kramer correction is applied to the results of the ANOVA [Motulsky, 1995]. The test essentially compares all group means to find significant differences. The standard deviations of all samples are pooled, to account for the fact that by comparing multiple means at a certain significance level, chances to find significant differences and to mistakenly reject the null hypothesis (Type I error) increase with the number of comparisons. Applying the Tukey-Kramer test to the results of Task I indicates a significantly larger mean absolute angle mismatch in the *de-panned* condition without panning than in any of the other conditions. This again falsifies hypothesis I. Both conditions in subtask I yield a significantly larger mean absolute angle mismatch than the conditions in subtask II. This proves hypothesis II, i.e. that panning improves the localisation performance.

To separate the impact of the recording condition and the panning on the localisation performance, a two-way ANOVA is performed on the data, with the recording condition as one factor and the subtask as the other. The two-way ANOVA provides the advantage of indicating whether there is interaction between the tested factors, i.e. whether there is a synergistic effect. The results show a significant effect of the panning on the localisation performance, $F(1, 1) = 80.79$, $p_{Pan} = 0.0000$. This proves hypothesis II, that panning improves the localisation performance. The test condition, i.e. whether the *static* or *de-panned* recording is used, has no significant impact on the performance, $F(1, 1) = 1.30$, $p_{Cond} = 0.2547$. This contradiction with the results of the analysis of subtasks I and II separately is explained through the significant impact of interaction, which reveals a synergistic effect, $F(1, 1) = 13.48$, $p_{Int} = 0.0003$. In other words, the impact of the test condition depends upon whether panning is used or not: If no panning is used, de-panning has a negative effect on the localisation performance. If head

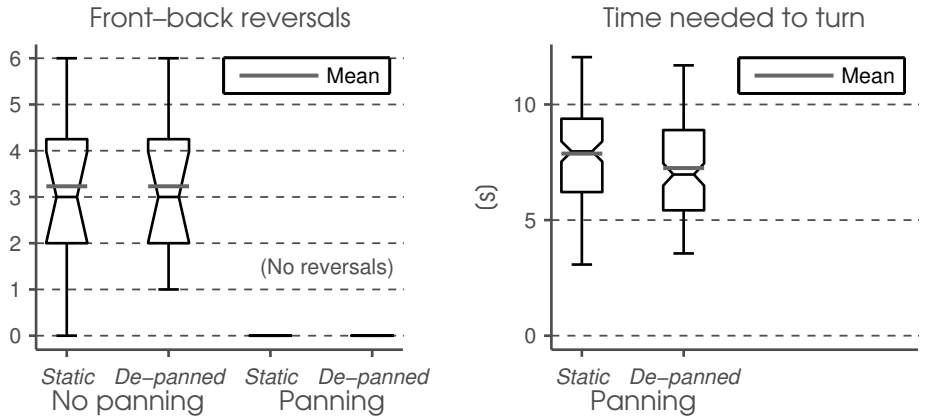


Figure 5.6: *Front–back reversals and time needed to turn towards speakers.* More than 80 percent of front–back reversals stem from mistakenly perceiving a source to be in the back. No front–back reversal occurred with panning enabled, in either condition. The time needed to turn towards a speaker was limited to about 12 seconds by the duration of the speech sample.

	Front–back reversals			Time needed to turn				
	static	*de-pan.*	p_{Cond}	*static* (sec)	*de-pan.* (sec)	p_{Cond}	p_{Subj}	p_{Int}
Mean	3.2	3.2	1.0000	7.9	7.2	0.0023	0.0000	0.0398
Median	3	3	0.7630	8.0	7.0	0.0022	0.0000	

Table 5.5: *p-Values front–back reversals and time needed to turn.* The front–back reversal rates in subtask I are normally distributed and have equal mean, therefore no difference is found. The time needed to turn towards the speaker differs significantly between the test conditions. However, a significant difference between subjects and a significant interaction cast doubt on this result.

tracking and panning are enabled, however, the *de-panned* recording yields better results. The results are summarised in table 5.4.

The mean number of front–back reversals in subtask I is equal in both tested conditions: 3.2 out of 10 (cf. fig. 5.6). This is close to chance level, as 2 out of the 10 tested directions were at the extreme right ($-90°$), where no reversal can occur. Most of the reversals (83 percent in the *static* and 85 percent in the *de-panned* case) occurred when a source was mistakenly perceived to be in the back. The chance of this kind of error is increased by the fact that frontal source directions prevailed in the test.

To check whether the number of front–back reversals is normally distributed, a Lilliefors normality test [Lilliefors, 1967] is applied to the data. It reveals that the null hypothesis, i.e. that the data comes from a normal distribution, cannot be rejected for both the *static* ($p = 0.4932$) and the *de-panned* condition ($p = 0.3287$). An F-test indicates that the null hypothesis of equal variances cannot be rejected for the two samples, $p = 0.8227$. A paired t-test fails to reject the null hypothesis, i.e. that the observations from the *static* and *de-panned* condition are taken from distributions with equal mean (cf. table 5.5), $p = 1.0000$. The de-panning does not have an effect on the number of front–back reversals. A Friedman analysis confirms this result, $\chi^2(1) = 0.09$, $p = 0.7630$. The difference between test subjects is not analysed, as only two measures per subject were collected in subtask I.

In subtask II, no front–back reversal was observed. All test subjects managed to correctly identify whether a source was in front or in the back in both tested conditions in all trials. This further supports hypothesis II, i.e. that panning improves localisation performance, as it

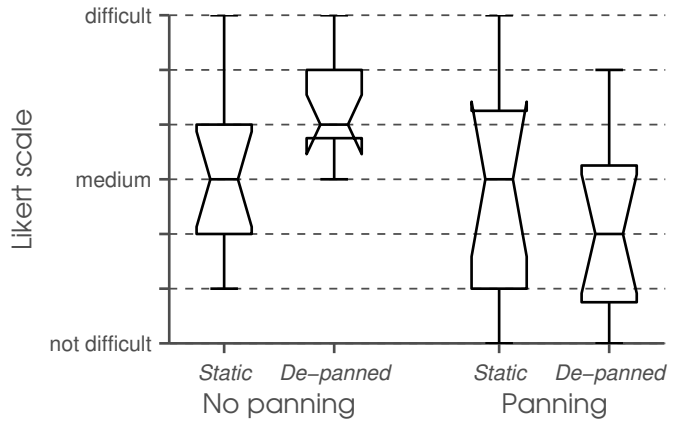


Figure 5.7: *Perceived difficulty.* Localising speakers in the *de-panned* condition without panning is perceived to be significantly more difficult than in the *de-panned* condition with panning enabled.

	No panning	Panning	Comparison
	p_{Cond}	p_{Cond}	p_{Cond}
Friedman	0.0067	0.2059	0.0023

Table 5.6: *Perceived difficulty.* In subtask I, a Friedman analysis indicates the mean rank of the perceived difficulty in the *de-panned* case to be significantly larger than in the *static* case. No significant difference is found in subtask II. Comparing both subtasks, the Friedman analysis indicates a significant difference between the test cases. A Tukey-Kramer post test reveals the *de-panned* case without panning to be perceived as significantly more difficult than the *de-panned* case with panning.

significantly reduces front–back reversals compared to a scenario without panning.

The mean time needed to turn towards the speaker in subtask II is 7.9 seconds with the *static* recording, and 7.2 seconds with the *de-panned* recording (cf. fig. 5.6). Though a two-way ANOVA indicates the difference to be significant, $F(1, 12) = 9.53$, $p = 0.0023$, it also reveals a highly significant difference between subjects, $F(1, 12) = 14.44$, $p = 0.0000$, and a significant interaction between the impacts of the tested condition and subject, $F(1, 12) = 1.86$, $p = 0.0398$ (cf. tab 5.5). A Friedman analysis yields similar results for the median time needed to turn towards the speaker: Both the difference between the *static* and *de-panned* condition, $\chi^2(1) = 9.37$, $p = 0.0022$, and between test subjects, $\chi^2(12) = 107.9$, $p = 0.0000$, is highly significant. Due to the significant difference between subjects, and the significant interaction revealed by the two-way ANOVA, it is questionable to conclude that the time needed to turn towards a speaker is shorter on average when using a *de-panned* recording than with a *static* recording. Task I of the user study was mainly designed to analyse the localisation performance in terms of accuracy, not speed. The test subjects were not instructed to take decisions fast. As the same test phrase was used throughout Task I, test subjects became familiar with the speech sample and its duration and timed their answers accordingly.

The Likert scores are a measure for the perceived difficulty of the tests in Task I. Gardner and Martin argue for the interpretation of the Likert scale as an ordinal, rather than an interval scale [Gardner and Martin, 2007]. Furthermore, the authors point out the possibility of subjects' responses being nonlinear and biased by the interpretation of the phrases used in the scale. The authors suggest avoiding parametric analyses. Instead, their nonparametric equivalents should be used. To compare the perceived difficulty in each subtask, a Friedman analysis is performed on the medians of the perceived difficulty (see fig. 5.7). The null hypothesis is rejected in the first subtask, indicating that localisation in the *static* condition is perceived to be significantly

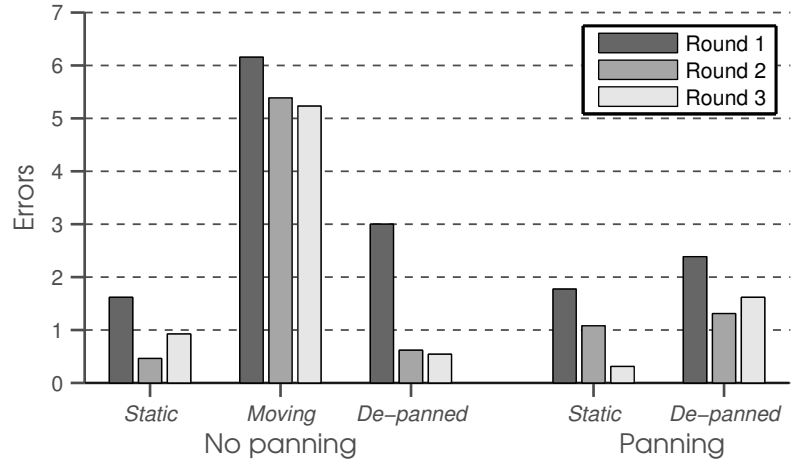


Figure 5.8: *Mean error rates.* The mean error rates in the *moving* condition are significantly higher in all three rounds than in all other conditions. The performance of test subjects significantly improved from the first round to the second.

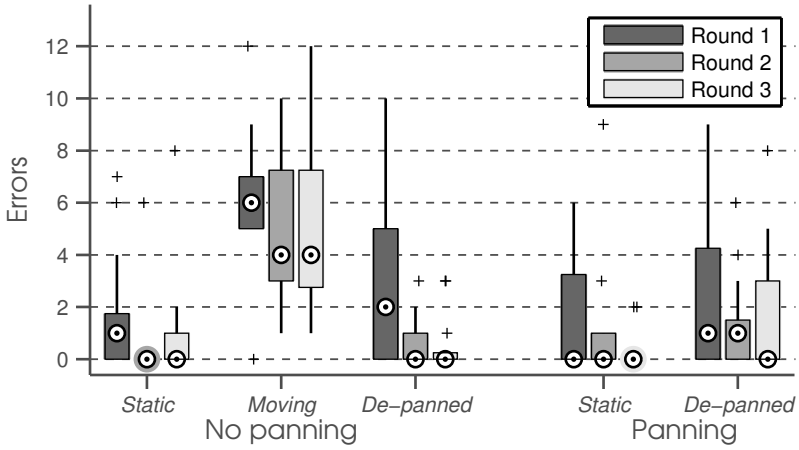


Figure 5.9: *Median error rates.* In rounds II and III, the median error rates in the *moving* condition are significantly higher than in all other conditions.

less difficult than in the *de-panned* condition, $\chi^2(1) = 7.36$, $p = 0.0067$. No significant difference between conditions is found in subtask II, $\chi^2(1) = 1.6$, $p = 0.2059$. When comparing both subtasks, the Friedman analysis indicates a significant difference between all tests in Task I, $\chi^2(3) = 14.5221$, $p = 0.0023$. A post test with Tukey-Kramer correction reveals the speaker localisation in the *de-panned* case without panning to be perceived significantly more difficult than speaker localisation in the *de-panned* case with panning. The results are summarised in table 5.6.

5.3.3 Task II – speaker segregation

Error rates

The performance in Task II is measured in terms of the number of correctly identified speaker turns. An error occurs each time a turn of the speaker in question is missed or one of the three other speakers is mistaken for the speaker in question. The total error rate is the sum of the missed and the wrongly identified turns. In each test case, subjects had to identify 5 turns of the speaker in question, whereas 15 turns were from the other three speakers. Thus, a total of 5

	No panning			Panning			Comparison		
	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3
ANOVA	0.0033	0.0000	0.0001	0.3985	0.5845	0.1025	0.0003	0.0000	0.0000
Friedman	0.0242	0.0000	0.0001	0.0588	0.3173	0.0652	0.0131	0.0000	0.0000

Table 5.7: *p-Values error rates.* The error rates are compared between the tested conditions in both subtasks separately and in combination. p_i indicates the p-value for rejecting the null hypothesis in round i. Significant differences are found only when comparing the *moving* condition to other conditions. A Tukey-Kramer post test reveals that in rounds II and III the *moving* condition leads to significantly higher mean and median error rates than all other conditions.

turns could be missed and 15 marked wrong in each test case, yielding a maximum of 20 errors per tested condition.

In subtask I, the speakers are not registered with the test subject's environment. Referring to hypothesis III, performance is assumed to be better when interaural cues are present in the recording, aiding the subjects to segregate the speakers. In the *moving* condition, recorded with just one loudspeaker for all speakers, no segregation based on interaural cues is possible. The test subjects thus have to identify the speaker in question based on his voice, which is assumed to be more difficult than identifying him based on his direction. This is expected to manifest itself in higher error rates in the *moving* condition than in the *static* or *de-panned* condition, where interaural cues are present.

Subtask II provides test subjects with the possibility to turn towards the speaker in question. According to hypothesis IV, this is assumed to improve the speaker segregation, resulting in lower error rates compared to subtask I without panning. Both subtasks are repeated three times, to identify learning effects. It is expected that performance improves with each round, as test subjects become acquainted with the speaker voices and the test procedure. To minimise learning effects within each round, the tested conditions are shuffled (cf. table 5.2).

Perceived difficulty

Test subjects marked the perceived difficulty of each tested condition in each round on a seven-step Likert scale (cf. fig. 5.4). The scores are interpreted as a ranking of the tested conditions by the test subjects in terms of their relative perceived difficulty. The *moving* condition is expected to be rated more difficult than the other conditions.

Statistical analysis

The performance is analysed for each subtask separately and in comparison, for each round. To analyse improvement due to repetition, performance is also compared between rounds. The data is analysed using a two-way ANOVA and a Friedman analysis.

The most striking result of Task II is the speaker segregation performance with the *moving* recording in subtask I (cf. fig. 5.8 and fig. 5.9). As expected, test subjects had difficulties identifying the speaker in question in a recording lacking interaural cues. Error rates in this condition where high in all three rounds. Both ANOVA and Friedman analysis indicate a significant difference between the mean and median error rates, respectively, among the three tested conditions of subtask I, in all three rounds. In round I, a post test with Tukey-Kramer correction of the ANOVA results indicates that the mean error rate is significantly higher in the *moving* condition than in the *static* condition. In rounds II and III, the *moving* recording yields significantly higher mean error rates than both the *static* and *de-panned* recording. No significant difference of the mean error rates is found between *static* and *de-panned* conditions

	p_{Cond}	p_{Pan}	p_{Int}
Two-way ANOVA	0.1287	0.5468	0.6449

Table 5.8: *Effect of recording condition and panning on speaker segregation performance.* No significant impact of either the recording condition, the head tracking and panning, or interaction effects are found in Task II. The *moving* condition is excluded from this comparison.

	p_{Cond}	p_{Rnd}	p_{Int}
Two-way ANOVA	0.0000	0.0031	0.8699

Table 5.9: *Effect of test round on speaker segregation performance.* A highly significant difference between the test rounds is found, indicating that repetition of the test has a significant impact on the performance of test subjects (see also fig. 5.8). A Tukey-Kramer post test indicates a significant improvement from round I to round II. No significant interaction between the test condition and the test round is found.

in the three rounds. Applying the Tukey-Kramer post test to the Friedman analysis indicates the median error rates for the *moving* condition to be significantly higher than for the *static* and *de-panned* condition, in rounds II and III. No significant difference is found between the *static* and *de-panned* condition. This proves hypothesis III, i.e. that speaker segregation improves if interaural cues are present to distinguish speakers spatially. It also indicates that de-panning does not deteriorate the segregation performance significantly compared to the *static* recording.

In subtask II, neither the ANOVA nor the Friedman analysis indicates a significant difference between the tested conditions *static* and *de-panned*. This is in accordance with the results of subtask I.

When comparing both subtasks, ANOVA and Friedman analysis indicate a significant performance difference between the tested conditions in all three rounds. A Tukey-Kramer post test is applied to the ANOVA results. It indicates that the *moving* condition leads to significantly higher mean error rates compared to the other conditions, both with and without panning, in all three rounds. No significant difference is found between the other four conditions, i.e. *static* and *de-panned* with and without panning. Similar conclusions can be drawn from a Tukey-Kramer post test of the Friedman analysis results. Performance in the *moving* condition is significantly worse than in other conditions. In rounds II and III, test subjects performed significantly worse in the *moving* condition than in all other conditions. No significant difference is found between the mean ranks of the *static* and *de-panned* conditions in both subtasks in all three rounds. Thus, the impact of interaural cues on the segregation performance is found to be significant in test cases with and without head tracking and panning, producing lower error rates compared to a test case without interaural cues. The results are summarised in table 5.7.

To analyse the impact of panning and the recording condition on the performance, a two-way ANOVA is performed. The *moving* condition is excluded from this analysis, as it is only tested without panning. The analysis indicates no significant impact of the recording condition, $F(1,1) = 0.36$, $p_{Cond} = 0.5468$, or panning, $F(1,1) = 2.33$, $p_{Pan} = 0.1287$, on the segregation performance (see table 5.8). There is no significant interaction between the two factors, $F(1,1) = 0.21$, $p_{Int} = 0.6449$. Thus, hypothesis IV, i.e. that panning improves the segregation performance, is falsified.

A two-way ANOVA is applied to investigate upon the impact of the test round on the segregation performance. It is found to be significant, $F(2,4) = 5.96$, $p_{Rnd} = 0.0031$, indicating the presence of learning effects. A Tukey-Kramer post test reveals significantly higher mean error

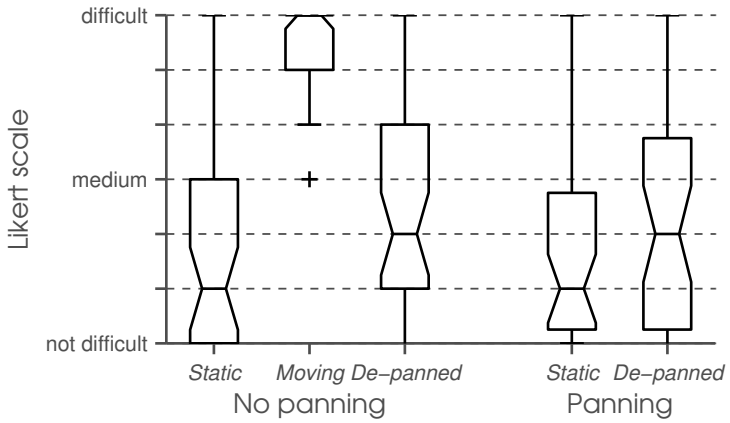


Figure 5.10: *Perceived difficulty.* The *moving* condition is perceived to be significantly more difficult than all other conditions.

rates in round I than in rounds II and III. Thus, after the first trial, segregation performance of the test subjects improved. No significant interaction effects were found between the test round and the test condition, $F(2, 8) = 0.48$, $p_{Int} = 0.8699$. The improvement after round I is independent of the test condition. No significant difference is found between the mean error rates in rounds II and III, indicating that performance did not significantly improve after the second round. In fact, some test subjects performed worse in round III than in round II in the same test condition. As task II was quite demanding in terms of the concentration required, this effect may partly be attributed to increasing fatigue of the test subjects. The results are summarised in table 5.9.

The Likert scores provide a subjective measure for the perceived difficulty of the test conditions in Task II (cf. fig. 5.10). The scale is interpreted as an ordinal scale, and a nonparametric analysis is applied to the scores of each subtask separately and in comparison. A Friedman analysis indicates a significant difference between the test conditions in subtask I, $\chi^2(2) = 53.12$, $p_{Cond} = 0.0000$. Due to missing entries, the data of one test subject was excluded in this analysis. A Tukey-Kramer post test reveals that the mean rank of the perceived difficulty of the *moving* condition is significantly larger than the mean ranks of both the *static* and *de-panned* condition. No significant difference between the *static* and *de-panned* condition is found either in subtask I or subtask II. When comparing both subtasks, the Friedman analysis indicates a significant difference between all five test cases, $\chi^2(2) = 60.90$, $p_{Cond} = 0.0000$. Due to missing entries, the data of two test subjects was excluded. A Tukey-Kramer post test reveals the *moving* condition to yield a significantly larger mean rank of the perceived difficulty than all other conditions. No significant difference is found between the other conditions. This implies that test subjects perceived the speaker segregation to be significantly more difficult in the *moving* condition, where no interaural cues were present in the recording to spatially segregate speakers, than in the *static* and *de-panned* case. The results are summarised in table 5.10.

5.3.4 Comments of test subjects

One of the most stated problems in the speaker localisation task was inside-the-head locatedness (IHL). Test subjects reported difficulties to localise sound sources that were straight ahead, as they often lacked externalisation. This was said to be confusing. Some test subjects pointed out a lack of depth in the *de-panned* recording. Whereas the sound sources appeared to be positioned on a "clear circle" in the *static* recording, in the *de-panned* recording they seemed to be positioned on a "straight line", ranging from the far left to the far right of the listener. This made it more difficult to map sources to a virtual circle than in the *static* case. One subject

	No panning	Panning	Comparison
	p_{Cond}	p_{Cond}	p_{Cond}
Friedman	0.0000	0.0976	0.0000

Table 5.10: *Perceived difficulty.* In subtask I, Friedman analysis and Tukey-Kramer post test indicate the *moving* condition to be perceived significantly more difficult than the other two conditions. No significant difference is found in subtask II. Comparing both subtasks, the Friedman analysis indicates a significant difference between all four test cases, with the *moving* condition being perceived significantly more difficult than all other conditions.

commented on the *static* recording as having "more depth" and a "thicker sound" and hence as being more pleasant than the *de-panned* recording.

When asked to turn towards the speaker, one subject stated that the approximate direction of the speaker could be determined immediately, but turning towards him required some "searching" process, to "balance" the sound on both ears. One subject commented on this task as being "fun", and stated that closing the eyes made the task easier.

Some test subjects named the hiss in the recordings as an additional localisation cue. Before the voice from a speaker could be heard, the hiss preceding it gave a hint as to his direction. The same cue was also used by some subjects in the speaker segregation task to identify the speaker in question. Most test subjects pointed out difficulties to distinguish speakers in the *moving* recording. Some test subjects said they became more acquainted with the voice of the speaker in question towards the end, and managed to segregate the speakers based on their accents or articulations. In the other test conditions test subjects reported to rely mainly on the direction when segregating different speakers.

Only one test subject named the head tracking as a helpful factor in the speaker segregation task. Another subject stated that turning towards the speaker in question made the segregation task indeed more difficult, as it was easier to localise and identify a speaker a bit off the centre. Yet another test subject pointed out IHL as the main cue for segregating the speakers: After turning towards the speaker in question, he was not externalised anymore, which clearly separated him from the other speakers in the recording.

5.4 Discussion

The *static* case, made with several loudspeakers at fixed positions, and recorded without head movement, represents the "ideal" case of a binaural recording, preserving the spatial cues of all speakers. In the *de-panned* recording, simulating a situation where the KAMARA headset user moves the head during the recording, interaural cues are restored by compensating for the head movements through the de-panning algorithm. If no panning is applied during playback to register the recorded speakers with the environment, the *de-panned* recording yields a significantly larger mean and median absolute angle mismatch between the perceived and the actual direction of the recorded speakers than the *static* recording. This indicates that the de-panning algorithm cannot fully restore the spatial cues contained in the recording. Hypothesis I, i.e. that the localisation performance with a *de-panned* and *static* recording is equal, is thus falsified. Test subjects perceived localisation with the *de-panned* recording to be significantly more difficult than with the *static* recording. This may be related to the fact that some test subjects perceived the speakers in the *de-panned* recording to be positioned on a line, whilst in the *static* recording they appeared to reside on a circle around the listener, with distinct directions.

With head tracking and panning enabled, the mean and median absolute angle mismatch decreased significantly, which proves hypothesis II. Test subjects localised speakers significantly

more accurately by turning towards them than by indicating their directions. The reduced localisation blur achieved by facing the virtual speakers implies that registering virtual sources with the environment through panning may lead to better spatial separability of the sources. This is seen as a major benefit in a telecommunication scenario. When comparing the two test conditions with panning enabled, the *de-panned* condition leads to a significantly better localisation performance. As test subjects turn towards the de-panned speaker, their head orientation approximately matches the head orientation during the recording, therefore nearly unprocessed audio is delivered to the test subjects (c.f. fig. 4.6d). Turning towards a virtual source recorded off the centre, as in the *static* case, increases the localisation blur significantly, as the panning algorithm fails to fully restore the spatial cues.

No effect of the recording condition on the number of front–back reversals is found. The *de-panned* recording does not yield a higher rate of reversals than the *static* recording. We assume front–back reversals to be mainly a result of the ambiguity of interaural cues in general, not of the processing involved in generating them. A more striking finding, however, is the fact that with head tracking and panning enabled, no front–back reversal occurred in any of the 260 observations. This is a strong argument for hypothesis II, i.e. that panning improves the localisation performance. When a test subject turns the head to search for the virtual sound source, the interaural cues change accordingly, indicating unambiguously whether the source is in front or in the back. Even test subjects without any prior experience with spatial audio and head tracking instinctively interpreted these motional cues correctly.

The time needed to turn towards the speaker is similar in both conditions, and presumably longer than in natural listening conditions. As one test subject stated, the direction of the virtual speaker could be determined immediately, but some "searching" was needed before confirming the direction. This searching manifested itself in subjects "overshooting" the correct direction multiple times to either side, before settling to the final direction estimate. This behaviour can also be observed in listening tests where subjects are asked to localise real sound sources [Blauert, 1996]. The panning applied to the binaural recording supports this instinctive reaction to localise sound sources. However, the responsiveness of the panning algorithm and the interaural cues generated by it could still be improved to match or approximate the natural listening situation. This should minimise and hence accelerate the necessary "searching" process. It should be pointed out, however, that the test was not designed to measure the localisation speed. This also explains the relatively slow response of test subjects. In fact, in Task II most test subjects were able to localise the speaker and turn towards him in much shorter time. The test phrase in this case was only about two seconds long, thus test subjects were driven to react quickly, and managed to localise the speaker much faster than in Task I.

The results obtained from the speaker segregation task prove the importance of interaural cues to segregate multiple speakers, thus proving hypothesis III. The *moving* condition, which contains little or no interaural cues to separate speakers, leads to significantly higher mean and median error rates than the *static* and *de-panned* cases, which contain natural or algorithmically restored interaural cues. Even after being presented with the same recording for the third time in round III, the median error rate of test subjects when trying to identify the 5 turns of the speaker in question is 4. Some subjects stated their choices in the *moving* case to be based on pure guessing, others marked no turn at all. For all other cases the median error rate in round III drops to 0, indicating that more than 50 percent of the test subjects managed to identify all speaker turns correctly. The result is supported by the perceived difficulty, with the *moving* condition rated significantly more difficult than all other conditions. This underlines the importance of interaural cues to segregate multiple speakers.

No significant differences are found between the *static* and *de-panned* case regarding the speaker segregation. Whilst the de-panning has a negative effect on the speaker localisation, it does not deteriorate the speaker segregation performance. Compared to an unprocessed binaural recording with no or misleading interaural cues, such as the *moving* recording, de-panning

significantly improves speaker segregation, and theoretically yields the same performance as the ideal case of a *static* recording devoid of head movements.

With head tracking and panning enabled, no significant difference is found in the speaker segregation performance of the *static* and *de-panned* case. When turning towards a speaker, in case of the *de-panned* recording, the test subject is presented with a nearly unprocessed recording, which is considered to be the ideal case. When turning towards a speaker in the *static* recording, however, panning has to be applied, in case the speaker was not recorded right in front. The panning basically adjusts the interaural cues to match those of a binaural recording made facing the speaker. As there is no significant difference between the panned *static* recording and the unprocessed *de-panned* recording, the panning algorithm does not significantly deteriorate the speaker segregation performance.

To summarise these results: De-panning restores the interaural cues of each speaker in a recording, hence significantly improves the segregation performance. The performance is not significantly different from the ideal case of a *static* recording. The localisation performance of a *de-panned* recording, however, is significantly worse than that of a *static* recording. Panning, on the other hand, adjusts interaural cues to register speakers with the environment of the listener, yielding a segregation performance not significantly different from the ideal case of an unprocessed recording.

The segregation performance improved significantly from round I to round II. This is attributed to the fact that test subjects became acquainted with the test procedure and the a priori unfamiliar voices of the speakers used in the test. No significant improvement from round II to round III is found, indicating that learning effects vanish after round I.

The comments of test subjects suggest that there are still some issues of the algorithms related to audio artefacts. The lack of externalisation of the processed audio was a common problem, resulting in the *de-panned* recording being perceived as "flat" or suffering from IHL. It was also stated that the unprocessed *static* recording was preferred over the *de-panned* recording. Thus, the de-panning seems to have a negative effect on the perceived audio quality. The fact that IHL was also stated to make the localisation task more difficult indicates that externalisation (or the lack thereof) might also affect the performance of test subjects. On the other hand, audio artefacts introduced by the algorithms served as additional cues. One test subject reported to identify the speaker in question in the segregation task based on IHL. When turning towards a speaker, the panning algorithm mixes high frequency content of both channels, reducing interaural decorrelation and thus externalisation. IHL can thus serve as a cue to detect a speech signal that is panned to the front. Another artefact used as a cue was the background hiss. Subjects often stated to rely on the hiss preceding each speaker in the *de-panned* recording to determine his direction. The reason for this "spatialised" hiss is that de-panning is applied not only to the speech signal but inevitably also to the monaural wideband background noise in the recording. The de-panning introduces interaural cues to this background noise, which can be interpreted by test subjects as localisation cues. As the recording for the *de-panned* condition was made with just one loudspeaker right in front of the KAMARA user, the speakers in the unprocessed recording contain little or no interaural cues. Therefore, after de-panning, speech signal and background noise are enhanced with almost identical interaural cues. This explains, how localising the background hiss could indeed serve as a valid cue for determining the direction of the speech signal. A possible solution to this problem could be to remove background noise before processing the recording, or mask the processed noise with monaural noise.

Chapter 6

Summary and conclusions

6.1 Summary

Telecommunication describes the act of exchanging thoughts and interacting over distance. Mobile phones and VoIP softwares are conventional tools to enable telecommunication. Previous studies have shown face-to-face communication to outperform telecommunication systems in various aspects [Billinghurst et al., 2002, Lindeman et al., 2009]. In an attempt to approach the performance and naturalness of face-to-face communication, the applicability of audio augmented reality (AAR) to such systems is studied. Means are suggested for its implementation, in terms of the theoretical background, the enabling technologies and the necessary audio processing. As an example application, an AAR-enhanced teleconference scenario is devised. In the test scenario, a user is presented with the binaural recording of a remote meeting, recorded at the ears of one of the remote participants via a binaural headset.

Before playing the binaural recording back to the user as an overlay of the real acoustic environment, the recording needs to be processed. Head movements on both the local and the remote end distort the perceived directions of the recorded sound sources. Algorithms are presented to compensate for these head movements. The de-panning algorithm adjusts the interaural cues of a binaural recording, restoring the directions of the recorded sounds. The panning algorithm registers the recorded sound sources with the environment, by adjusting the interaural cues according to the head movements of the listener. To reduce computational complexity and processing artefacts, a method is presented to merge the algorithms to a single processing stage.

To evaluate the performance of the proposed algorithms, a user study was conducted. The study was designed study the impact of the algorithms

1. on the localisation and

2. on the segregation

of a virtual sound source. Objective and subjective measures were obtained in a *within–subjects* test from 13 participants of the study.

6.2 Conclusions

The KAMARA headset provides a simple and effective way to integrate spatial audio into a telecommunication system. The binaural recording preserves the spatial cues of recorded sound sources, yielding a listening experience similar to the natural auditory perception of an environment. Head movements distort the spatial cues and thus the perceived directions of the recorded sound sources. The proposed de-panning algorithm successfully restores the perceived

directions. Test subjects were able to localise a speaker in a de-panned binaural recording, even though the de-panning increased the localisation blur significantly. The localisation accuracy, in terms of the mean absolute angle mismatch, was about 7° worse than in the ideal case – an unprocessed recording devoid of any head movements.

The panning algorithm significantly improved the localisation performance of test subjects. Panning adjusts the binaural playback according to head movements of listeners, allowing them to localise a virtual source by turning towards it. The user study results prove that the panning algorithm was successful in registering the binaurally recorded source with the environment, which provides a natural way of embedding virtual audio content into the auditory perception. The test subjects interacted with the system intuitively, using head rotations to "search" for the virtual source. No significant performance difference was found between subjects, even though about half of the test subjects had no previous experience with spatial audio or head tracking. These results imply that the proposed system is suitable also for "naïve" users. By registering the virtual sources with the environment, no front–back reversal occurred, i.e. all test subjects correctly determined whether a source was in front or in the back, which is a remarkable result for a binaural localisation task.

Interaural cues are shown to affect the ability of test subjects to segregate multiple virtual sources. The segregation performance was significantly better with recordings containing interaural cues than with a recording with no cues. In case of misleading spatial cues, i.e. arbitrary changes in the perceived directions of the sources due to head movements, the performance is expected to be even worse. No significant difference was found between the recordings containing interaural cues, regardless of how these cues were obtained. The de-panned recording, in which the spatial cues were algorithmically restored, did not lead to a significantly worse performance than the ideal case, an unprocessed binaural recording of sound sources separated in space, devoid of head movements. Thus, even though de-panning yielded a significantly worse localisation performance compared to the ideal case, it did not cause a significantly worse segregation performance. Yet, with de-panning, both the localisation and the segregation performance are better compared to an unprocessed binaural recording containing no or distorted interaural cues. In a telecommunication scenario, the de-panning algorithm restores the perceived directions of speakers and enhances the ability of a listener to segregate the participants of a meeting. This is assumed to improve the listening comfort and the ability to follow a remote conversation, which is a major argument for the use of AAR in a telecommunication scenario.

The simplicity of the proposed de-panning and panning algorithms has several advantages. The algorithms run on a standard PC, with a responsiveness that was found to be sufficient for the test scenario. System lag was an issue only in the case of fast head movements, due to the limited updated rate of the head tracking device. The processing is based on simple ITD and head shadowing models, hence the system does not require an HRTF dataset. This makes it transferable and relatively robust against individual HRTF variations. On the downside, the current implementation provides few possibilities for adjustments of the processing parameters, besides the head radius and a wetness factor to adapt the ILD correction to the room reverberance.

Transmitting a binaural recording of one's environment through a KAMARA headset is a simple yet effective way to share auditory perception over distance. Tackling issues related to head movements with the algorithms proposed in this work allowed both experienced an inexperienced users to localise virtual sources in a binaural recording. This significantly improved the ability of test subjects to segregate multiple sources, with minimal requirements in terms of the computing resources and hardware equipment. The proposed implementation of an AAR system might serve as a valuable tool to enhance existing telecommunication systems and help overcome the gap to face-to-face communication.

6.3 Future outlook

The proposed implementation is a proof-of-concept showing the potential of implementing AAR for telecommunication through panning and de-panning of binaural recordings. Various aspects of the setup could be refined. The simple ILD and ITD models underlying the algorithms do not account for individual differences among subjects. Parametrisation of the filters could yield better results in terms of the localisation accuracy of virtual sources. More accurate ILD modelling could minimise motional cues, by accounting for the fine structure of the spectral changes cause by head movements.

Improvements could also be made to the sound quality of the system. Test users reported audio artefacts, introduced by the de-panning and panning algorithms. To tackle the problem of spatialised background hiss, noise removal and masking techniques could be considered. A more sophisticated algorithm mixing the high frequency channels of the input signals should be devised to preserve interaural decorrelation and avoid the lack of externalisation of virtual sources, when panning is applied. Some test subjects of the user study reported lags with fast head movements. Improving the responsiveness of the system by minimising the system delay and increasing the update rate of the head tracking device could tackle this issue. This would be of particular benefit in applications requiring immediate feedback to user interaction, for example in an AAR navigation scenario.

A central aspect of AAR is the combination of real and virtual auditory content. An issue further to be investigated upon is the mixing of a binaural recording from a remote end and the pseudo-acoustic environment, perceived through a KAMARA headset. Registering binaurally recorded sound sources with the environment, as proposed in this work, is the first step towards a seamless integration of remote sound sources into the auditory perception. A fully embedded virtual environment is assumed to maximise the sense of presence and immersion of a user, and thus the communication performance. A user study could be conducted analysing the performance of the system in collaborative tasks, compared to face-to-face collaboration.

A limitation of the current system is that the positions of sound sources need to be known beforehand. To obtain a self-contained system, the speaker identification and localisation algorithms developed at the Department of Signal Processing and Acoustics in the course of this project could be integrated. The interoperability of the algorithms has been shown in a demonstration setup. An issue not considered in this work is how to process the KAMARA user's own voice, as it contains little or no interaural cues, and is prone to IHL. A possible approach is to process it with a BRIR acquired on the fly via a clap or finger snap [Gamper and Lokki, 2009]. Finger snaps could also be used as an intuitive way to define the desired perceived directions of remote speakers.

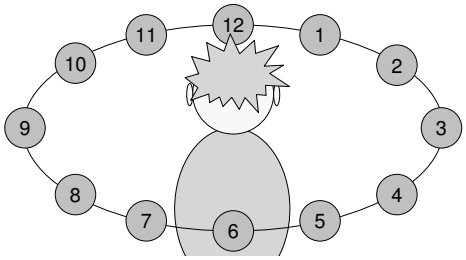
Currently, the system supports only one virtual source at a time. If multiple simultaneous sound sources are present in the recording, the algorithms fail. Thus, one of the biggest challenges for future developments is the automatic segregation and (de-)panning of multiple sound sources. The sources could for instance be separated by analysing the direction of sound segments in frequency bands, a technique employed in Directional Audio Coding (DirAC) [Pulkki and Faller, 2006]. A similar approach could tackle the problem of motional cues caused by head movements in the presence of strong room reflections.

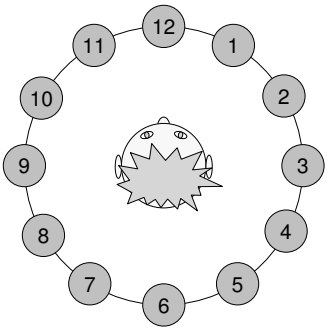
The example of a telecommunication scenario proves the potential of AAR to enhance the performance of the human auditory system as an information channel. As the human is a multisensory being, enhancing the current implementation with multimodal feedback, e.g. through the senses of vision or touch, seems to be a promising way to improve overall performance of the system. Integration of the AAR system with existing AR systems is left to future research.

Appendix A

User study questionnaire

A.1 Task I – speaker localisation





Without panning

a) Write down the direction of the speaker you hear (1–12).

How difficult was this task? ☐ ☐ ☐ ☐ ☐ ☐ ☐
(not difficult) (medium) (difficult)

b) Write down the direction of the speaker you hear (1–12).

How difficult was this task? ☐ ☐ ☐ ☐ ☐ ☐ ☐
(not difficult) (medium) (difficult)

Comments: _____

With panning

c) Look at the speaker you hear.

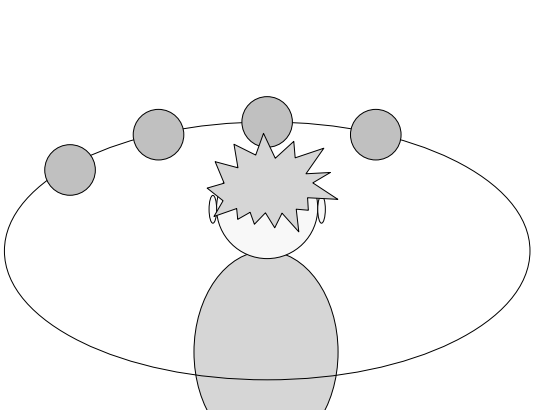
How difficult was this task? ☐ ☐ ☐ ☐ ☐ ☐ ☐
(not difficult) (medium) (difficult)

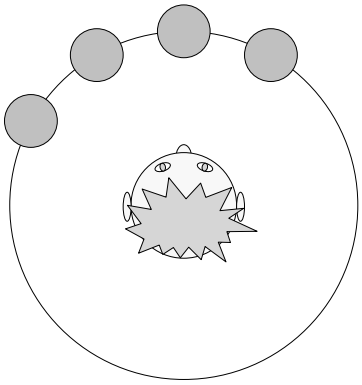
d) Look at the speaker you hear.

How difficult was this task? ☐ ☐ ☐ ☐ ☐ ☐ ☐
(not difficult) (medium) (difficult)

Comments: _____

A.2 Task II – speaker segregation



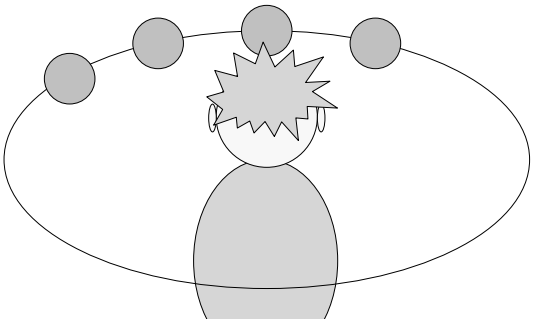


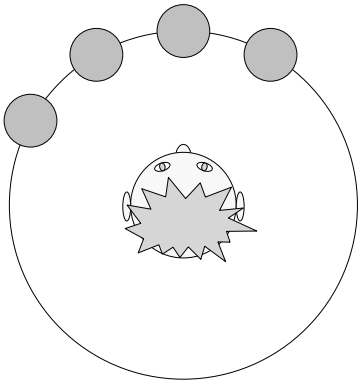
Without panning

Mark the words that you hear from the first speaker.

1: agency	1: attitude	1: economically
2: coexist	2: approach	2: museum
3: actually	3: credit	3: mediocrity
4: colleges	4: careful	4: helpless
5: divorced	5: eternal	5: dark
6: curiosity	6: academic	6: grains
7: dry	7: doors	7: enough
8: damage	8: compile	8: harms
9: company	9: discount	9: famous
10: coeducational	10: cheese	10: developed
11: archeological	11: cast	11: dishes
12: cry	12: coins	12: development
13: data	13: forbidden	13: new
14: greasy	14: diploma	14: evening
15: dark	15: coverage	15: greasy
16: curiosity	16: desk	16: emphasized
17: church	17: enjoy	17: declining
18: compounded	18: evening	18: nevada
19: fuming	19: composure	19: graph
20: deadline	20: climbing	20: postponed
How difficult was this task?	How difficult was this task?	How difficult was this task?
☐ ☐ ☐ ☐ ☐ ☐ ☐	☐ ☐ ☐ ☐ ☐ ☐ ☐	☐ ☐ ☐ ☐ ☐ ☐ ☐
(not diff.) (medium) (diff.)	(not diff.) (medium) (diff.)	(not diff.) (medium) (diff.)

Comments: _____





With panning

Mark the words that you hear from the first speaker.

1: one
2: greasy
3: loved
4: earthquake
5: good
6: dark
7: innocence
8: mechanic
9: long
10: departure
11: never
12: nothing
13: mates
14: power
15: money
16: programs
17: penalty
18: price
19: minor
20: including

How difficult was this task?

☐ ☐ ☐ ☐ ☐ ☐ ☐

(not diff.) (medium) (diff.)

1: relaxed
2: sufficiently
3: norwegian
4: suit
5: garbage
6: soft
7: salads
8: pass
9: tied
10: points
11: window
12: particularly
13: year
14: ship's
15: today
16: sun
17: seeds
18: seldom
19: sweaters
20: worship

How difficult was this task?

☐ ☐ ☐ ☐ ☐ ☐ ☐

(not diff.) (medium) (diff.)

Comments: _____

Bibliography

[Algazi et al., 2001a] Algazi, V., Duda, R., Morrison, R., and Thompson, D. (2001a). Structural composition and decomposition of HRTFs. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 103–106.

[Algazi et al., 2001b] Algazi, V. R., Avendano, C., and Duda, R. O. (2001b). Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3):1110–1122.

[Algazi et al., 2001c] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001c). The CIPIC HRTF database. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 99–102.

[Atal and Schroeder, 1966] Atal, B. S. and Schroeder, M. R. (1966). Apparent sound source translator. US patent 3236949.

[Avery et al., 2005] Avery, B., Thomas, B. H., Velikovsky, J., and Piekarski, W. (2005). Outdoor augmented reality gaming on five dollars a day. In *AUIC '05: Proceedings of the Sixth Australasian conference on User interface*, pages 79–88, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

[Azuma et al., 2001] Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., and MacIntyre, B. (2001). Recent advances in augmented reality. *Computer Graphics and Applications, IEEE*, 21(6):34–47.

[Azuma, 1997] Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, 6(4):355–385.

[Bang & Olufsen, 1992] Bang & Olufsen (1992). Music for Archimedes. CD B&O 101.

[Bazerman et al., 2000] Bazerman, M. H., Curhan, J. R., Moore, D. A., and Valley, K. L. (2000). Negotiation. *Annual Review of Psychology*, 51(1):279–314.

[Bederson, 1995] Bederson, B. B. (1995). Audio augmented reality: a prototype automated tour guide. In *CHI '95: Conference companion on Human factors in computing systems*, pages 210–211, New York, NY, USA. ACM.

[Begault, 1992] Begault, D. R. (1992). Perceptual effects of synthetic reverberation on three-dimensional audio systems. *Journal of the Audio Engineering Society*, 40(11):895–904.

[Behringer et al., 1999] Behringer, R., Chen, S., Sundareswaran, V., Wang, K., and Vassiliou, M. (1999). A novel interface for device diagnostics using speech recognition, augmented reality visualization, and 3D audio auralization. In *IEEE International Conference on Multimedia Computing and Systems*, volume 1, pages 427–432.

[Beracoechea et al., 2008] Beracoechea, J. A., Torres-Guijarro, S., García, L., and Casajús-Quirós, Francisco J.and Ortiz, L. (2008). Subjective intelligibility evaluation in multiple-talker situation for virtual acoustic opening-based audio environments. *Journal of the Audio Engineering Society*, 56(5):339–356.

[Billinghurst et al., 2002] Billinghurst, M., Kato, H., Kiyokawa, K., Belcher, D., and Poupyrev, I. (2002). Experiments with face-to-face collaborative ar interfaces. *Virtual Reality*, 6(3):107–121.

[Blauert, 1996] Blauert, J. (1996). *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. The MIT Press.

[Bloom, 1977] Bloom, P. J. (1977). Creating source elevation illusions by spectral manipulation. *Journal of the Audio Engineering Society*, 25(9):560–565.

[Bovbjerg et al., 2000] Bovbjerg, B. P., Christensen, F., Minnaar, P., and Cheng, X. (2000). Measuring the head-related transfer functions of an artificial head with a high directional resolution. In *The AES 109th Convention Preprints*. Paper number 5264.

[Bronkhorst, 2000] Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128.

[Brooks, 1996] Brooks, Jr., F. P. (1996). The computer scientist as toolsmith II. *Commun. ACM*, 39(3):61–68.

[Brungart and Simpson, 2001] Brungart, D. S. and Simpson, B. D. (2001). Distance-based speech segregation in near-field virtual audio displays. In Hiipakka, J., Zacharov, N., and Takala, T., editors, *Proceedings of the 7th International Conference on Auditory Display (ICAD2001)*, pages 169–174.

[Burkhard and Sachs, 1975] Burkhard, M. D. and Sachs, R. M. (1975). Anthropometric manikin for acoustic research. *The Journal of the Acoustical Society of America*, 58(1):214–222.

[Chanda et al., 2006] Chanda, P., Park, S., and Kang, T. I. (2006). A binaural synthesis with multiple sound sources based on spatial features of head-related transfer functions. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1726–1730.

[Cherry, 1953] Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5):975–979.

[Christensen et al., 2000] Christensen, F., Jensen, C. B., and Møller, H. (2000). The design of VALDEMAR - an artificial head for binaural recording purposes. In *the AES 109th Convention Preprints*. Paper number 5253.

[CIPIC/IDAV Interface Laboratory, 2004] CIPIC/IDAV Interface Laboratory (2004). The CIPIC HRTF database. `http://interface.cipic.ucdavis.edu/CIL_html/CIL_HRTF_database.htm`.

[Cohen et al., 1993] Cohen, M., Aoki, S., and Koizumi, N. (1993). Augmented audio reality: telepresence/VR hybrid acoustic environments. In *Proceedings of the 2nd IEEE International Workshop on Robot and Human Communication*, pages 361–364.

[Cohen and Wenzel, 1995] Cohen, M. and Wenzel, E. M. (1995). The design of multidimensional sound interfaces. In *Virtual environments and advanced interface design*, pages 291–346. Oxford University Press, Inc., New York, NY, USA.

[Cooper and Bauck, 1989] Cooper, D. H. and Bauck, J. L. (1989). Prospects for transaural recording. *Journal of the Audio Engineering Society*, 37(1/2):3–19.

[Dalenbäck et al., 1996] Dalenbäck, B.-I., Kleiner, M., and Svensson, P. (1996). Auralization, virtually everywhere. In *The AES 100th Convention Preprints*. Paper number 4228.

[Drullman and Bronkhorst, 2000] Drullman, R. and Bronkhorst, A. W. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *The Journal of the Acoustical Society of America*, 107(4):2224–2235.

[Duda et al., 1999] Duda, R. O., Avendano, C., and Algazi, V. R. (1999). An adaptable ellipsoidal head model for the interaural time difference. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1999)*, pages 965–968, Washington, DC, USA. IEEE Computer Society.

[Echtler et al., 2003] Echtler, F., Sturm, F., Kindermann, K., Klinker, G., Stilla, J., Trilk, J., and Najafi, H. (2003). The intelligent welding gun: Augmented reality for experimental vehicle construction. In Ong, S. and Nee, A., editors, *Virtual and Augmented Reality Applications in Manufacturing, Chapter 17*. Springer Verlag.

[Eckel, 2001a] Eckel, G. (2001a). Immersive audio-augmented environments: the LISTEN project. In *Information Visualisation, 2001. Proceedings. Fifth International Conference on*, pages 571–573.

[Eckel, 2001b] Eckel, G. (2001b). The vision of the LISTEN project. In *Seventh International Conference on Virtual Systems and Multimedia*, pages 393–396.

[Fastl, 2004] Fastl, H. (2004). Towards a new dummy head? In *Proceedings of 33rd International Congress on Noise Control Engineering INTER-NOISE 2004, Prague, Czech Republic*.

[Feiner et al., 1997] Feiner, S., Macintyre, B., Höllerer, T., and Webster, A. (1997). A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. *Personal and Ubiquitous Computing*, 1(4):208–217.

[Fels and Vorländer, 2004] Fels, J. and Vorländer, M. (2004). Artificial heads for children. In *Proc. of the 19th ICA (International Congress on Acoustics)*, pages 3457–3458.

[Furness, 1986] Furness, T. A. (1986). Augmented reality: A class of displays on the reality–virtuality continuum. In *The Super Cockpit and its Human Factors Challenges*, pages 48–52, Boston, Massachusetts, USA.

[Gamper and Lokki, 2009] Gamper, H. and Lokki, T. (2009). Instant BRIR acquisition for auditory events in audio augmented reality using finger snaps. In *Proceedings of the International Workshop on the Principles and Applications of Spatial Hearing (IWPASH)*.

[Gardner and Martin, 2007] Gardner, H. J. and Martin, M. A. (2007). Analyzing ordinal scales in studies of virtual environments: Likert or lump it! *Presence: Teleoperators and Virtual Environments*, 16(4):439–446.

[Gardner, 1998] Gardner, W. (1998). *3-D audio using loudspeakers*. Kluwer Academic Publishers.

[Gardner and Martin, 1995] Gardner, W. G. and Martin, K. D. (1995). HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America*, 97(6):3907–3908.

[Garofolo et al., 1993] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993). DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM.

[Genuit, 1987] Genuit, K. (1987). Method and apparatus for simulating outer ear free field transfer function. US patent 4672569.

[Gilkey and Anderson, 1997] Gilkey, R. H. and Anderson, T. R., editors (1997). *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum Associates, Mahwah, New Jersey.

[Griesinger, 1990] Griesinger, D. (1990). Binaural techniques for music reproduction. In *Proceedings of the 8th international conference of the Audio Engineering Society*, pages 197–207.

[Hammershøi and Møller, 1996] Hammershøi, D. and Møller, H. (1996). Sound transmission to and within the human ear canal. *The Journal of the Acoustical Society of America*, 100(1):408–427.

[Härmä et al., 2004] Härmä, A., Jakka, J., Tikander, M., and Karjalainen, M. (2004). Augmented reality audio for mobile and wearable appliances. *Journal of the Audio Engineering Society*, 52(6):618–639.

[Härmä et al., 2003] Härmä, A., Jakka, J., Tikander, M., Karjalainen, M., Lokki, T., Nironen, H., and Vesa, S. (2003). Techniques and applications of wearable augmented reality audio. In *The AES 114th Convention Preprints*. Paper number 5768.

[Hartmann and Wittenberg, 1996] Hartmann, W. M. and Wittenberg, A. (1996). On the externalization of sound images. *The Journal of the Acoustical Society of America*, 99(6):3678–3688.

[Herre and Disch, 2007] Herre, J. and Disch, S. (2007). New concepts in parametric coding of spatial audio: From SAC to SAOC. In *IEEE International Conference on Multimedia and Expo*, pages 1894–1897.

[Hiipakka et al., 2009] Hiipakka, M., Karjalainen, M., and Pulkki, V. (2009). Estimating pressure at eardrum with pressure-velocity measurement from ear canal entrance. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '09)*, pages 289–292.

[Hill and Lewicki, 2006] Hill, T. and Lewicki, P. (2006). *Statistics : methods and applications*. StatSoft, Tulsa, Okla.

[Hindus et al., 1996] Hindus, D., Ackerman, M. S., Mainwaring, S., and Starr, B. (1996). Thunderwire: A field study of an audio-only media space. In *Computer Supported Cooperative Work*, pages 238–247. ACM Press.

[Hirahara et al., 2007] Hirahara, T., Sagara, H., and Otani, M. (2007). Sound localization with scaled dummy-heads on a telehead. In *Proc. of the 19th ICA (International Congress on Acoustics)*, pages 1–4.

[Huang and Hsieh, 2007] Huang, C. R. and Hsieh, S. F. (2007). Robust 3-D crosstalk canceller design. In *IEEE International Conference on Multimedia and Expo*, pages 1882–1885.

[Huopaniemi and Riederer, 1998] Huopaniemi, J. and Riederer, K. A. J. (1998). Measuring and modeling the effect of source distance in head-related transfer functions. In *Proceedings of the ICA/ASA '98 Conference*. Paper number 2988.

[IEEE, 2001] IEEE (2001). IEEE standard for inertial sensor terminology. *IEEE Std 528-2001*.

[Iida, 2008] Iida, K. (2008). Estimation of sound source elevation by extracting the vertical localization cues from binaural signals. In *Proceedings of Meetings on Acoustics*, volume 4. ASA. Paper number 050002.

[Ircam & AKG Acoustics, 2002] Ircam & AKG Acoustics (2002). LISTEN HRTF database. `http://www.ircam.fr/equipes/salles/listen/index.html`.

[Jain, 2000] Jain, R. (2000). Real reality. *Computer Graphics and Applications, IEEE*, 20(1):40–41.

[Jot et al., 1995] Jot, J.-M., Larcher, V., and Warusfel, O. (1995). On the minimum-phase nature of head-related transfer functions. In *The AES 98th Convention Preprints*. Paper number 3980.

[Julier et al., 2000] Julier, S., Baillot, Y., Lanzagorta, M., Brown, D., and Rosenblum, L. (2000). Bars: Battlefield augmented reality system. In *NATO Symposium on Information Processing Techniques for Military Systems*, pages 9–11.

[Kapralos et al., 2008] Kapralos, B., Jenkin, M. R., and Milios, E. (2008). Virtual audio systems. *Presence: Teleoperators and Virtual Environments*, 17(6):527–549.

[Katz et al., 2007] Katz, B. F. G., Tarault, A., Bourdot, P., and Vézien, J.-M. (2007). The use of 3D-audio in a multi-modal teleoperation platform for remote driving/supervision. In *30th International Conference of the Audio Engineering Socitey*, Saariselkä, Finland.

[Keyrouz et al., 2007] Keyrouz, F., Diepold, K., and Keyrouz, S. (2007). Humanoid binaural sound tracking using Kalman filtering and HRTFs. In *Proc. 6th Intl. Work. on Robot Motion and Control*, Poland.

[Kim et al., 2007a] Kim, J., Kim, Y., and Ko, S. (2007a). A parametric model of head-related transfer functions for sound source localization. In *The AES 122nd Convention Preprints*. Paper number 7096.

[Kim et al., 2007b] Kim, J. S., Kim, S. G., and Yoo, C. (2007b). A novel adaptive crosstalk cancellation using psychoacoustic model for 3D audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, volume 1, pages 185–188.

[Kim and Choi, 2005] Kim, S.-M. and Choi, W. (2005). On the externalization of virtual sound images in headphone reproduction: A Wiener filter approach. *The Journal of the Acoustical Society of America*, 117(6):3657–3665.

[Kim et al., 2005] Kim, Y., Kim, S., Kim, J., Lee, J., and il Park, S. (2005). New HRTFs (head related transfer functions) for 3D audio applications. In *The AES 118th Convention Preprints*. Paper number 6495.

[Kleiner et al., 1993] Kleiner, M., Dalenbäck, B.-I., and Svensson, P. (1993). Auralization - an overview. *Journal of the Audio Engineering Society*, 41(11):861–875.

[Kuipers, 2002] Kuipers, J. B. (2002). *Quaternions and Rotation Sequences: A Primer with Applications to Orbits, Aerospace and Virtual Reality*. Princeton University Press.

[Larsen et al., 2008] Larsen, E., Iyer, N., Lansing, C. R., and Feng, A. S. (2008). On the minimum audible difference in direct-to-reverberant energy ratio. *The Journal of the Acoustical Society of America*, 124(1):450–461.

[Lehnert and Blauert, 1991] Lehnert, H. and Blauert, J. (1991). Virtual auditory environment. In *Fifth International Conference on Advanced Robotics (ICAR 1991), 'Robots in Unstructured Environments'*, pages 211–216 vol.1.

[Lentz et al., 2006] Lentz, T., Assenmacher, I., Vorländer, M., and Kuhlen, T. (2006). Precise Near-to-Head Acoustics with Binaural Synthesis. *Journal of Virtual Reality and Broadcasting*, 3(2).

[Lilliefors, 1967] Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J. Am. Stat. Assoc.*, 62:399–402.

[Lindau et al., 2008] Lindau, A., Maempel, H.-J., and Weinzierl, S. (2008). Minimum BRIR grid resolution for dynamic binaural synthesis. In *Proc. of the Acoustics '08, Paris*, pages 3851–3856.

[Lindeman et al., 2007] Lindeman, R., Noma, H., and de Barros, P. (2007). Hear-through and mic-through augmented reality: Using bone conduction to display spatialized audio. In *6th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2007)*, pages 173–176.

[Lindeman et al., 2008] Lindeman, R., Noma, H., and Goncalves de Barros, P. (2008). An empirical study of hear-through augmented reality: Using bone conduction to deliver spatialized audio. In *IEEE Virtual Reality Conference (VR '08)*, pages 35–42.

[Lindeman et al., 2009] Lindeman, R. W., Reiners, D., and Steed, A. (2009). Practicing what we preach: IEEE VR 2009 virtual program committee meeting. *Computer Graphics and Applications, IEEE*, 29(2):80–83.

[Lokki and Gröhn, 2005] Lokki, T. and Gröhn, M. (2005). Navigation with auditory cues in a virtual environment. *Multimedia, IEEE*, 12(2):80–86.

[Lokki et al., 2004] Lokki, T., Nironen, H., Vesa, S., Savioja, L., Härmä, A., and Karjalainen, M. (2004). Application scenarios of wearable and mobile augmented reality audio. In *The AES 116th Convention Preprints*. paper number 6026.

[Loomis et al., 1998] Loomis, J. M., Golledge, R. G., Klatzky, R., and Klatzky, R. L. (1998). Navigation system for the blind: Auditory display modes and guidance. *Presence*, 7:193–203.

[Lucasfilm Ltd., 2004] Lucasfilm Ltd. (2004). MPSE to present George Lucas with filmmaker's award. `http://www.lucasfilm.com/press/news/news20041026.html`.

[Lyons et al., 2000] Lyons, K., Gandy, M., G, M., and Starner, T. (2000). Guided by voices: An audio augmented reality system. In *Proceedings of Intl. Conf. on Auditory Display (ICAD) 2000*.

[MacDonald et al., 2006] MacDonald, J. A., Henry, P. P., and Letowski, T. R. (2006). Spatial audio through a bone conduction interface. *International Journal of Audiology*, 45(10):595–599.

[Macpherson and Middlebrooks, 2002] Macpherson, E. A. and Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, 111(5):2219–2236.

[Middlebrooks et al., 1989] Middlebrooks, J. C., Makous, J. C., and Green, D. M. (1989). Directional sensitivity of sound-pressure levels in the human ear canal. *The Journal of the Acoustical Society of America*, 86(1):89–108.

[Milgram et al., 1995] Milgram, P., Takemura, H., Utsumi, A., and Kishino, F. (1995). Augmented reality: A class of displays on the reality–virtuality continuum. In *Proceedings of the SPIE Conference on Telemanipulator and Telepresence Technologies*, volume 2351, pages 282–292, Boston, Massachusetts, USA.

[Mills, 1958] Mills, A. W. (1958). On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4):237–246.

[Minnaar et al., 2001] Minnaar, P., Olesen, S. K., Christensen, F., and Møller, H. (2001). The importance of head movements for binaural room synthesis. In Hiipakka, J., Zacharov, N., and Takala, T., editors, *Proceedings of the 7th International Conference on Auditory Display (ICAD2001)*, pages 21–25.

[Minnaar et al., 2000] Minnaar, P., Plogsties, J., Olesen, S. K., Christensen, F., and Møller, H. (2000). The interaural time difference in binaural synthesis. In *The AES 108th Convention Preprints*. Paper number 5133.

[Møller et al., 1999] Møller, H., Jensen, C. B., Hammershøi, D., and Sørensen, M. F. (1999). Evaluation of artificial heads in listening tests. *Journal of the Audio Engineering Society*, 47(3):83–100.

[Møller et al., 1995] Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. (1995). Head-related transfer functions of human subjects. *Journal of the Audio Engineering Society*, 43(5):300–321.

[Møller et al., 1996] Møller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D. (1996). Binaural technique: Do we need individual recordings? *Journal of the Audio Engineering Society*, 44(6):451–469.

[Moore et al., 2007] Moore, A. H., Tew, A. I., and Nicol, R. (2007). Headphone transparification: A novel method for investigating the externalisation of binaural sounds. In *The AES 123rd Convention Preprints*. Paper number 7166.

[Motulsky, 1995] Motulsky, H. (1995). *Intuitive Biostatistics*. Oxford University Press, USA, 1 edition.

[Mynatt et al., 1998] Mynatt, E. D., Back, M., Want, R., Baer, M., and Ellis, J. B. (1998). Designing audio aura. In *CHI '98: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 566–573, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

[Nageno, 2001] Nageno, K. (2001). Headphone. US patent RE37398 (reissued).

[Nardi and Whittaker, 2002] Nardi, B. A. and Whittaker, S. (2002). The place of face-to-face communication in distributed work. In *IN P. HINDS AND S. KIESLER (EDS.), DISTRIBUTED WORK*, pages 83–112. MIT Press.

[Otani et al., 2009] Otani, M., Hirahara, T., and Ise, S. (2009). Numerical study on source-distance dependency of head-related transfer functions. *The Journal of the Acoustical Society of America*, 125(5):3253–3261.

[Parodi, 2008] Parodi, Y. L. (2008). Analysis of design parameters for crosstalk cancellation filters applied to different loudspeaker configurations. In *The AES 125th Convention Preprints*. Paper number 7636.

[Peltola, 2009] Peltola, M. (2009). Augmented reality audio applications in outdoor use. Master's thesis, Helsinki University of Technology.

[Pentenrieder et al., 2007] Pentenrieder, K., Bade, C., Doil, F., and Meier, P. (2007). Augmented reality-based factory planning - an application tailored to industrial needs. In *6th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2007)*, pages 31–42.

[Piekarski and Thomas, 2002] Piekarski, W. and Thomas, B. (2002). ARQuake: the outdoor augmented reality gaming system. *Commun. ACM*, 45(1):36–38.

[Priwin et al., 2004] Priwin, C., Stenfelt, S., Granström, G., Tjellström, A., and Håkansson, B. (2004). Bilateral bone-anchored hearing aids (BAHAs): An audiometric evaluation. *The Laryngoscope*, 114(1):77–84.

[Puckette, 1996] Puckette, M. (1996). Pure data: another integrated computer music environment. In *Proceedings of the International Computer Music Conference*, pages 37–41.

[Pulkki, 2001] Pulkki, V. (2001). *Spatial sound generation and perception by amplitude panning techniques*. PhD thesis, Helsinki University of Technology.

[Pulkki and Faller, 2006] Pulkki, V. and Faller, C. (2006). Directional Audio Coding: Filterbank and STFT-Based Design. In *Preprint 120th Conv. Aud. Eng. Soc.*

[Qu et al., 2008] Qu, T., Xiao, Z., Gong, M., Huang, Y., Li, X., and Wu, X. (2008). Distance dependent head-related transfer function database of KEMAR. In *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*, pages 466–470.

[Rao et al., 2006] Rao, H., Mathews, V., and Park, Y.-C. (2006). Inverse filter design using minimax approximation techniques for 3-D audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, volume 5, pages 353–356.

[Rapanos, 2008] Rapanos, N. (2008). Latin squares and their partial transversals. In Kominers, S. D., editor, *Harvard College Mathematics Review*, volume 2, pages 4–12. Harvard College.

[Raspaud and Evangelista, 2008] Raspaud, M. and Evangelista, G. (2008). Binaural partial tracking. In *Int. Conference on Digital Audio Effects DAFX08*, volume 11, pages 123–128, Espoo, Finland.

[Regenbrecht et al., 2005] Regenbrecht, H., Baratoff, G., and Wilke, W. (2005). Augmented reality projects in the automotive and aerospace industries. *Computer Graphics and Applications, IEEE*, 25(6):48–56.

[Riederer, 1998] Riederer, K. (1998). Repeatability analysis of head-related transfer function measurements. *AES 105th convention, preprint 4846*.

[Riikonen et al., 2008] Riikonen, V., Tikander, M., and Karjalainen, M. (2008). An augmented reality audio mixer and equalizer. In *The AES 124th Convention Preprints*. Paper number 7372.

[Röber, 2009] Röber, N. (2009). *Interaction with Sound: Explorations beyond the Frontiers of 3D virtual auditory Environments*. PhD thesis, Fakultät für Informatik, Otto-von-Guericke Universität Magdeburg.

[Rocchesso, 2002] Rocchesso, D. (2002). Spatial effects. In [Zölzer et al., 2002], pages 137–200.

[Rohde et al., 1997] Rohde, P., Lewinsohn, P. M., and Seeley, J. R. (1997). Comparability of Telephone and Face-to-Face Interviews in Assessing Axis I and II Disorders. *Am J Psychiatry*, 154(11):1593–1598.

[Rossing and Fletcher, 2004] Rossing, T. D. and Fletcher, N. H. (2004). *Principles of vibration and sound*. Springer, New York :, 2nd ed. edition.

[Rozier et al., 2000] Rozier, J., Karahalios, K., and Donath, J. (2000). Hear&there: An augmented reality system of linked audio. In *Proceedings of the International Conference on Auditory Display (ICAD)*, pages 63–67.

[Rychtáriková et al., 2009] Rychtáriková, M., Bogaert, T. V. d., Vermeir, G., and Wouters, J. (2009). Binaural sound source localization in real and virtual rooms. *Journal of the Audio Engineering Society*, 57(4):205–220.

[Satarzadeh et al., 2007] Satarzadeh, P., Algazi, V. R., and Duda, R. O. (2007). Physical and filter pinna models based on anthropometry. In *The AES 122nd Convention Preprints*. Paper number 7098.

[Savioja et al., 1999] Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R. (1999). Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, 47(9):675–705.

[Sawhney, 1998] Sawhney, N. (1998). Contextual awareness, messaging and communication in nomadic audio environments. Master's thesis, Media Arts and Sciences, MIT Media Lab.

[Sawhney and Schmandt, 2000] Sawhney, N. and Schmandt, C. (2000). Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. *ACM Trans. Comput.-Hum. Interact.*, 7(3):353–383.

[Saxena and Ng, 2009] Saxena, A. and Ng, A. (2009). Learning sound location from a single microphone. In *IEEE International Conference on Robotics and Automation (ICRA '09)*, pages 1737–1742.

[Shilling and Cunningham, 2002] Shilling, R. and Cunningham, S. B. (2002). *Virtual auditory displays*. Handbook of Virtual Environments. Lawrence Erlbaum Associates, Mahwah NJ.

[Shinn-Cunningham, 1998] Shinn-Cunningham, B. (1998). Applications of virtual auditory displays. In *Proceedings of the 20th International Conference of the IEEE Engineering in Biology and Medicine Society*, volume 3, pages 1105–1108 vol.3.

[Shinn-Cunningham et al., 1997] Shinn-Cunningham, B., Lehnert, H., Kramer, G., Wenzel, E., and Durlach, N. (1997). Auditory displays. In [Gilkey and Anderson, 1997], pages 611–663.

[Shinn-Cunningham et al., 2005] Shinn-Cunningham, B. G., Kopco, N., and Martin, T. J. (2005). Localizing nearby sound sources in a classroom: Binaural room impulse responses. *The Journal of the Acoustical Society of America*, 117(5):3100–3115.

[Slaney, 1998] Slaney, M. (1998). A critique of pure audition. In *Computational auditory scene analysis*, pages 27–41. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

[Snik et al., 1995] Snik, A. F. M., Mylanus, E. A. M., and Cremers, C. W. R. J. (1995). The bone-anchored hearing aid compared with conventional hearing aids. Audiologic results and the patients' opinions. *Otolaryngologic Clinics of North America*, 28:73–83.

[Sottek and Genuit, 1999] Sottek, R. and Genuit, K. (1999). Physical modeling of individual head-related transfer functions (HRTFs). *Journal of the Acoustical Society of America*, 105. Paper number 1162.

[Stanley and Walker, 2006] Stanley, R. M. and Walker, B. N. (2006). Lateralization of sounds using bone-conduction headsets. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*, pages 1571–1575.

[Stern et al., 2006] Stern, R. M., Wang, D., and Brown, G. (2006). Binaural sound localization. In Wang, D. and Brown, G. J., editors, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, pages 147–187. Wiley-IEEE Press.

[Sundareswaran et al., 2003] Sundareswaran, V., Wang, K., Chen, S., Behringer, R., McGee, J., Tam, C., and Zahorik, P. (2003). 3d audio augmented reality: implementation and experiments. In *Second IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 296–297.

[Sutherland, 1968] Sutherland, I. E. (1968). A head-mounted three dimensional display. In *AFIPS '68 (Fall, part I): Proceedings of the December 9–11, 1968, fall joint computer conference, part I*, pages 757–764, New York, NY, USA. ACM.

[Tappan, 1964] Tappan, P. W. (1964). Proximaural loudspeakers (-nearphones-). In *The AES 16th Convention Preprints*. Paper number 358.

[The Oxford English Dictionary, 2006] The Oxford English Dictionary (2006). "augmented reality n." def. OED Online, `http://dictionary.oed.com/cgi/entry/50014757`.

[Toshima et al., 2004] Toshima, I., Aoki, S., and Hirahara, T. (2004). An acoustical tele-presence robot: Telehead II. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*, volume 3, pages 2105–2110 vol.3.

[Väljamäe et al., 2009] Väljamäe, A., Larsson, P., Västfjäll, D., and Kleiner, M. (2009). Auditory landmarks enhance circular vection in multimodal virtual reality. *Journal of the Audio Engineering Society*, 57(3):111–120.

[Vallino, 1998] Vallino, J. R. (1998). *Interactive Augmented Reality*. PhD thesis, University of Rochester.

[Walker et al., 2001] Walker, A., Brewster, S., Mcgookin, D., and Ng, A. (2001). Diary in the sky: A spatial audio display for a mobile calendar. In *proceedings of IHM-HCI 2001*, pages 531–540. Springer.

[Walker and Lindsay, 2005] Walker, B. N. and Lindsay, J. (2005). Navigation performance in a virtual environment with bonephones. In Brazil, E., editor, *Proceedings of the 11th International Conference on Auditory Display (ICAD2005)*, pages 260–263, Limerick, Ireland.

[Warusfel and Eckel, 2004] Warusfel, O. and Eckel, G. (2004). LISTEN – augmenting everyday environments through interactive soundscapes. In *IEEE Workshop on VR for public consumption*, Chicago.

[Wenzel, 1992] Wenzel, E. M. (1992). Localization in virtual acoustic displays. *Presence: Teleoperators and Virtual Environments*, 1(1):80–107.

[Wenzel et al., 1993] Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123.

[Wightman and Kistler, 1989] Wightman, F. L. and Kistler, D. J. (1989). Headphone simulation of free-field listening. I: Stimulus synthesis. *The Journal of the Acoustical Society of America*, 85(2):858–867.

[Wightman and Kistler, 1997] Wightman, F. L. and Kistler, D. J. (1997). Factors affecting the relative salience of sound localization cues. In [Gilkey and Anderson, 1997], pages 1–23.

[Williamson et al., 2007] Williamson, J., Smith, R. M., and Hughes, S. (2007). Shoogle: excitatory multimodal interaction on mobile devices. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 121–124, New York, NY, USA. ACM.

[Yost, 1993] Yost, W. A. (1993). Perceptual models for auditory localization. In *The AES 12th Convention Preprints*, pages 155–168.

[Yost, 1997] Yost, W. A. (1997). The cocktail party problem: Forty years later. In [Gilkey and Anderson, 1997], pages 329–347.

[Zahorik et al., 1995] Zahorik, P., Wightman, F., and Kistler, D. (1995). On the discriminability of virtual and real sound sources. In *Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on*, pages 76–79.

[Zalis et al., 2005] Zalis, M. E., Perumpillichira, J. J., Kim, J. Y., Del Frate, C., Magee, C., and Hahn, P. F. (2005). Polyp Size at CT Colonography after Electronic Subtraction Cleansing in an Anthropomorphic Colon Phantom1. *Radiology*, 236(1):118–124.

[Zhu and Pan, 2008] Zhu, J. and Pan, Z. (2008). Occlusion registration in video-based augmented reality. In *VRCAI '08: Proceedings of The 7th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, pages 1–6, New York, NY, USA. ACM.

[Zimmermann and Lorenz, 2008] Zimmermann, A. and Lorenz, A. (2008). LISTEN: a user-adaptive audio-augmented museum guide. *User Modeling and User-Adapted Interaction*, 18(5):389–416.

[Zölzer et al., 2002] Zölzer, U., Amatriain, X., Arfib, D., Bonada, J., De Poli, G., Dutilleux, P., Evangelista, G., Keiler, F., Loscos, A., Rocchesso, D., Sandler, M., Serra, X., and Todoroff, T. (2002). *DAFX:Digital Audio Effects*. John Wiley & Sons.

[Zotkin et al., 2004] Zotkin, D., Duraiswami, R., and Davis, L. (2004). Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia*, 6(4):553–564.

[Zotkin et al., 2003] Zotkin, D., Hwang, J., Duraiswaini, R., and Davis, L. (2003). HRTF personalization using anthropometric measurements. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 157–160.