Augmented Reality Audio for Mobile and Wearable Appliances*

AKI HÄRMÄ, JULIA JAKKA, MIIKKA TIKANDER, AND MATTI KARJALAINEN, *AES Fellow*

Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, FIN-02015, HUT, Finland

TAPIO LOKKI, *AES Member*

Helsinki University of Technology, Telecommunications Software and Multimedia Laboratory, FIN-02015, HUT, Finland

AND

JARMO HIIPAKKA, *AES Member,* **AND GAËTAN LORHO,** *AES Member*

Nokia Research Center, Audio-Visual Systems Laboratory, FIN-00045, NOKIA GROUP, Finland

The concept of augmented reality audio characterizes techniques where a real sound environment is extended with virtual auditory environments and communications scenarios. A framework is introduced for mobile augmented reality audio (MARA) based on a specific headset configuration where binaural microphone elements are integrated into stereo earphones. When microphone signals are routed directly to the earphones, a user is exposed to a pseudoacoustic representation of the real environment. Virtual sound events are then mixed with microphone signals to produce a hybrid, an augmented reality audio representation, for the user. An overview of related technology, literature, and application scenarios is provided. Listening test results with a prototype system show that the proposed system has interesting properties. For example, in some cases listeners found it very difficult to determine which sound sources in an augmented reality audio representation are real and which are virtual.

0 INTRODUCTION

The era of wearable audio appliances started with the introduction of the portable cassette player more than two decades ago. The development of digital technology led to portable CD players and finally to fully digital MP3 players [1]. Currently other devices may be considered belonging to the same category. For instance, digital cellular phones have developed considerably in recent years [2]. While speech communication is the main application, many manufacturers have recently integrated a digital audio player to a phone to enable high-quality audio playback. However, the basic application scenario for wideband audio is still the same as in early walkmans. Of the currently available audiocentric devices, hearing aids may be considered the most wearable. The number of users of these devices is constantly increasing in developed countries. With digital technology the quality of hearing aid devices has improved significantly while the prices are dropping, thus preparing the way for an even higher number of users. Yet another application that relates to this is personal active hearing protectors.

We may have multiple wearable audio appliances, but only one pair of ears. At some point it makes sense to integrate all of these functions into the same physical device. Mechanical and electrical integration is already feasible. However, in application scenarios there are many interesting new possibilities and problems to explore [3]. Also, the progress in speech and audio technology, computing, and communications predicts the introduction of completely new types of intelligent and interactive audio and speech applications. For example, spatial auditory displays that can provide the user with different types of information in the form of spatialized sound events have been introduced [4].

We consider a device that a user could be wearing at all times. It would resemble portable audio players in some

*Manuscript received 2003 August 6; revised 2004 January 26 and March 15.

respect and also provide speech and audio communications services, for example, over a mobile or wireless network. But at the same time it would also make it possible for a user to hear and interact with the real acoustic environment in a natural way, a principle proposed earlier by many authors [5]–[7]. Thus it would facilitate and even make easier ordinary speech communication with other people, enable safe navigation in traffic, and permit the operation of machines where acoustic feedback is important. This is the case particularly in assistive devices for the blind [8]. In addition there would be a large number of new functions that provide information and communication channels which are not available in a natural acoustic environment or in current appliances.

The possibility to hear the natural acoustic environment around a user differentiates the concept of augmented reality (AR) audio from the traditional concept of a virtual reality (VR) audio environment, where a user is typically immersed into a completely synthetic acoustic environment [9], [10]. The computer graphics community has established concise definitions for different types of realities. Augmented reality is produced by adding synthetic objects into the real environment [11]. In augmented virtuality (AV) objects from the real world are embedded into a virtual reality scene. Finally, mixed reality (MR) includes VR, AR, AV, and a continuum between them [12], [13]. The framework presented in this paper could be used to implement AV or MR audio applications. However, in this study the focus is on augmented reality audio [6], where the mixing of real and virtual environments can be most easily understood as a process of adding virtual audio objects to the real or a modified real acoustic environment around a user.

The proposed system for mobile augmented reality audio (MARA) requires specific transducer systems and auralization techniques.[1] In the prototype system introduced in this paper the transducer configuration is based on a headset where miniature microphones are integrated into earphone elements in both ears. When microphone sounds are routed directly to the earphones, a user can perceive a representation of the real acoustic environment. Since the experience may differ from the open-ear case, we call this representation the pseudoacoustic environment. It has been demonstrated that users can adapt to a modified binaural representation within a reasonable time [14]. Virtual and synthetic sound events, such as the speech of a remote user, music, audio markers, or simply user interface sounds are superimposed onto the pseudoacoustic sound environment in a device that may be called an augmentation mixer. At one extreme, virtual sounds can be combined with the pseudoacoustic signals in the augmentation mixer in such a way that a user may not be able to determine which sound sources are local and which are rendered artificially by means of digital signal processing. Listening test results presented in this paper demonstrate that in some cases this can be achieved relatively easily using personalized in-situ head-related room impulse responses (HRIRs) and even generic head-related transfer functions (HRTFs) combined with a synthetic room model.

Real-time implementation of a MARA system requires low latency for audio and seamless integration of audio streams, signal processing algorithms, and network communications. Implementation aspects of the software system are beyond the scope of this paper. A modular and flexible software architecture based on the Mustajuuri system [15] for testing different MARA application scenarios was introduced in [3].

This paper starts with an introductory section where we put the MARA concept into a wider context and define the central parts of the proposed framework. We also review the literature and give an organized representation of previous works and ideas in related application fields. In Section 3 we propose a transducer system for MARA and present listening test results conducted to evaluate the performance of the system. In Section 5 we study some specific signal processing techniques needed in building applications and give some ideas about user interfaces needed for mobile augmented reality audio applications.

1 REAL, VIRTUAL, AND AUGMENTED AUDIO ENVIRONMENTS

The basic difference between real and virtual audio environments is that virtual sounds are originating from another environment or are created artificially. Augmented reality audio (or augmented audio reality) combines these aspects in a way where real and virtual sound scenes are mixed so that virtual sounds are perceived as an extension to the natural ones. At one extreme an augmented reality audio system should pass a test that is closely related to the classical Turing test for artificial intelligence [16]. That is, if a listener is unable to determine whether a sound source is part of the real or a virtual audio environment, the system implements a subjectively perfect augmentation of the listener's auditory environment. At the other extreme, virtual auditory scenes could be rendered in high quality such that they are easily separable from real ones by their characteristics which are not possible in normal acoustic environments.

Fig. 1(a) illustrates a user in a real acoustic environment and Fig. 1(b) a headphone-based virtual acoustic system where the sound environment is created using a computer.

1.1 Externalization in Virtual Spatial Audio

In the current paper the focus is on the development of techniques for mobile and wearable applications. Hence it is clear that transducers used for producing virtual sounds must be wearable. In practice, headphones or earphones are probably the most convenient alternatives, although other types of wearable transducer systems can also be considered [17]. Headphones have been used successfully in many virtual reality applications [18].

Headphone auralization often produces an incorrect localization of virtual sound sources. For example, common problems are front–back confusion and a perception that sources at the front are elevated. In augmented reality audio, probably the most severe problem is a perceived effect of having the virtual source localized inside the

[1] In [3] we called this wearable augmented reality audio (WARA).

listener's head. This is usually called intracranial, or inside-the-head locatedness (IHL) [19]. The spatial localization of sound sources that are perceived to be inside the listener's head is termed lateralization. It has been demonstrated that a listener can make a clear distinction in headphone listening between localized (that is, sounds outside the head) and lateralized sound sources and that these two types can coexist in the listener's experience [20].

The effect of lateralized sound in headphone listening can be produced using amplitude and delay differences in two headphone channels corresponding to each source. In order to make a sound source externalized, more sophisticated binaural techniques are needed [21]. Differences in the two ear signals due to head-related transfer functions (HRTFs) play an important role. In the laboratory environment personalized HRTFs can be used to produce a realistic illusion of an externalized source [22]. However, there may be significant variability in HRTFs among subjects, depending on how the HRTFs have been measured [23].

HRTF responses measured in a free field apply only in free-field conditions, which is an unnatural sounding environment for most people. Additional acoustic cues such as the amount of reverberation in virtual sound and unwanted sounds in the environment make it easy for a listener to resolve the Turing test [16]. Externalization is related to the perception of auditory distance such that there is a continuum in perceived locations of sources from inside the head to any external position [19], [22]. In [24] it was found that the direct-to-reverberant energy ratio dominates auditory distance cues for unfamiliar sounds while intensity is the most important distance cue in speech signals. It is well known that the use of artificial reverberation can help in forming an externalized sound image in headphone listening [25], [26]. Similarly, sounds recorded in echoic environments are perceived more distant than dry recordings [27]. Control of signal level and reverberation are clearly necessary for a successful auralization of virtual sound sources.

Dynamic cues related to head turning and other movements of a listener or a source are important because listeners use them both intentionally [28] and unintentionally [29] in a listening situation. The virtual source should be stable in relation to the environment. Unnatural changes in the acoustic properties of the incoming signal can easily damage the perceptual illusion of a virtual source [30]. In headphone listening this requires that the auralization processor be controlled by the position and orientation of the listener's head (see [31]). One such system, called binaural room scanning (BRS), has been implemented in [32].

Multimodality aspects such as the connection of a sound event to a visible real-world object and the user's expectations concerning the performance of the system also affect the externalization and localization of virtual sound sources [33]. For familiar sounds the type of sound is also of importance. For example, whispered and shouted voices gave completely different estimates for the distance of a speaker in listening tests reported in [34].

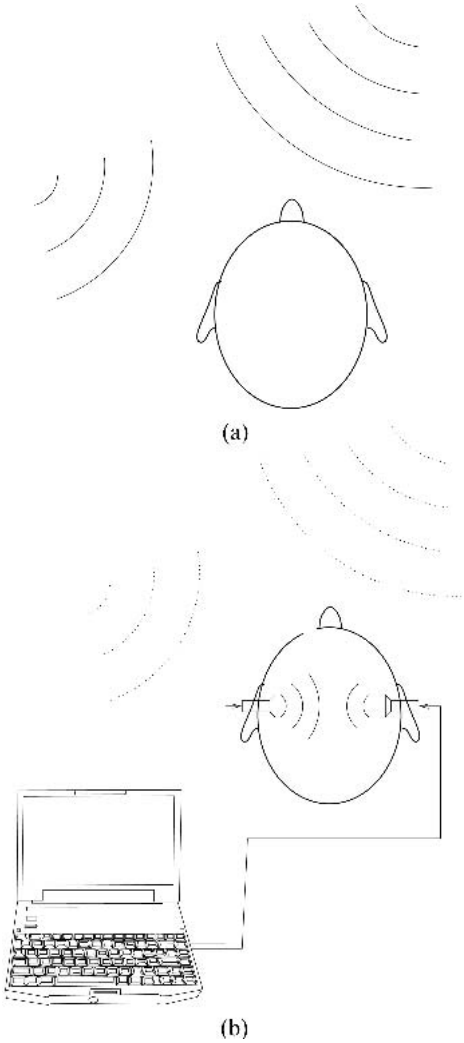

Fig. 1. (a) Listener in real environmental. (b) Listener in virtual environment.

1.2 Virtual Reality Environment

In this paper we assume that the virtual sound environment is created using techniques of auralization [35]–[39]. Typically this involves binaural processing using HRTF filters.

Let us first formulate the open-ear case illustrated in Fig. 1(a). Denoting the z transform of an input signal at the entrance of the left ear canal by $y_{\text{ear}}(z)$, the filtering of a source signal $x(z)$ from a certain position is given by

$$y_{\text{ear}}(z) = H_{\text{l}}(z)x(z) \tag{1}$$

where $H_{\text{l}}(z)$ is the HRTF corresponding to the position of the source. This is illustrated in Fig. 2(a). Assuming that $H_{\text{l}}(z)$ is an estimated transfer function, one may use it to render a signal $a(z)$ corresponding to the same location of

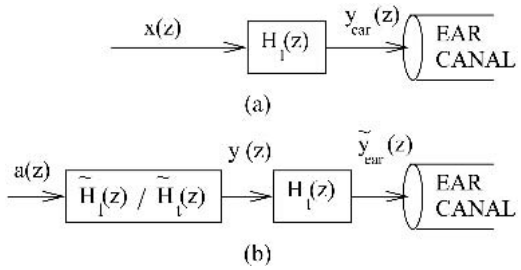

Fig. 2. (a) Signal paths in open-ear case. (b) Signal paths in typical headphone auralization.

a source. This configuration is shown in Fig. 1(b). In headphone auralization one must take into account the transfer function from a headphone element to the ear canal (see, for example, [40]). This transfer function is sometimes called ear canal transfer function (ECTF) [41], or earphone transfer function (ETF) [42]. Fig. 2(b) illustrates a typical configuration for headphone auralization where the ETF is denoted by $H_t(z)$. In computing the actual output signal of the system $y_l(z)$, the effect of the ETF has been canceled from the estimated HRTF $\tilde{H}_l(z)$.

It has been verified that there are significantly larger individual differences in ETFs than in HRTFs [41], [23].

The headphone auralization scheme illustrated in Fig. 2(b) is a typical system used in auditory displays. However, in augmented reality audio it does not necessarily lead to good results in the externalization of sound sources in a reverberant pseudoacoustic environment. It is probably necessary to bring also some early reflections and reverberation to the virtual sound environment to make it match better with the local environment.

In some applications the rendered virtual sound environment should be independent of the position and rotation of the user's head. The sound source should be localized and often somehow connected to the real environment around the user. In these applications some system for finding the position and orientation of a user is needed.

1.3 Pseudoacoustic Environment

The pseudoacoustic environment is a modified representation of the real acoustic environment around a user. In this paper we consider a specific binaural headset where a small microphone element has been integrated into each earphone. This system is illustrated in Fig. 3(a) and the actual ear piece in Fig. 4.

The signal entering the ear canal in one ear may be characterized by the block diagram of Fig. 5(a) and the following equation:

$$\tilde{y}_{ear}(z) = [H_m(z)H_t(z) + E(z)]x(z) \quad (2)$$

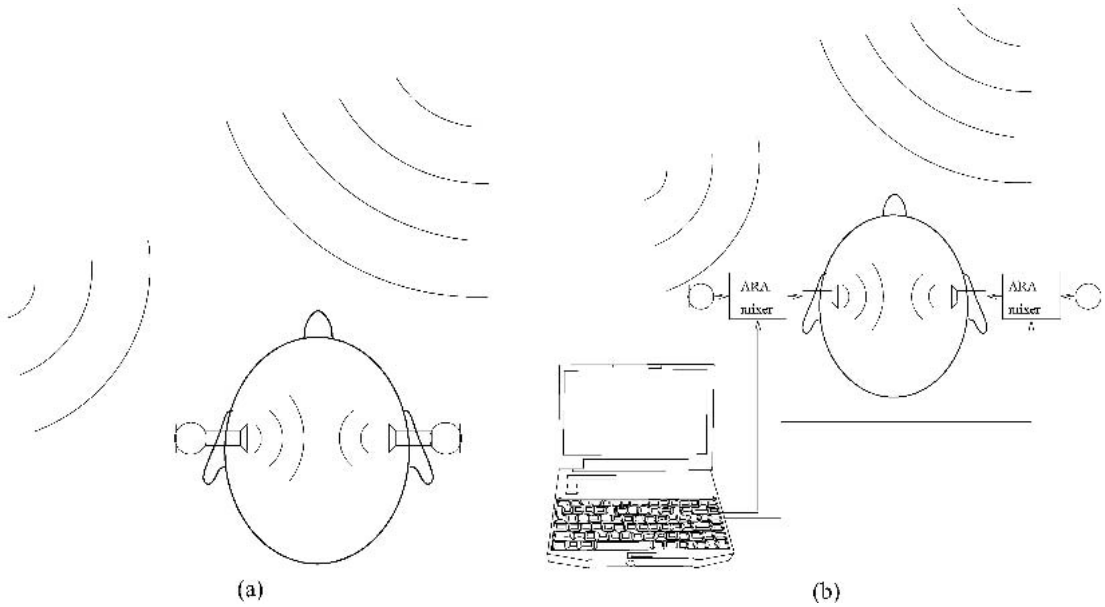

Fig. 3. (a) Listener in pseudoacoustic environment. (b) Listener in augmented environment.

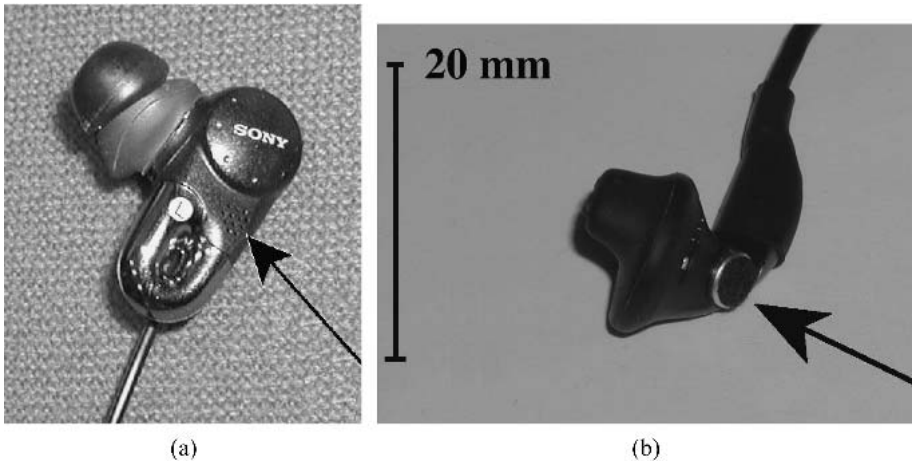

Fig. 4. Headsets used in tests. (a) Model I, constructed from a pair of Sony MDR-NC10 noice canceling headphones. (b) Model II with an open-type earphone constructed from a Sony MDR-ED268LP earphone and an electret microphone element. Positions of the microphone elements are indicated by arrows.

where $H_t(z)$ represents the ETF, as in Fig. 2. $H_m(z)$ is a transfer function from a certain position in space to a microphone element in our microphone–earphone system, that is, it is basically similar to an HRTF but estimated using the microphone mounted into an earphone element close to the ear canal. Finally, $E(z)$ is a transfer function representing the leakage of direct sound from the source location to the ear canal. It should be noted that in the present discussion we simplify notation by using $E(z)$ while it would be more appropriate to express the leakage sound as an analog acoustic signal $E(\omega)$.

In many applications it is most convenient to try to make the pseudoacoustic environment as identical to the real environment as possible. It would be possible to estimate $H_t(z)$ and use it to cancel the effect of the earphone by replacing $H_m(z)$ with $H_m(z)/\tilde{H}_t(z)$. Equalization filters $\tilde{H}_t(z)$ could be estimated in a measurement where a probe microphone was inserted into the ear canal. However, this is difficult and would probably lead to highly individualized filters, specific for some particular positioning of equipment [43].

The leakage effect of $E(z)$ is even more difficult to handle. Adding the compensation filter $\tilde{H}_t^{-1}(z)$ would typically call for a digital implementation of filtering. This would add delay to the signal path. Since $E(z)$ represents delay-free mechanical or acoustic leakage of sound to the ear canal, any additional delay in the transfer path from microphone to the earphone would mean that the leakage sound would arrive at the user's ear canal before the earphone signal. That is, any attempt to cancel the leakage transfer function $E(z)$ from $\tilde{y}_{ear}(z)$ would be impossible because the compensation filter would be noncausal. However, at low frequencies the latency is a smaller problem, and therefore cancellation using DSP could be used. In the current paper we only try to control the signal level and do some coarse equalization of signals to make the pseudoacoustic environment sound as natural as possible. Accordingly some difference is expected to remain in the user's spatial impression. It has been demonstrated in many experiments that listeners can adapt to atypical [14] or supernormal binaural inputs [44]. Modified HRTF filters have also been used intentionally in virtual audio applications to reduce problems related to the front–back confusion [45].

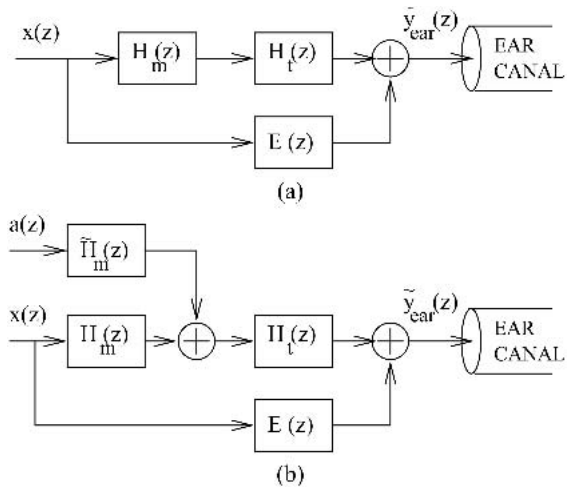

Fig. 5. (a) Signal paths in pseudoacoustic reproduction. (b) Signal paths in augmented reality audio rendering.

There are basically two practical ways to reduce the problem of leakage. First the design of the microphone–earphone element could be made carefully enough to effectively attenuate the leakage sound. Second, the level of the pseudoacoustic sound can be made high such that it efficiently masks the contribution of the leakage.

Let $p_{ear}(t)$ be a sound pressure signal of a pure tone in the ear canal. It is composed of

$$p_{ear}(t) = 10^{G/20}p(t) + 10^{A/20}p(t - \phi) \tag{3}$$

where $p(t)$ is the pressure signal outside the ear, G is a gain in dB provided by the microphone–earphone system, and A is attenuation in the leakage path. The phase term ϕ represents the difference between pseudoacoustic and leakage sounds. If for a certain partial $\phi = 0$, the magnitude of the partial is increased by 201°g $(10^{G/20} + 10^{A/20})$ dB. If, on the other hand, $\phi = \pm\pi$, the amplitude of the partial is attenuated by 201°g $(10^{G/20} - 10^{A/20})$ dB. In particular if $A = G$, the partial vanishes from the ear canal signal. The difference between the two cases gives the worst-case spectrum change in the ear canal signal $p_{ear}(t)$. The general rule of thumb is that changes in signal spectrum greater than 1–2 dB are noticeable [46], [47]. However, the just noticeable difference for moderate and smooth changes in the spectrum envelope, such as spectrum tilting, is in the range of 2–4 dB [48]. In binaural listening conditions the just noticeable difference has been found to be slightly higher [49] than in a monoaural case. The amplification G needed to compensate for a leakage attenuated by A dB to have a maximum spectrum difference of D dB can be derived easily and is given by

$$G = A + 20 \log 10\left(\frac{1 + 10^{D/20}}{10^{D/20} - 1}\right). \tag{4}$$

The difference $|G - A|$ as a function of D is plotted in Fig. 6. If the attenuation of the leakage path is 25 dB or higher, the spectrum difference never exceeds 1 dB. On the other hand, if the attenuation is only 5 dB, we may need a gain of 20 dB for the pseudoacoustic signal to compensate for the worst-case spectrum coloration occurring when there are both coherent summation and cancellation of frequency components in a spectrum. In listening tests in Section 4 we compare two different earphone models of very different attenuation characteristics.

The proposed model could also be used to produce a modified representation of reality, which could be advantageous, more convenient, or just an entertaining feature for a user. For example, in some cases the system would provide hearing protection, hearing aid functions, noise reduction, spatial filtering, or it would emphasize important sound signals such as alarm and warning sounds.

1.4 Augmented Reality Audio Environment

An augmented audio environment is produced by superimposing the virtual sound environment onto the pseudoacoustic environment. First the pseudoacoustic environ-

ment should be delivered to a user in such a way that its loudness and binaural properties are acceptable for the user. Second the virtual sound environment should be mixed carefully with the local environment to produce a coherent perception for the user. The goal is to find the best mixing rules for local and virtual environments, leading to a meaningful fusion of these two main components of augmented reality audio. In Fig. 3(b) the mixing of the pseudoacoustic and virtual audio environments is performed in a device called augmented reality audio (ARA) mixer.

Fig. 5(b) shows a block diagram of the proposed augmented reality audio system. The pseudoacoustic environment is provided for the user as explained. Virtual audio events are added to the earphone signals using HRTF-type directional filters $\tilde{H}_{\mathrm{m}}(z)$, approximating responses $H_{\mathrm{m}}(z)$ from a source location to a microphone in the earpiece.

Comparing the two block diagrams of Fig. 5 one may see that if $\tilde{H}_{\mathrm{m}}(z) = H_{\mathrm{m}}(z)$, the only difference between pseudoacoustic sounds and virtual sounds is that the effect of leakage by $E(z)$ is missing from virtual sounds. In listening tests reported in Section 4 we will study the importance of this difference.

The true HRTF given by $H_{\mathrm{l}}(z)$ in Eq. (1) is replaced by the transfer function of Eq. (2), which represents a linear modification of input signals. In [44] it was found that listeners can accommodate linear transformations more easily than arbitrary nonlinear mappings of binaural input signals.

2. POSITIONING OF VIRTUAL SOUND OBJECTS

In typical virtual audio applications audio objects are rendered in relation to the user's head. In principle the location of the user's head sets up a virtual coordinate system, or a map of the virtual auditory scene. Sound objects in a pseudoacoustic environment are placed in the physical environment around the user and therefore, conceptually, positioned according to a physical coordinate system. Thus putting virtual audio objects onto the pseudoacoustic environment also means that we are superimposing one coordinate system on another.

There are many different ways of setting the coordinate system for virtual audio objects, but there are also many different ways of characterizing the physical coordinates of objects in the environment. Restricting the discussion to users on the planet earth, the physical coordinate system may be based on global latitudes and longitudes. However, it may also be a local coordinate system inside a building or a moving vehicle, for example, in an automobile application.

In some applications it is necessary to make the two coordinate systems match so that virtual sound sources appear in distinct locations in the physical environment. In other applications it may be sufficient that virtual sources are floating somewhere around the user.

2.1 Virtual Coordinates

The most common case of a floating virtual coordinate system is the one where the only anchor point relative to which the event is localized is the user's head. Usually, virtual sound sources are rendered to different directions. For example, information services such as news, calendar events, e-mails, or other types of messages can be remapped to the virtual acoustic space a user [50], [17]. The spatial calendar application introduced in [51] is a typical example. Here speech messages representing calendar events are rendered around the user so that noon appears in the front of the user, 3 p.m. at the right and 6 p.m. behind the user.

Immersive virtual reality applications also use specific virtual coordinate systems. Usually the coordinate system is related to the geometry of a graphical virtual reality scene [52]–[54]. In computer game applications using spatial audio techniques, the virtual coordinate system is spanned by the game scene [55] and sometimes combined with information on the physical location of a user [56].

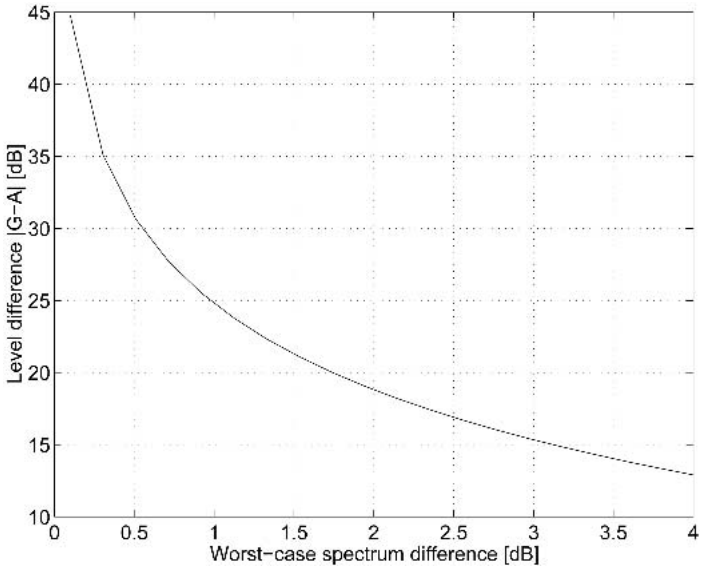

Fig. 6. Difference between gain of pseudoacoustic signal and attenuation of leakage sound $|G - A|$ versus worst-case maximum spectrum deviation D in ear canal signal $p_{\mathrm{ear}}(t)$.

Telepresence is another case of a floating virtual coordinate system. The idea was originally approached in 1933, when binaural hearing was examined by letting test persons with two receivers listen to sound examples through microphones in a dummy head's ears, the dummy head being in another room [57]. Conventional telepresence applications (see, for example, [6]) are similar to virtual auditory display systems in the sense that they aim at producing an immersive experience for a user. That is, a user is typically disconnected from the surrounding acoustic environment. The bidirectional augmented telepresence application illustrated in Fig. 7 is an interesting special case where a binaural telepresence signal is combined with a pseudoacoustic environment [3]. For the user on the right, the system would combine the local pseudoacoustic environment with a remote pseudoacoustic environment to produce a specific type of augmented audio reality experience. The coordinate system associated with the binaural signal rendered onto the pseudoacoustic environment represents the physical environment of the other person.

Virtual audio teleconferencing systems based on headphone auralization have been studied by many authors [58]–[62]. Improved intelligibility in spatially separated talkers [59] is typically linked to Cherry's concept of the cocktail party effect [63]. In multiparty teleconferencing the positioning of each talker can be done freely. Typically it is based on some predefined "assumed virtual environment" [58] or a map of a virtual meeting room [60]. In the system of Fig. 7 participants at the remote end are placed in their physical positions, but the alignment of the two coordinate systems can be done freely.

2.2 Natural Coordinates

When placing virtual audio objects in definite locations in the physical environment around a user, it is necessary to make the coordinate system used for rendering virtual sounds match with a map of the physical environment. At one extreme it would be possible to place a virtual audio object to any physical object in the universe. For example, one application introduced in [3] was based on the concept of a localized audio message, an acoustic Post-it sticker, for which one can assign any location in a physical coordinate system. This idea has earlier been introduced in museum and exhibition audio guide systems [64], [65]. It is also related to the concept of situated computing [66]. In [64] a recorded message was played to a user when he/she was in a certain location in a museum. In [65] the system was tracking the location of a user continuously, and an auralized dynamic sound scape and different spoken messages were played through wireless headphones for the user depending on the location.

One potential application for MARA is a navigation aid system for the visually impaired (see, for example, [67], [68], [8]. In this application the map of the physical space can be global or local.

A typical example for a local physical coordinate system are virtual auditory displays for air crews on simulated night mission flights [4] and collision alarm systems for flight pilots [69]. Here the associated physical coordinate system is moving with the airplane. In both cases the matching between virtual and physical coordinate systems is very critical.

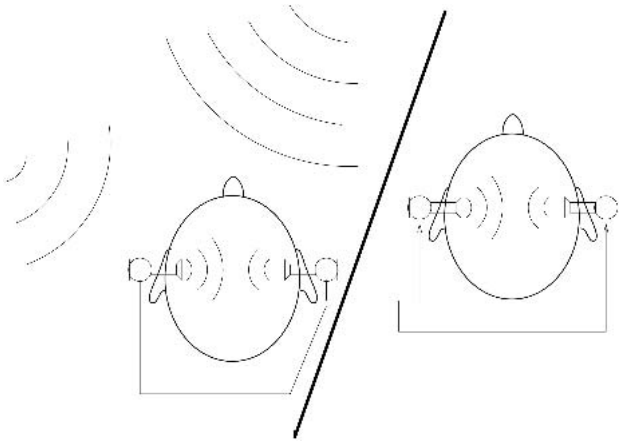

Fig. 7. Listener in augmented environment and another user experiencing telepresence based on binaural reproduction.

3 MOBILE AUGMENTED REALITY AUDIO HEADSETS

In this study we focus on a rather specific type of microphone–earphone system. There are many alternatives, but this was chosen as a starting point because it seems to have many beneficial properties and it directly facilitates the testing of rather unconventional applications and services. In many cases (for example, in [70], [71]) enclosed headphones are used and the microphones are attached outside the enclosures. Even if the microphone signals are played back to let the sound pass the headphones, important binaural cues are lost due to the enclosures and the positioning of the microphones. Sawhney and Schmandt have proposed a system [17] where the user is wearing a special collar that has a small loudspeaker on each shoulder and a microphone placed on the chest. This setup does not interfere with binaural hearing, but the audio quality of the loudspeakers is fairly poor compared to the headphones. In addition, the sound is radiated to other subjects in the environment.

Small in-ear headphones are already widely used in portable audio and with mobile phones as hand-free headsets. Adding microphones to this style of headphones allows the microphones to be located very close to the ear canals. In this way binaural information can be captured accurately.

3.1 The Proposed Transducer System

In the proposed system microphones are mounted directly on earplug-type earphone elements and are therefore placed very close to the ear canals. Ideally, the whole system could fit in the user's ear and there would be no additional wires. The wearability of this device would be the same as for a hearing aid device. Earplug-type earphones have a low power consumption, which results in extended operating times and low weight of the device.

For testing we constructed two different types of headsets—model I and model II—shown in Fig. 4. The average

attenuation of direct sound in model I measured in an anechoic chamber with an audiometer for four subjects is plotted in Fig. 8 (lower curve). In model I the earphones are of the earplug type and provide 10–30 dB attenuation of direct sound. In model II the earphones are placed at the entrance of the ear canal and provide only 1–5 dB attenuation. The average attenuation measured with the audiometer is shown as the upper curve of Fig. 8.

The use of earplug-type headphones makes it possible to control the signals entering a listener's ears more accurately than, for example, with open headphones. One may mute or amplify sounds selectively if the direct acoustic propagation of sound into ears is suppressed by blocking the ear canal with the device. However, model II may be more convenient for a typical user. The attenuation of external sounds is one of the parameters that need to be studied by comparing different transducer systems.

In both headsets the microphones are located in the ears or very close to the ears on both sides. When microphone signals are routed directly to earphones the system exposes a user to a binaural representation of the real acoustic environment around the user. However, in practice it is almost impossible to position and tune transducers so that the signals entering the listener's ears are identical to those in the open-ear case. The microphone elements are outside the ear, and therefore interaural time differences are larger and the effect of the pinna is smaller and distorted by the presence of the microphone–earphone element. The acoustic load to the ear is changed; see [72] for measurement data with the same earphone element as in model II.

A generic block diagram of the MARA system sketched in Section 1 is shown in Fig. 9. The signals entering the user's ears are composed of pseudoacoustic input signals, captured with microphones, and virtual sounds that may be speech signals from remote talkers or some other signals, such as recorded or synthesized announcements, advertisements, instructions, warnings, or music. Mixing of the two compounds is performed in the augmented reality audio (ARA) mixer. The preprocessing block shown above the head in Fig. 9 is used to produce the output signal of the system. Microphone signals are recorded and some preprocessing may be applied to produce a monophonic signal which, in a communications applications, is transmitted to another user. Alternatively, the binaurally recorded signal (remote pseudoacoustic signal for the other user, see Fig. 7) could be transmitted in its original form. Both options have been implemented in the software system introduced in [3].

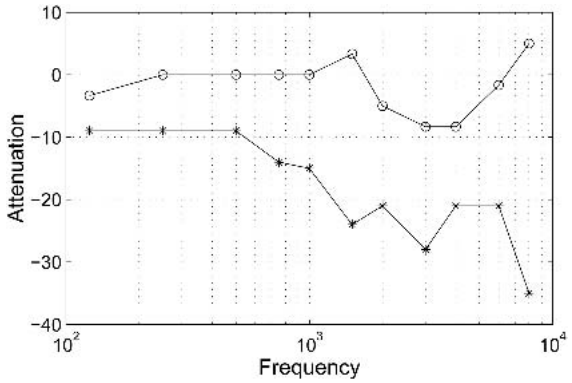

Fig. 8. Attenuation in model I (lower curve) and model II (upper curve) headsets.

4 LISTENING TEST

In MARA a goal is to produce an experience for the listener where the pseudoacoustic sound environment and virtual sound events merge into one perceived auditory environment. In the spirit of the Turing test discussed earlier, virtual sounds should be rendered with such a precision that the listener cannot be able to say which sound sources are real acoustic sources in the environment and which ones only played from the earphones the user is wearing. It is obvious that this is difficult to achieve in other than laboratory environments, where the listener's head can be assumed immobilized and the acoustic transfer functions from a source to the listener's ears can be measured accurately. Therefore we designed a listening experiment to see how close to the subjectively perfect augmented reality audio display we can get with the proposed system.

The basic configuration is illustrated in Fig. 10. We measure acoustic transfer functions from a loudspeaker to microphones in the headset and use the same responses to create virtual sounds for testing. In an ideal case the signals entering the user's ear canals in the case of the loudspeaker, that is, the pseudoacoustic case, and the virtual sound reproduction through the headset only should be identical. However, a real system is far from ideal. For example, measurement noise, leakage of the headset in loudspeaker reproduction, and small movements of the listener change the transfer functions.

In typical real-world applications transfer functions from a source location to the headset worn by the user are unknown or can be approximated only roughly. Therefore

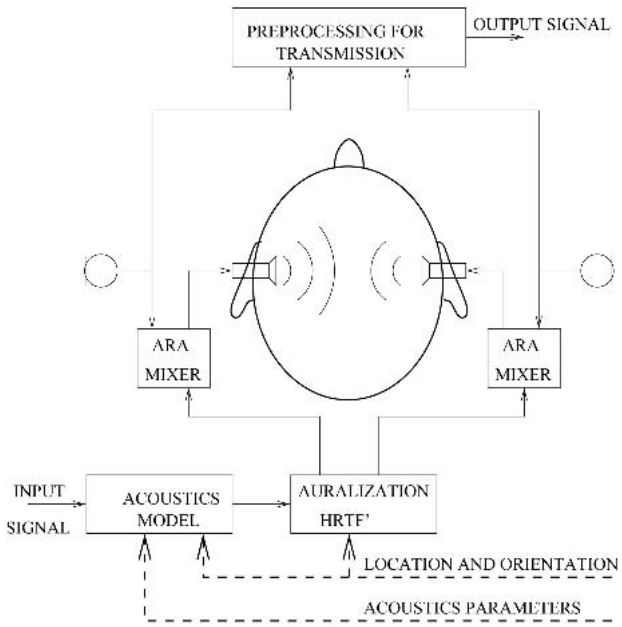

Fig. 9. Generic system diagram applicable for most augmented reality audio applications.

we also wanted to test how the results would change if measured transfer functions were replaced by generic synthetic responses.

4.1 Test Setup and Samples

The test was carried out in a standardized listening room [73] using an SGI Octane workstation with eight-channel audio output. Two of the playback channels were mixed to earphone signals together with directly routed pseudo-acoustic signals from the ear microphones. Other channels were connected to a set of Genelec 1030 loudspeakers in the listening room. This configuration made it possible to conduct listening experiments using both loudspeakers and earphones so that switching between earphones and loudspeakers will be free of clicks or changes in noise level. The test was an AB hidden reference test where samples A and B were at random either virtual sound presented through earphones or real sound from a loudspeaker. Listeners were allowed to switch between A and B at any time, and user responses were collected using the GuineaPig 2 listening test system [74]. The graphical user interface of the listening test system is shown in Fig. 11.

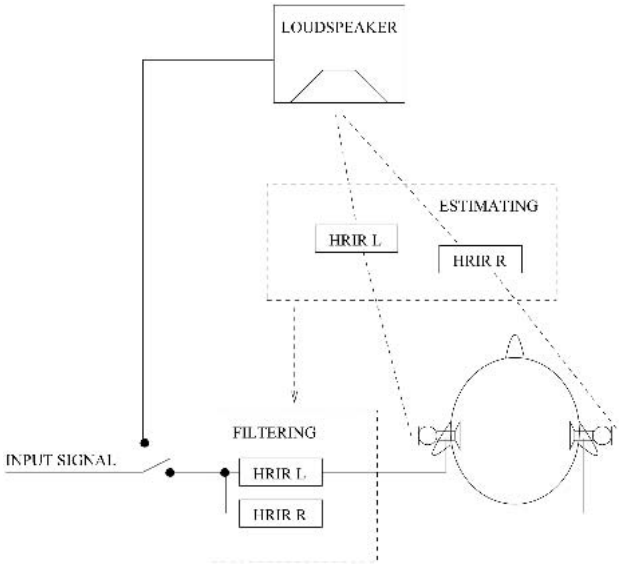

Fig. 10. Listening test configuration. Head-related room impulse responses (HRIRs) were measured for each listener. HRIRs were then used to synthesize test signals for headset playback. In listening tests signals were switched between loudspeaker and earphone playback.

The test setup consisted of a chair with a head support for the test subject and three loudspeakers placed 2 m away from the test subject, one in the front and the others 30° to each side of the subject. In order to remove the cue of identifying the virtual sound source by a slight error in localization, dummy loudspeakers were added to the loudspeaker array. A somewhat similar test setup based on in-situ binaural room impulse response (BRIR), measurements has been used earlier in studies on the externalization of virtual sources (for example, in [22, 41]).

In the beginning of each listening test session binaural impulse responses from the loudspeakers to the headset microphones in the test subject's ears were measured in situ. This includes the room response as well as the test subject's head-related impulse response modified by the earphone and the microphone location. The excitation signal used was a 1-second logarithmic frequency sweep signal. The binaural impulse response was then computed by deconvolution between recorded microphone signals and the excitation signal. The impulse response was truncated to 0.2 second, which is slightly less than the T_{60} reverberation time of the room [73].

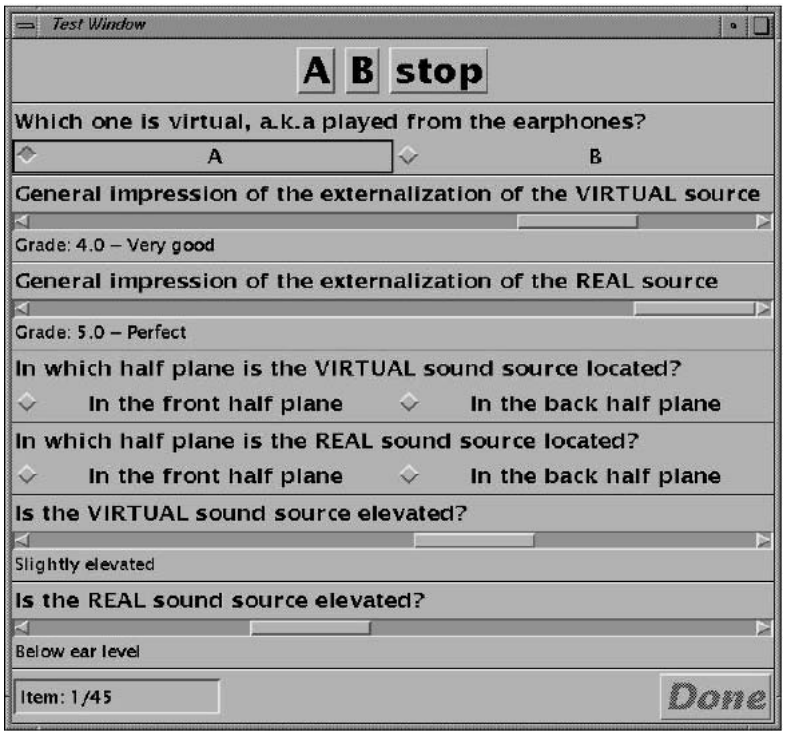

Fig. 11. Graphical user interface of listening test system.

Two sets of virtual sound samples were used in the test, all generated from ten original source signals. These dry source signals are listed in Table 1. The first set of virtual sound samples was generated by convolution between the measured binaural impulse response and the dry source signals. In the following these are called specific test samples because they were generated separately for each listener and each listening test session.

A second set of generic test samples was generated beforehand and it was identical for all subjects. These were produced by convolving the original test samples with impulse responses composed of a generic directional HRTF response and the output of a computational room model. The generic HRTF responses were measured from one subject in the same room, but they were truncated to 1 ms such that they represent the contribution of the direct sound only. The room model was based on a well-known concept of direct sound, early reflections, and statistical late reverberation (see [75] for a description of the software). Early reflections were synthesized using an image source model of the room [76], and late reverberation was produced using a reverberation algorithm proposed in [77].

4.2 Listening Test Procedures

In most tests the subjects compared specific or generic test samples played through the headset earphones against the original samples played back with the loudspeakers in the listening room. Both specific and generic virtual sound samples were used in each listening test session, and they were presented in random order. Most listeners participated only in one listening session, and there were no repetitions of the stimuli.

The duration of a sample was at maximum about 5 seconds. The sound pressure level in loudspeaker reproduction was adjusted so that it was approximately 60 dB (SPL) for test samples in the listening position. The level of the pseudoacoustic reproduction was adjusted by one of the authors (Ms. Jakka) to a fixed level such that the level of the pseudoacoustic environment was slightly amplified over the open-ear case but not annoying. Virtual sounds were played at a level where the loudness was subjectively close to the loudness of pseudoacoustic reproduction.

In the test there were a total of 36 sample pairs. In each pair one sample was played from a loudspeaker and the other, either a specific or a generic virtual sound sample, through the headset. The virtual sound sample was rendered either to the same direction as the active loudspeaker or to a different direction. Subjects were allowed to listen to samples A and B as many times as they wished, and they were forced to give various gradings to samples before proceeding to the next pair of samples (see Fig. 11). First listeners were asked to mark which one of the two signals (A or B) they thought was played from the loudspeaker. Next they had to grade the externalization of both signals. Finally they were asked to characterize the elevation of both signals using a slider in the user interface and indicate if the sound appeared in the front or the back hemisphere.

The grading scale for the externalization was 1 to 5, where 1 corresponds to hearing the sound originating from inside the head and 5 to when the sound appears at the loudspeakers. A similar scale was used also in [22]. In the listening test reported in this paper the subjects were not asked to assess the perceived location of a source in the frontal horizontal plane. Subjects were asked to grade the externalization, report possible front–back confusion, and give a grading for the elevation of assumed virtual and pseudoacoustic sources. The elevation was ranked on a grading scale from 0 to 5 so that the value 2 represented the ear level, 5 was for a source above the head of the listener, and 0 was given if the perceived source position was close to the floor.

An additional test was performed with babble-talk-type low-level background noise, which was played back from one of the loudspeakers during the test. This setup was intended to imitate a natural listening environment with distracting sources.

4.3 Results

The performance of the model II headset was tested with 19 subjects. Four subjects were considered naive and the rest expert listeners. Naive listeners were students at HUT who had no prior experience in spatial audio listening tests. Expert listeners were authors of this paper or personnel of the laboratory who all had participated in several related listening tests. Five subjects were female and fourteen were male. The test with babble-talk background noise was performed with six subjects. In addition, the model I headset was tested with three authors of this paper, who also participated in all other tests as subjects.

Table 1. Test sequences and mean externalization grades with 95% confidence intervals in parentheses.

Item	Description	Loudspeaker	Specific Virtual Test Sample	Generic Virtual Test Sample
1	Sound of breaking glass 1	4.9 (4.7, 5)	3.8 (3.3, 4.3)	2.7 (2.2, 3.3)
2	Sound of breaking glass 2	4.7 (4.5, 5)	2.7 (1.8, 3.5)	3.5 (3, 4)
3	Female singer	4.7 (4.5, 4.9)	3.1 (2.3, 3.9)	3.1 (2.6, 3.7)
4	Male speaker	4.8 (4.6, 5)	3.3 (2.7, 3.9)	2.7 (2.1, 3.2)
5	White noise	4.7 (4.5, 4.9)	3.9 (3.3, 4.5)	3.2 (2.7, 3.8)
6	Bell sounds	4.6 (4.3, 4.8)	3.8 (3.2, 4.3)	3.8 (3.2, 4.3)
7	Tuba long note	4.7 (4.5, 4.9)	4 (3.5, 4.6)	3.4 (2.9, 3.9)
8	Electric guitar chord	4.6 (4.3, 4.9)	3.7 (2.9, 4.5)	3.9 (3.3, 4.5)
9	English horn single note	4.9 (4.8, 5)	4 (3.4, 4.6)	4.1 (3.7, 4.6)
10	Pop music	4.8 (4.6, 4.9)	2.7 (2.1, 3.3)	2.5 (1.9, 3.1)

4.3.1 Source Confusion

In this paper the cases where the listener thought that the virtual sound came from the loudspeaker or vice versa are called source confusion. The percentage of 50% would mean that listeners cannot discriminate virtual sounds from the pseudoacoustic sounds played from the loudspeakers. Fig. 12 illustrates the percentages of source confusion for the ten test sequences of Table 1. The data are plotted separately for four different conditions indicated in the figure caption. The percentage in source confusion over all sequences and conditions was 18.5% for the model I headset with a 95% confidence interval of (9.0–28.0%). For the model II headset the mean was 14% with a confidence interval of (9.6–18.5%). In both types of headsets source confusion was equally common in both specific and generic virtual sound samples, and the differences between specific and generic responses were found statistically insignificant (p values of 0.96 and 0.92, respectively).

A high value of source confusion is approached in several test sequences reported in Fig. 12. It seems that the highest percentage of source confusion was found for the earplug-type model I headset. However, recalling that the number of test subjects for the model I headset was only three, the comparison between average percentages may not give a good accuracy. In the paired T-test results the difference between the source confusion percentages in the two headsets was not found statistically significant (p value of 0.38). The percentages for naive and expert listeners over all tested conditions using the model II headset were 13.9% and 14.1%, respectively, but the difference, in terms of the paired T test, was statistically insignificant (p value of 0.95), and the confidence intervals were equal.

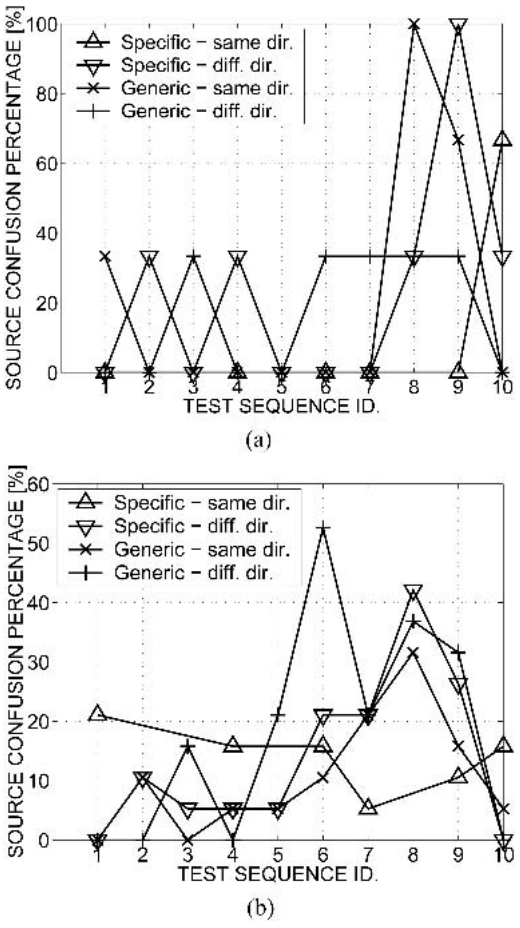

Fig. 12. Percentages of source confusion (a) Model I headset. (b) Model II headset. △, ▽—data with specific responses where virtual source is rendered to same direction as loudspeaker and to different direction, respectively. ×, +—similar setting using generic responses.

Source confusion is very common in narrow-band sounds. For the musical instrument sound samples 6, 7, 8, and 9 (see Table 1), the percentage of source confusion is 24.2% with a 95% confidence interval of (17.1–31.3%). In other sequences the percentage is only 6.8% with a confidence interval of (3.5–10.1%).

If a virtual sound sample in an AB test is rendered to a different direction than the loudspeaker, the source confusion is more likely to occur. In the average of both specific and generic cases of virtual sounds the source confusion percentages are 11.8% and 15.8% for the cases where the sounds are coming from the same direction and a different direction, respectively. That is, source confusion is more likely in the case where the virtual sound is rendered to a different direction than the loudspeaker. In a paired T test the difference was found to be statistically significant (p value of 0.0001).

The experiments with babble-talk background noise show little effect on the percentage of source confusion. This was only tested with the model II headset. The percentage of source confusion with babble-talk noise over all test sequences is 15.6%, but compared to the case where no background noise is present (14.1%) the difference is statistically weak (p value of 0.43). Several subjects reported that the background noise may increase differences at extremes. When identifying the location of the source is difficult, the background noise makes it even more difficult, whereas the easy cases become easier.

4.3.2 Externalization Grades

The median externalization grades for all listeners, both headsets, for the cases of real source and for the two different types of virtual sound items (specific and generic) are shown in Fig. 13. Since there were only three subjects testing the model I headset, all externalization grades were plotted in Fig. 13(a). The median and the lower and upper quartile values of the externalization grades for the model II headset, collected from 19 subjects, are illustrated in Fig. 13(b). In the model I headset the sounds from the loudspeakers and the specific virtual sounds are close to each other, whereas the generic samples gave lower grades. The mean externalization grades in the three cases (real, specific, and generic) are 3.9, 3.6, and 2.8, respectively. Note that the samples played from the loudspeakers (top panel) often gave grades less than 5, showing that the sound from the loudspeaker was not well externalized in using the model I headset. In fact, in some cases subjects reported that when using the model I headset both virtual and pseudoacoustic signals were localized inside the head of the listener.

The one-way ANOVA analysis followed by a paired T test of the data from the model I headset indicates that the difference in externalization grades for real and specific responses was only weakly significant (p value of 0.1). The model I headset has a high attenuation of the leakage, and therefore the cases of sound from a loudspeaker and virtual sounds produced using specific responses are similar. Recall that in the case of ideal transducers with no leakage, and with an immobilized listeners' head, the signals entering the subject's ear canals would be identical in both cases. The fact that the samples auralized using generic responses gave clearly lower grades than the real and specific sources was found to be a statistically significant phenomenon (p value of 0.001) in the one-way ANOVA test.

Fig. 13(b) shows the median statistics of the listening tests with the model II headset. Unlike in the model I case, the loudspeaker (real) received often almost the maximum grade whereas both specific and generic virtual samples received lower grades. Average values over all data are 4.7 (4.6–4.8), 3.5 (3.3–3.7), and 3.3 (3.1–3.5) for real, specific, and generic cases, respectively. The 95% confidence interval are indicated in parenthesis. The high externalization grades in the case of sounds from the loudspeaker may partially result from subjects biasing in the case where they are sure which sample was played from the loudspeaker. It is possible that in these cases subjects often gave the externalization grade of 5 for the real sound source, even if they were instructed to give the grade by the perceived externalization of the source.

The range of the confidence intervals in the mean results suggests that there are significant differences between test sequences. The average range of the confidence intervals for the loudspeaker case is only 0.2 unit on the externalization grade scale, but in both cases of virtual sound samples the confidence interval has a range of 0.4 unit. The difference between virtual sounds (specific and generic) and those played from loudspeakers was statistically significant, whereas the difference between the two virtual cases, that is, specific and generic, was only weakly significant (p value of 0.1). The lower and upper quartile values marked by a whisker around the mean value give an

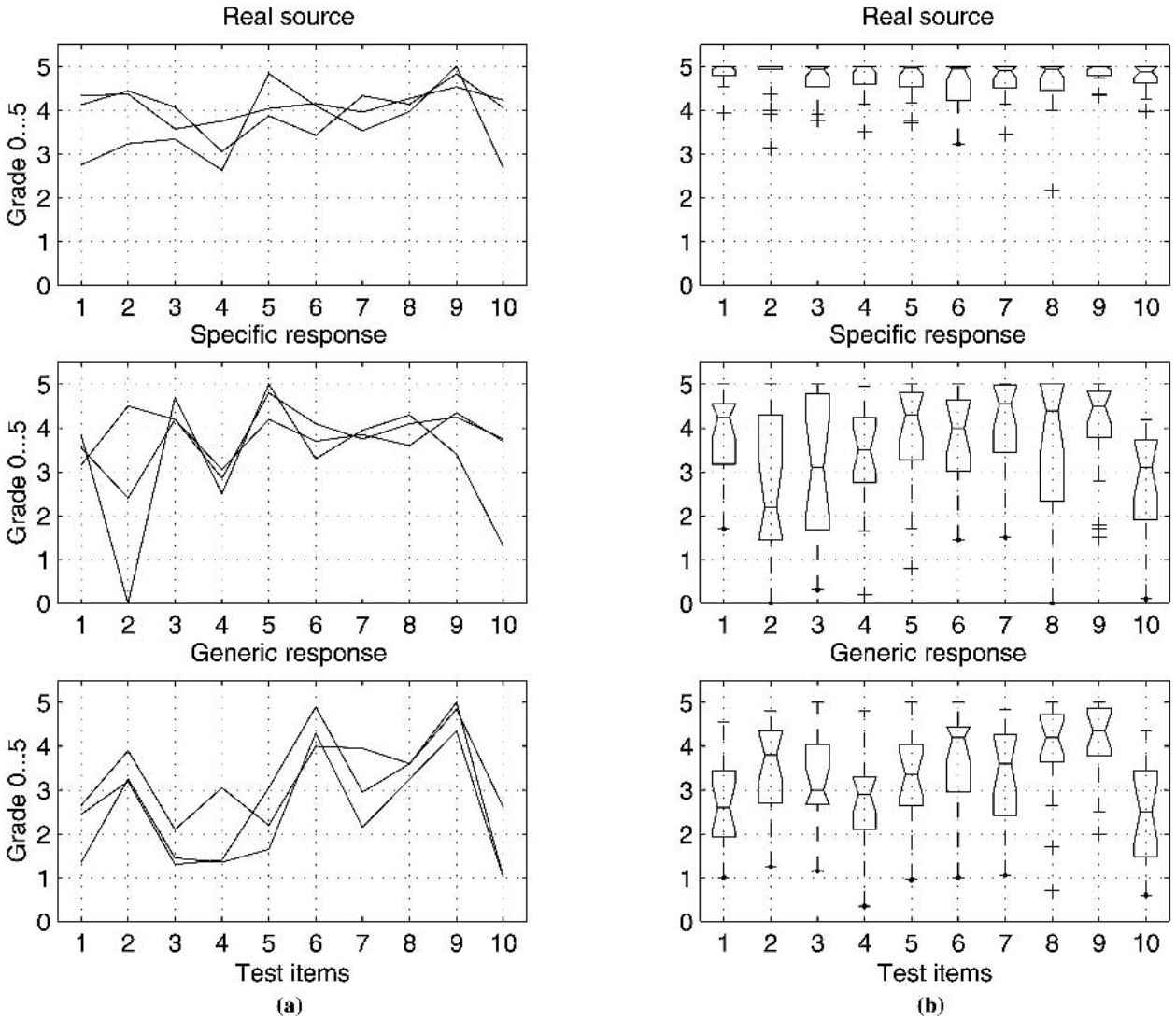

Fig. 13. Median values of externalization grades over all subjects for reference sound from a loud-speaker (top) and for virtual sound produced by filtering with specific (middle) and generic (bottom) responses. (a) Model I headset. (b) Model II headset. Median value is indicated by a horizontal line inside a box representing the region between lower and upper quartile values of data.

indication of the intersubject differences. The average ranges between the lower 25% quartile and the upper 75% quartile in the real, specific, and generic cases, averaged over all test sequences, were 0.4, 1.9, and 1.5. That is, the largest intersubject differences were found in virtual sounds auralized using specific in-situ responses, measured individually for each subject.

4.3.3 Front–Back Confusion and Elevation Phenomena

Front–back confusion is a well-known phenomenon in localization experiments. In the current tests this occurred in both loudspeaker reproduction and virtual sounds. In the case of a real sound from a loudspeaker the percentage of front–back confusion in a model II headset was 1.3% (0.3–2.3%). However, for the virtual sound samples produced using specific and generic HRIRs the percentages were 23% (17.9–28.1%) and 19% (15–22%), respectively. The significance that the percentage of front–back confusion in the current experiment was larger in the case of specific HRIRs than for generic responses is relatively low (p value of 0.19). This is a surprising finding because it is generally assumed that the use of individualized HRTFs helps in resolving front–back confusion. The difference between real and virtual sources in the percentage of front–back confusion was statistically significant.

There are large individual differences in the phenomenon of front–back confusion. In the bitmap of Fig. 14 the front–back confusion of virtual sound samples is shown in black. Here the columns represent the 36 test sequence pairs and the rows are the 19 subjects. For three test subjects, a majority of virtual samples were localized behind the head. Interestingly, those three subjects were originally ranked expert listeners. For two subjects even some of the loudspeaker samples were localized behind the head. In the test situation the front–back confusion can be explained by the fact that the pinna was partially blocked by the earphone and the head rotation was hindered.

An elevation of both real and virtual sources was commonly observed in the listening test. The average values and the 95% confidence intervals for the model II headset for real, specific, and generic samples were 2.28 (2.2–2.4), 2.3 (2.0–2.7), and 2.5 (2.1–2.9), where the value 2 corresponds to the ear level and 5 would be given if the perceived position of the source were above the head of the listener. In the paired T test the difference between loudspeaker sounds and specific virtual sounds was not significant, but the difference between loudspeaker sounds and those rendered using generic responses was found somewhat significant (p value of 0.04). The elevation of sound sources also varies individually, which may occur because of the ear plugs fitting into different ears in different positions. For two test subjects the virtual source samples were on average elevated above the ear level, and for one subject some of the real source samples were elevated. The extreme value of 5.0 occurred in 0.5% and 1.7% of the real and virtual samples, respectively. In the model I headset the mean grades for the elevation in the cases of real and virtual sounds were 2.4 (2.3–2.5) and 2.5 (2.3–2.6), showing a slightly larger elevation than in the case of the model II headset. Here the difference between the two types of sound, virtual and loudspeaker, was found statistically insignificant.

It could be expected that the sound samples played from the front center loudspeaker direction would not externalize as well as those played from the sides. The average externalization grade in the model I headset for virtual sounds played from the center loudspeaker was 3.4 (3.1–3.6), whereas for the directions of 30° to the left and right the average was 3.6 (3.4–3.7). For real samples played from the center loudspeaker the average externalization grade was 4.6 (4.4–4.7), and for the two other loudspeakers it was 4.65 with a 95% confidence interval of (4.60–4.77). Even if the mean values do show this trend, in terms of the paired T test the degradations in the externalization grades in the two cases are not significant (p values of 0.2 and 0.1).

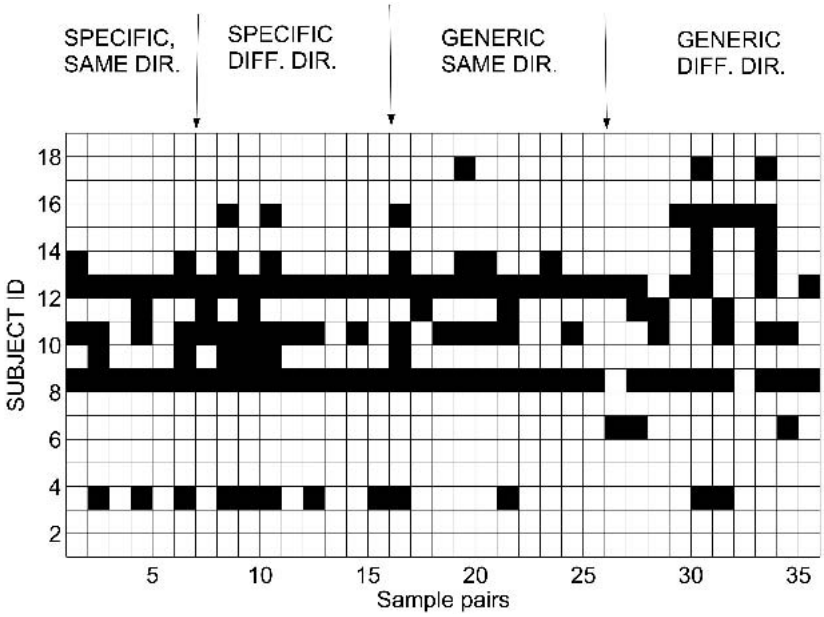

Fig. 14. Back–front confusion in virtual sound samples for 36 sample pairs and 19 subjects. Confusion is indicated by black color.

4.3.4 Discussion of Results

Perceived locations of virtual and pseudoacoustic sources were not compared in the listening tests. Many subjects reported that virtual and pseudoacoustic sources were often externalized to the same degree, but their perceived locations were slightly different, even if the virtual source was rendered to the position of the loudspeaker. Nevertheless, source confusion is common for many of the test sequences. For example, the average percentage of source confusion in the model II headset was 14%. Assuming that there were an equal number of right guesses in the listening test, we may anticipate that in 28% of the test cases a listener was not able determine which sound was played from a loudspeaker and which was a virtual sound from the earphones. Source confusion occurs both with binaural responses measured in situ and as with more generic binaural responses with a nonpersonalized direct sound response and a computationally modeled room response. In the case of the model II headset the difference between the two methods was found statistically insignificant.

In the current paper we did not perform comprehensive tests to assess the intrasubject differences, that is, how results vary when the same test is repeated several times for the same subject. However, in preliminary tests we found that the intrasubject differences seem to be smaller than the intersubject differences that have been reported here.

Since only three subjects were tested with the model I headset the comparison of the two different headsets may not give accurate results. Nevertheless it seems that source confusion is more common with the model I headset. However, the lack of externalization of the real source sound may be a larger problem in model I than in model II. In both headsets the difference between specific and generic virtual samples is small. In a real-world application we would most likely always use generic HRIRs rather than in-situ measured responses. However, the difference in externalization grades between specific and generic responses was found only weakly significant in the model II headset whereas in the model I headset a larger difference was observed. Therefore we may argue that the model II headset may be the more promising alternative as a transducer with respect to the MARA objective of pseudoacoustically imitating the real audio environment.

It is probable that the major factor causing the difference between virtual and pseudoacoustic sounds is the leakage of direct sound to the ear past the headset system in the case of loudspeaker playback. The leakage is naturally completely absent in the virtual production case. The frequency responses of the earphones may not be as flat as would be desired, which may cause coloration of the virtual sound. Measurement of the actual transfer characteristics of different types of headsets using probe microphone methods is work currently in progress, and it will probably give a more accurate picture of the leakage effect.

In order to study the perception of the augmented audio environment more carefully, the listening test subjects should be made to adapt to the pseudoacoustic environment prior to the test. This could be done by letting subjects wear the headsets for a longer period of time, such as two hours or a day, before performing the test. In addition, the listening test should be repeated at different reproduction levels for the pseudoacoustic environment.

5 ENABLING MOBILITY AND WEARABILITY

The reported results show that the proposed transducer configuration may be useful in implementing mobile augmented audio reality applications. However, room acoustics and the location of a user in the room was fixed. Listeners' heads were also immobilized, and the implementation of the system on a desktop computer platform is currently far from being mobile and wearable [3]. In the following sections we give a brief overview on processing principles and technologies needed to make the system adapt dynamically to new environments and movements of a user. In particular, we discuss techniques needed to control the top and bottom blocks of the schematic diagram of Fig. 9: acoustics model, dynamic auralization, and preprocessing for transmission. We will also give some ideas on the user interface and the relation of the present work to future mobile communications technology.

5.1 Processing of Binaural Signals

The block diagram of a MARA system proposed in this paper is illustrated in Fig. 9. In a dynamic setting it is necessary to provide the system control information that should be obtained from the environment. Details of binaural signal processing are beyond the scope of this paper.

Acoustic parameters, such as the amount of reverberation and the geometric properties of the real environment around the user may be needed for a successful augmentation of the pseudoacoustic environment. There are basically two mechanisms for acquiring this information. In some applications this can be derived off-line, and it would be available for the system as information related to that particular location. This would rely on a global localization system and an extensive database. This is basically the approach taken in [65]. A more generic alternative would be based on an on-line estimation of the acoustic parameters from microphone input signals.

Probably the most important tool for the analysis of binaural signals is the cross-correlation function between two microphone signals. In a free field with two microphones the cross-correlation function corresponding to a single source position is an impulse function, where the time of the impulse indicates the time delay between the two microphone signals. When the head is placed between the two microphones, the cross-correlation function is changed such that it is convolved with the cross-correlation function of the two HRIRs. In practice the effect of the head is that we get a slightly blurred cross-correlation peak.

In a diffuse reverberant field with two omnidirectional microphones the cross-correlation function is a rectangular pulse [78], [79]. Head diffraction again changes the correlation function slightly [80]. However, due to the folding effect, a diffuse field produces almost uniform distribution of energy in the cross-correlation function. The magnitude

of correlation peaks versus total energy in the cross-correlation function could be used to estimate the ratio of direct-to-diffuse sound in the field.

Locations of peaks in the cross-correlation function indicate the directions of sources in the acoustic field. Estimation of the peak positions of cross-correlation functions is a popular method in the localization of sources from microphone array signals [81]. A source localization method based on an array of two microphones has been studied, for example, in [82], and binaural microphone arrays have been used in such studies as [83] and [84].

In localization, the array (headset) is assumed to be static and the system is locating static or moving sources. If, on the other hand, the sound source or sources are assumed to be static and the headset worn by the user is used to locate these static sources, then the exact position and orientation of the user can be solved. A technique for head tracking based on binaural microphone signals of the proposed system was recently introduced [85]. This approach has the advantage that an unlimited number of users can use the same sound sources for positioning.

The bidirectional telepresence application illustrated in Fig. 7 has some specific problems that have been studied in a recent publication [86]. One of the problems comes up when a remote user starts talking. The user at the near end will hear the remote talker localized inside his/her head. One way to avoid this rather unnatural experience is to detect the speech activity of the remote user and separate and rerender his/her voice outside the other user's head. Since the distance from the user's mouth to the two ear microphones is the same, one can use efficiently the difference between the frequency-selective sum and difference of the two microphone signals in the detection and separation of the voice of the user. In fact, summing of the two signals is essentially beamforming [87], which in the case of two microphones can give up to 6 dB amplification for the speech of a user and also other sounds sources in the median plane. When this is done separately and adaptively in different frequency regions, the cancellation of sources in other directions can be performed at the same time. Frequency-selective sum–difference beamforming is actually a special case of the celebrated generalized sidelobe canceler (GSC) introduced in [88] and used for spatial voice activity detection in [89].

The use of beamforming techniques for a head-worn microphone array or for binaural microphone signals has been proposed for hearing aid devices by many authors [90], [91]. Similar services could also be provided with the proposed MARA system. In addition, localization and beamforming could be combined with sound recognition techniques to facilitate spatial decomposition of the auditory space at a higher conceptual level. Automatic speech recognition is a classical example, but the recognition of music or other types of environmental sounds [92]–[94] or sound environments [95] using the binaural input signals could be considered.

5.2 Context and Location Aware Services

Many of the applications and services discussed in this paper are based on the knowledge of the location and orientation of the user. The context of the services is strongly related to the user's location. Tracking of a user's location can be performed on two different levels: globally and locally. For some services it is sufficient to track the user's global location in the range of meters. This information can, for example, be in front of the Helsinki railway station, at an information desk, or in the science fiction section in a library. Outdoors global positioning is often achieved by using satellite positioning systems. At the moment there are two radio navigation satellite networks for global positioning, the Global Positioning System (GPS) and the Global Navigation Satellite System (GLONASS). There are also regionally restricted navigation networks, such as EGNOS/GALILEO which covers Europe.

The global satellite systems do not work indoors because the buildings block the satellite signal. In cities there may also be problems because of the canyoning effect caused by high buildings around the user. For indoor positioning there are a variety of methods available. In the Active Badge system [96] a user wears a badge that emits an infrared signal, which is detected by a sensor. The signal can contain a globally unique identifier (GUID) and thus will allow identification. The Active Bat location system [97] uses an ultrasound time-of-flight lateralization technique to provide the physical location of an Active Bat tag. The tags emit ultrasonic pulses, synchronized by radio frequencies, to a grid of ceiling-mounted sensors. The distance measurement data are then forwarded to a central unit, which calculates the position. The tags can also have GUIDs for recognition. Instead of central computing, the Cricket Location Support System [98], developed at MIT, lets the user calculate the location. This can be done, for example, in a PDA. The infrastructure of the system consists of transmitters, beacons, which emit both radio and ultrasound signals, and receivers, which listen to radio and ultrasound pulses. Radio frequencies are used for synchronization. RADAR [99], developed by the Microsoft Research group, is an RF-based localization system using the IEEE 802.11 WaveLAN wireless networking technology. The system consists of multiple base stations and wireless transmitters worn by the user. The transmitter emits RF signals, and the signal strength is measured at multiple base stations. These data are compared to the empirical measurement data, and thus the user location can be estimated. SpotOn [100] is an example of an ad hoc application where no infrastructure is implemented. The system implements lateralization with low-cost tags, and the user location is estimated by detecting the radio signal attenuation. Technology based on Radio-frequency identification (RFID) tags provides new interesting possibilities for the localization of objects in the environment [101], [102]. An overview of related positioning systems can be found in [103].

In some services it is essential to track the user's local position as well. By this we mean continuous tracking of the user's exact position and head orientation. This information is needed for a successful augmentation of the acoustic environment. In order to track head orientation and movement the user needs to wear some sort of sensors

or transmitters. Modern head trackers can be divided into five categories based on their tracking method: acoustic, electromagnetic, inertial, mechanical, and optical. In acoustic head tracking the tracking can be done by detecting the flight times or phase differences for arriving sound in the microphone array. In electromagnetic tracking, a stationary element emits a pulsed magnetic field. A sensor worn by the user senses the field and reports the position and orientation to the source. Inertial trackers have miniature gyroscopes inside the sensor. The rapidly spinning wheel tries to resist any movement. The resistance can be measured and converted to position and orientation. Mechanical trackers have a physical connection between the reference point and a target. The orientation of the connection point can easily be measured and turned into the desired coordinates. In optical tracking an array of cameras locates a set of LEDs in known positions. The LEDs can be worn by the user and the cameras can be placed in the environment, or vice versa. An overview of related tracking devices can be found in [103].

For contextual services it is important that the positioning infrastructure can exchange information with the user. The hardware worn by the user can send, for example, identification data and the user's preferred language to the service provider. In return the service provider can offer personalized services to the user. Also the acoustical parameters of the environment (for example, reverberation time and reference points for positioning) could be transmitted to the user's hardware for proper augmentation of the services. As the user changes place, for example, going from a library to a movie theater, the system should automatically detect the available services. This requires common protocols for all related services to communicate with each other. One such architecture is proposed in [104].

The headset proposed in this paper is a candidate for acquiring information on the surroundings of the user in conjunction with other devices such as a mobile phone or a positioning device. The headset is also a very good way to make the aggregate information available to the user. Some services may have been tied to a certain location. Once the infrastructure has identified the user, these services can be presented via the headset as the user approaches them. Audio events can also be easily rendered to a desired position in the acoustic environment. As an example a user can be in a store where special offers are played to the user as he or she walks in the store. After shopping the user goes, with navigation instruction played by the headset, to a cafe. Someone might have had left an acoustic Post-it message to the user in the cafe. As the cafe's infrastructure identifies the user, the message is played through the headset.

5.3 User Interface

One of the basic ideas of the headset configuration described in this paper is that the user may wear the headset at all times. Because of the pseudoacoustic playback, it is not necessary to take off the headphones for live discussion with other people. This would make the headset an ideal means for communicating contextual information to the user. It would also facilitate using a mobile device without seeing the screen of the device, that is, eyes-free usage, and may also solve the current limitations on mobile device usage as a secondary task, during critical primary tasks such as driving or cycling.

Eyes-free interaction between the device and its user would naturally require efficient means for both input and feedback information. For input there are two main ways to communicate with the device: speech and haptics. Speech would be ideal for rapid and natural communication. However, natural speech communication requires very sophisticated techniques, still mostly in their basic research phase, and simpler speech input methods may often be irritating to the user. For example, navigating a hierarchical menu structure using speech prompts is both irritating and slow, although fairly easy to implement. There are also many usage situations where speech is not an optimal input modality, as talking aloud may not be socially acceptable.

For feedback from the device, both speech and nonspeech sounds can be used effectively. Text-to-speech (TTS) synthesis is a viable option for getting complex information to the user, while nonspeech (earcons and auditory icons) output is suited well for simpler concepts, such as alerts and simple user action feedback [105]. A practical experiment of applying sonification by earcons to a mobile phone user interface was carried out in [106]. The hierarchical structure of a mobile phone menu with four levels was represented with the help of structured musical tones and implemented in a mobile phone. The result of a user study indicated that the sonification was found irritating but also helpful in some situations.

Spatial audio is also an important aspect of auditory user interface presentation. In a similar way to the spatial organization of information in graphical user interfaces, where multiple visual objects can be easily displayed and managed, space could be exploited to present the auditory objects of an interface to the user wearing the MARA headset system. User interface design for spatial auditory displays has been considered by Cohen and Ludwig [107], who developed an audio window system where the user can monitor and manipulate spatially separated sound items.

The binaural impulse response technique employed for virtual source generation in Section 3.2 could be easily applied to the user interface (UI) applications on the MARA system. However, the common problems of externalization and perception of distance encountered in virtual audio presentations over headphones should be considered carefully to guarantee perceptual validity and high usability of these audio UI solutions. The use of echoic stimuli to improve externalization and control distance perception would need careful perceptual assessment in order to ensure a consistent audio UI presentation for different source distances while keeping a uniform sense of space for different audio objects of the UI. Recently, the issue of distance control, for instance, has been investigated by Martens [108], who presented a technique for the control of a virtual sound source range.

The design of a near-field virtual display is also an interesting option for audio UI applications, as nearby virtual sources such as sources perceived in the region of

space within "arm's reach," offer a potential for multimodal interaction with sound objects using gestural control, as suggested in [109]. An interesting application of this idea was presented by Cohen et al. [6], who developed the Handy Sound system using a DataGlove and Polhemus devices to control an audio UI. A distinction between distant virtual sources and nearby sources can also be made on the basis that nearby sources may relate more to the listener's private space or "personal space," as described in [108]. In most applications a spatial audio UI would correspond better to this type of user space.

Another issue to consider in audio UI design relates to the structure of the UI displayed to the user. This topic has been studied in several research works on auditory mapping of graphical UI and design for nonvisual UI. The GUIB (textual and graphical user interfaces for blind people) project [110] and the Mercator project [111] illustrate two different approaches to auditory UI presentation, as described in [112]. In the GUIB system a direct analogy to the spatial presentation in visual interfaces is proposed with the spatial audio technique. This is the approach also chosen in the audio window system presented by Cohen and Ludwig [107]. The Mercator project utilizes another approach focusing on the presentation of object hierarchies in a UI, rather than a pixel-by-pixel level of the interface. These two approaches have been combined subsequently with the idea of a hierarchical navigation system based on direct manipulation with spatial audio [113].

The design of an audio interface display depends mainly on the type of UI application considered. However, in the context of the MARA system, the mobile usage factor would require an easy to use and efficient audio UI solution. In the case of a hierarchical menu structure presentation, as commonly found on mobile devices, the spatial UI design presented in [114] could easily be adopted. This nonvisual interaction solution was successfully applied to a music play-list navigation and generation for an audio player application [115]. Due to the limitations in sound localization with nonindividual HRTFs, that is, front–back confusion and elevation problems, this auditory display uses only a small number of sound positions presented along a line from left to right. This simple approach ensures that the spatial sound presentation will work reliably for any user with normal hearing. In the implementation, stereo amplitude panning can be combined with a simple stereo widening technique [116] to externalize the left- and rightmost spatial locations. With this scheme it is possible to present any traditional hierarchical menu structure purely by audio, considering both speech and nonspeech sounds, depending on the UI application.

5.4 Infrastructure Support for MARA

The headset presented in this paper is essential for the concepts of mobile augmented reality audio. However, there are many applications where some type of support from the underlying infrastructure would be beneficial or required. The headset is just the acoustic front end for many applications. For location-based services, the possibilities of the headset providing the necessary information are remote, so the applications have to rely on the information provided by other systems. As already discussed previously, positioning services may be provided by the mobile network, by satellite positioning systems, or by several other methods. Also locations relative to a local environment, such as a car, are best provided by the environment itself. Knowledge of the acoustics around the headset can be acquired partially from the microphone signals of the headset itself. However, this information could also be contained in the environment. For the telepresence application mentioned in Section 2.1 the most critical infrastructure requirement is to be able to transmit a high-quality full-duplex stereophonic signal. This is something that the current mobile network infrastructures and standard codecs do not support. These applications can currently be tested with nonstandardized configurations in IP-based networks. In addition, to make the headset itself truly wearable, data transmission to and from the headset should preferably be wireless. There are many ways to implement this but, for example, the current Bluetooth profiles do not support high-quality duplex transmission. Still, interoperability here would be a key for enabling the use of the proposed headset as the only pair of headphones the user has to wear.

6 CONCLUSIONS

In this paper we presented an overview of the concept of mobile augmented reality audio (MARA). In the case of open ears a listener may perceive the acoustic environment in its natural form. When wearing a headset with binaural microphones and earphones user are exposed to a modified representation of the acoustic environment around them, which here is called the pseudoacoustic environment. A traditional example of a pseudoacoustic environment is that of a person wearing a binaural hearing aid device. The virtual sound environment is a synthesized binaural representation or a binaural recording. In augmentation of the user's acoustic environment, virtual sound objects are rendered in a natural or a pseudoacoustic representation of the sound field around the user. In this paper, we focused on the latter case, where virtual sound objects are combined with a pseudoacoustic representation of the acoustic environment around a user. This is obtained using a specific headset where miniature microphone elements are integrated into small earphone elements. Pseudoacoustic representation is produced by routing microphone signals to the earphones.

The proposed transducer system for MARA was studied in listening tests where the subjects were presented real sounds from loudspeakers and virtual sounds through microphone–earphone systems in the same setting. Virtual sounds were produced by filtering with in-situ HRIRs or nonindividual HTRF responses combined with a synthetic room model. It was found that even experienced listeners often cannot discriminate virtual sounds from test sounds coming from loudspeakers in a listening room. In addition, the difference between using individualized or generic HRTF filters was small or vanishing. The small difference may be related to the fact that the effect of the pinna is relatively small in the proposed system because micro-

phones are placed just a few millimeters outside the ear canal and the headset is partly covering the pinna. However, a more fascinating hypothesis is that combining virtual sounds with a pseudoacoustic sound environment actually makes rendering virtual sounds easier. This hypothesis needs to be tested by listening tests in the future.

The results are encouraging and suggest that the proposed transducer system may provide a potential framework for the development of applications of mobile augmented reality audio. No long-term exposure tests with the proposed headset were performed to see how users can accommodate the modified representation of the real acoustic environment. This will also be part of future work.

7 ACKNOWLEDGMENT

A. Härmä's work was partially supported by the Academy of Finland. The authors are grateful to the Nokia Research Center for financial support. The authors wish to thank Hanna Järveläinen for help in the analysis of the listening test results.

8 REFERENCES

[1] K. Brandenburg, "Perceptual Coding of High Quality Digital Audio," in *Applications of Digital Signal Processing to Audio and Acoustics,* M. Kahrs and K. Brandenburg, Eds. (Kluwer, Boston/Dortrecht/London, 1998), pp. 39–83.

[2] J. Rapeli, "Umts: Targets, System Concept, and Standardization in a Global Framework," *IEEE Personal Commun.,* vol. 2, pp. 20–28 (1995 Feb.).

[3] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, H. Nironen, and S. Vesa, "Techniques and Applications of Wearable Augmented Reality Audio," presented at the 114th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 51, p. 419 (2003 May), convention paper 5768.

[4] R. D. Shilling and B. Shinn-Cunningham, "Virtual Auditory Displays," in *Handbook of Virtual Environments Technology,* K. Stanney, Ed. (Lawrence Erlbaum Assoc., Mahwah, NJ, 2000), ch. 4.

[5] M. W. Krueger, *Artificial Reality II* (Addison-Wesley, Reading, MA, 1991).

[6] M. Cohen, S. Aoki, and N. Koizumi, "Augmented Audio Reality: Telepresence/VR Hybrid Acoustic Environments," in *Proc. IEEE Int. Workshop on Robot and Human Communications* (Tokyo, Japan, 1993 Nov.), pp. 361–364.

[7] C. Müller-Tomfelde, "Hybrid Sound Reproduction in Audio-Augmented Reality," in *Proc. AES 22nd Int. Conf. on Virtual, Synthetic and Entertainment Audio* (Espoo, Finland, 2002 June), pp. 58–63.

[8] R. W. Massof, "Auditory Assistive Devices for the Blind," in *Proc. Int. Conf. on Auditory Display* (Boston, MA, 2003 July), pp. 271–275.

[9] C. Kyriakakis, "Fundamental and Technological Limitations of Immersive Audio Systems," *Proc. IEEE,* vol. 86, pp. 941–951 (1998 May).

[10] C. Kyriakakis, P. Tsakalides, and T. Holman, "Surrounded by Sound," *IEEE Signal Process. Mag.,* vol. 16, pp. 55–66 (1999 Jan.).

[11] T. Caudell and D. Mizell, "Augmented Reality: An Application of Heads-up Display Technology to Manual Manufacturing Processes," in *Proc. 25th Hawaii Int. Conf. on System Sciences,* vol. II (Hawaii, 1992 Jan.), pp. 659–669.

[12] P. Milgram and F. Kishino, "A Taxonomy of Mixed Reality Visual Display," *IEICE Trans. Inform. and Sys.,* vol. E77-D, no. 12, pp. 1321–1329 (1994).

[13] H. Tamura, H. Yamamoto, and A. Katayama, "Mixed Reality: Future Dreams Seen at the Border between Real and Virtual Worlds," *IEEE Computer Graphics and Appl.,* vol. 21, pp. 64–70 (2001 Nov./Dec.).

[14] R. Held, "Shifts in Binaural Localization after Prolonged Exposures to Atypical Combinations of Stimuli," *J. Am. Psychol.,* vol. 68 (1955).

[15] T. Ilmonen, "Mustajuuri—An Application and Toolkit for Interactive Audio Processing," in *Proc. Int. Conf. on Auditory Display* (Espoo, Finland, 2001 July).

[16] A. M. Turing, "Computing Machinery and Intelligence," *Quart. Rev. Psychol. Phil.,* vol. 109 (1950 Oct.).

[17] N. Sawhney and C. Schmandt, "Nomadic Radio: Speech and Audio Interaction for Contextual Messaging in Nomadic Environments," *ACM Trans. Computer–Human Interaction,* vol. 7, pp. 353–383 (2000 Sept.).

[18] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia* (Academic Press, New York, 1994).

[19] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1999).

[20] G. Plenge, "On the Differences between Localization and Lateralization," *J. Acoust. Soc. Am.,* vol. 56, pp. 944–951 (1972 Sept.).

[21] B. M. Sayersu and E. C. Cherry, "Mechanism of Binaural Fusion in the Hearing of Speech," *J. Acoust. Soc. Am.,* vol. 29, pp. 973–987 (1957 Sept.).

[22] W. M. Hartmann and A. Wittenberg, "On the Externalization of Sound Images," *J. Acoust. Soc. Am.,* vol. 99, pp. 3678–3688 (1996 June).

[23] D. Hammershøi and H. Møller, "Methods for Binaural Recording and Reproduction," *Acta Acustica* (with *Acustica*), vol. 88, pp. 303–311 (2002 May/June).

[24] P. Zahorik, "Assessing Auditory Distance Perception Using Virtual Acoustics," *J. Acoust. Soc. Am.,* vol. 111, pp. 1832–1846 (2002 Apr.).

[25] N. Sakamoto, T. Gotoh, and Y. Kimura, "On 'Out-of-Head Localization' in Headphone Listening," *J. Audio Eng. Soc.,* vol. 24, pp. 710–716 (1976 Nov.).

[26] D. R. Begault, "Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems," *J. Audio Eng. Soc.,* vol. 40, pp. 895–904 (1992 Nov.).

[27] R. A. Butler, E. T. Levy, and W. D. Neff, "Apparent Distance of Sounds Recorded in Echoic and Anechoic Chambers," *J. Experim. Psychol.,* vol. 6, pp. 745–750 (1980).

[28] P. Minnaar, S. K. Olesen, F. Christensen, and H. Moller, "The Importance of Head Movements for Binaural Room Synthesis," in *Proc. Int. Conf. on Auditory Display,* (Espoo, Finland, 2001 July), pp. 21–25.

[29] D. R. Perrott, H. Ambarsoom, and J. Tucker, "Changes in Head Position as a Measure of Auditory Localization Performance: Auditory Psychomotor Coordination under Monoaural and Binaural Listening Conditions," *J. Acoust. Soc. Am.,* vol. 82, pp. 1637–1645 (1987 Nov.).

[30] R. S. Pellegrini, "Comparison of Data- and Model-Based Simulation Algorithms for Auditory Virtual Environments," presented at the 106th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 47, p. 530 (1999 June), preprint 4953.

[31] D. R. Begault, A. S. Lee, E. M. Wenzel, and M. R. Anderson, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," presented at the *108th Convention* of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 48, p. 359 (2000 Apr.), preprint 5134.

[32] U. Horbach, A. Karamustafaoglu, R. Pellegrini, P. Mackensen, and G. Theile, "Design and Application of a Data-Based Auralization System for Surround Sound," presented at the 106th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 47, p. 528 (1999 June), preprint 4976.

[33] T. Lokki and H. Järveläinen, "Subjective Evaluation of Auralization of Physics-Based Room Acoustics Modeling," in *Proc. Int. Conf. on Auditory Display* (Espoo, Finland, 2001 July), pp. 26–31.

[34] J. W. Philbeck and D. H. Mershon, "Knowledge about Typical Source Output Influences Perceived Auditory Distance," *J. Acoust. Soc. Am.,* vol. 111, pp. 1980–1983 (2000 May).

[35] F. L. Wightman, "Headphone Simulation of Free-Field Listening I: Stimulus Synthesis," *J. Acoust. Soc. Am.,* vol. 85, pp. 858–867 (1989).

[36] M. Kleiner, B. I. Dalenbäck, and P. Svensson, "Auralization—An Overview," *J. Audio Eng. Soc.,* vol. 41, pp. 861–875 (1993 Nov.).

[37] J. Huopaniemi, "Virtual Acoustics and 3-D Sound in Multimedia Signal Processing," Ph.D. thesis, Rep. 53, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, Espoo, Finland (1999).

[38] L. Savioja, "Modeling Techniques for Virtual Acoustics," Ph.D. thesis, Rep. TML-A3, Helsinki University of Technology, Telecommunications Software and Multimedia Laboratory, Espoo, Finland (1999).

[39] T. Lokki, "Physically Based Auralization—Design, Implementation, and Evaluation," Ph.D. thesis, Rep. TML-A5, Helsinki University of Technology, Telecommunications Software and Multimedia Laboratory, Espoo, Finland (2002).

[40] H. Møller, "Fundamentals of Binaural Technology," *Appl. Acoust.,* vol. 36, pp. 171–218 (1992).

[41] S. Yano, H. Hokari, and S. Shimada, "A Study on Personal Difference in the Transfer Functions of Sound Localization Using Stereo Earphones," *IEICE Trans. Fundamentals,* vol. E83-A, pp. 877–887 (2000 May).

[42] H. Mano, T. Nakamura, and W. L. Martens, "Perceptual Evaluation of an Earphone Correction Filter for Spatial Sound Reproduction," in *Proc. 5th Int. Conf. on Humans and Computers (HC2002),* (Aizu-Wakamatsu, Japan, 2002 Sept.).

[43] A. Kulkarni and H. S. Colburn, "Variability in the Characterization of the Headphone Transfer Function," *J. Acoust. Soc. Am.,* pp. 1071–1074 (2000 Feb.).

[44] B. G. Shinn-Cunningham, N. L. Durlach, and R. M. Held, "Adapting to Supernormal Auditory Localization Cues. II Constraints on Adaptation of Mean Response," *J. Acoust. Soc. Am.,* vol. 103, pp. 3667–3676 (1998 June).

[45] N. Gupta, A. Barreto, and C. Ordonez, "Spectral Modification of Head-Related Transfer Functions for Improved Virtual Sound Spatialization," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. II (Orlando, FL, 2002 May), pp. 1953–1956.

[46] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models* (Springer, New York, 1990).

[47] B. C. J. Moore, *Introduction to the Psychology of Hearing,* 4th ed. (Academic Press, New York, 1997).

[48] P. Rao, R. van Dinther, R. Vedhuis, and A. Kohlrausch, "A Measure for Predicting Audibility Discrimination Thresholds for Speech Envelope Distortions in Vowel Sounds," *J. Acoust. Soc. Am.,* vol. 109, pp. 2085–2097 (2001 May).

[49] P. M. Zurek, "Measurement of Binaural Echo Suppression," *J. Acoust. Soc. Am.,* vol. 66, pp. 1750–1757 (1979 Dec.).

[50] N. Sawhney and C. Schmandt, "Design of Spatialized Audio in Nomadic Environments," in *Proc. Int. Conf. on Auditory Display* (1997 Nov.).

[51] A. Walker, S. A. Webster, D. McGookin, and A. Ng, "A Diary in the Sky: A Spatial Audio Display for a Mobile Calendar," in *Proc. BCS IHM-HCI 2001* (Lille, France, 2001), pp. 531–540.

[52] B. MacIntyre and E. D. Mynatt, "Augmenting Intelligent Environments: Augmented Reality as and Interface to Intelligent Environments," *AAAI 1998 Spring Symp. Ser., Intelligent Environments Symp.* (Stanford University, Stanford, CA, 1998 Mar.).

[53] E. D. Mynatt, M. Back, R. Want, and R. Frederick, "Audio Aura: Light-Weight Audio Augmented Reality," in *Proc. 10th ACM Symp. on User Interface Software and Technology* (Banff, Alta., Canada, 1997 Oct.).

[54] J. Donath, K. Karahalios, and F. Viégas, "Visiphone," in *Proc. Int. Conf. on Auditory Display* (Atlanta, GA, 2000 Apr.).

[55] T. M. Drewes, E. D. Mynatt, and M. Gandy, "Sleuth: An Audio Experience," in *Proc. Int. Conf. on Auditory Display* (Atlanta, GA, 2000 Apr.).

[56] K. Lyons, M. Gandy, and T. Starner, "Guided by Voices: An Audio Augmented Reality System," in *Proc. Int. Conf. on Auditory Display* (Atlanta, GA, 2000 Apr.).

[57] H. Fletcher, "An Acoustic Illusion Telephonically Achieved," *Bell Labs. Rec.,* vol. 11, pp. 286–289 (1993 June).

[58] N. Koizumi, M. Cohen, and S. Aoki, "Design of Virtual Conferencing Environment in Audio Telecommunication," presented at the 92nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 40, p. 446 (1992 May), preprint 3304.

[59] D. R. Begault, "Virtual Acoustic Displays for Teleconferencing: Intelligibility Advantage for 'Telephone Grade' Audio," presented at the 98th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 43, p. 401 (1995 May), preprint 4008.

[60] S. H. Kang and S. H. Kim, "Realistic Audio Teleconferencing Using Binaural and Auralization Techniques," *ETRI J.,* vol. 18, pp. 41–51 (1996 Apr.).

[61] K. S. Phua and W. S. Gan, "Spatial Speech Coding for Multiteleconferencing," in *Proc. IEEE TENCON 1999,* pp. 313–316.

[62] J. Virolainen, "Design and Implementation of a Stereo Audio Conferencing System," Master's thesis, Helsinki University of Technology (2001).

[63] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and Two Ears," *J. Acoust. Soc. Am.,* vol. 25, pp. 975–979 (1953 Sept.).

[64] B. Bederson, "Audio Augmented Reality: A Prototype Automated Tour Guide," in *Human Computer in Computing Systems* (ACM, 1995), pp. 210–211.

[65] G. Eckel, "Immersive Audio-Augmented Environments: The LISTEN Project," in *Proc. 5th Int. Conf. on Information Visualisation,* (IEEE, 2001), pp. 571–573.

[66] R. Hull, P. Neaves, and J. Bedford-Roberts, "Towards Situated Computing," in *Proc. 1st Int. Symp. on Wearable Computers* (Cambridge, MA, 1997 Oct.), pp. 146–153.

[67] N. Belkhamza, A. Chekima, R. Nagarajan, F. Wong, and S. Yaacob, "A Stereo Auditory Display for Visually Impaired," in *Proc. IEEE TENCON 2000,* vol. II, pp. 377–382.

[68] D. P. Inman, K. Loge, and A. Cram, "Teaching Orientation and Mobility Skills to Blind Children Using Computer Generated 3-D Sound Environments," in *Proc. Int. Conf. on Auditory Display* (Atlanta, GA, 2000 Apr.).

[69] D. R. Begault and M. T. Pittman, "Three-Dimensional Audio versus Head-Down Traffic Alert and Collision Avoidance System Displays," *J. Aviation Psychol.,* vol. 6, pp. 79–93 (1996 Feb.).

[70] F. Mueller and M. Karau, "Transparent Hearing," in *Proc. of CHI2002* (Minneapolis, MN, 2002 Apr. 20–25), pp. 730–731.

[71] K. K. J. Rozier and J. Donath, "Hear & There: An Augmented Reality System of Linked Audio," in *Proc. Int. Conf. on Auditory Display* (Atlanta, GA, 2000 Apr.).

[72] M. Vorländer, "Acoustic Load on the Ear Caused by Headphones," *J. Acoust. Soc. Am.,* vol. 107, pp. 2082–2088 (2000 Apr.).

[73] A. Järvinen, L. Savioja, H. Möller, A. Ruusuvuori, and V. Ikonen, "Design of a Reference Listening Room—A Case Study," presented at the 103rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 45, p. 1015 (1997 Nov.), preprint 4559.

[74] J. Hynninen and N. Zacharov, "GuineaPig—A Generic Subjective Test System for Multichannel Audio," presented at the 106th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 47, p. 513 (1999 June), preprint 4871.

[75] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating Interactive Virtual Acoustic Environments," *J. Audio Eng. Soc.,* vol. 47, pp. 675–705 (1999 Sept.).

[76] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.,* vol. 65, pp. 943–950 (1979).

[77] R. Väänänen, V. Välimäki, J. Huopaniemi, and M. Karjalainen, "Efficient and Parametric Reverberator for Room Acoustics Modeling," in *Proc. Int. Computer Music Conf. (ICMC'97)* (Thessaloniki, Greece, 1997 Sept.), pp. 200–203.

[78] R. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson Jr., "Measurement of Correlation Coefficients in Reverberant Sound Fields," *J. Acoust. Soc. Am.,* vol. 27, pp. 1072–1077 (1955 Nov.).

[79] I. Chun, B. Rafaely, and P. Joseph, "Experimental Investigation of Spatial Correlation in Broadband Reverberant Sound Fields," *J. Acoust. Soc. Am.,* vol. 113, pp. 1995–1998 (2003 Apr.).

[80] I. M. Lindevald and A. H. Benade, "Two-Ear Correlation in the Statistical Sound Fields of Rooms," *J. Acoust. Soc. Am.,* vol. 80, pp. 661–664 (1986 Aug.).

[81] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust Localization in Reverberant Rooms," in *Microphone Arrays: Signal Processing Techniques and Applications,* M. S. Brandstein and D. Ward, Eds. (Springer, New York, 2001), ch. 7, pp. 131–154.

[82] G. L. Reid and E. Milios, "Active Stereo Sound Localization," *J. Acoust. Soc. Am.,* vol. 113, pp. 185–193 (2002 Jan.).

[83] C. Schauer and H. M. Gross, "Model and Application of a Binaural Sound Localization System," in *Proc. Int. Joint Conf. on Neural Networks,* vol. 2 (Washington, DC, 2001 July), pp. 1132–1137.

[84] N. Roman and D. Wang, "Binaural Tracking of Multiple Moving Sources," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (Hong Kong, 2003 May).

[85] M. Tikander, A. Härmä, and M. Karjalainen, "Binaural Positioning System for Wearable Augmented Reality Audio," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)* (New Paltz, NY, 2003 Oct.).

[86] T. Lokki, H. Nironen, S. Vesa, L. Savioja, and A. Härmä, "Problem of Far-End User's Voice in Binaural Telephony," in *Proc. 18th Int. Cong. on Acoustics (ICA'2004)* (Kyoto, Japan, 2004 Apr.).

[87] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Mag.,* vol. 5, pp. 4–24 (1988 Apr.).

[88] L. J. Griffiths and C. W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beamforming," *IEEE Trans. Antennas and Propag.,* vol. AP-30, pp. 27–34 (1982 Jan.).

[89] M. W. Hoffman, Z. Li, and D. Khataniar, "Gsc-Based Spatial Voice Activity Detection for Enhanced Speech Coding in the Presence of Competing Speech," *IEEE Trans. Speech and Audio Process.,* vol. 9, pp. 175–178 (2001 Feb.).

[90] D. K. A. Wang and K. Yao, "Comparison of Microphone Array Designs for Hearing Aid," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 4 (Detroit, MI, 1995 May), pp. 2739–2742.

[91] J. G. Desloge, W. M. Rabinowitz, and P. M. Zurek, "Microphone-Array Hearing Aids with Binaural Output—Part I: Fixed-Processing Systems," *IEEE Trans. Speech and Audio Process.,* vol. 5, pp. 529–551 (1997 Nov.).

[92] J. P. Woodard, "Modeling and Classification of Natural Sounds by Product Code Hidden Markov Models," *IEEE Trans. Signal Process.,* vol. 40, pp. 1833–1835 (1992 July).

[93] M. Casey, "MPEG-7 Sound-Recognition Tools," *IEEE Trans. Circ. Sys. for Video Technol.,* vol. 11, pp. 737–747 (2001 June).

[94] A. Härmä, "Automatic Recognition of Bird Species

Based on Sinusoidal Modeling of Syllables," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003),* vol. V (Hong Kong, 2003 Apr.), pp. 535–538.

[95] A. Eronen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-Based Context Awareness—Acoustic Modeling and Perceptual Evaluation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. V (Hong Kong, 2003 Apr.), pp. 529–532.

[96] A. Harter and A. Hopper, "A Distributed Location System for the Active Office," *IEEE Network,* vol. 8, (1994 Jan.).

[97] A. J. A. Ward and A. Hopper, "A New Location Technique for the Active Office," *IEEE Personal Commun.,* vol. 4, pp. 42–47 (1997 Oct.).

[98] A. C. N. B. Priyantha and H. Balakrishnan, "The Cricket Location-Support System," in *Proc. 6th ACM MOBICOM* (Boston, MA, 2000 Aug.).

[99] P. Bahl and V. N. Padmanabhan, "Radar: An In-Building RF-Based User Location and Tracking System," in *Proc. IEEE Infocom 2000* (Tel-Aviv, Israel, 2000 Mar.).

[100] J. Hightower, R. Want, and G. Borriello, "Spoton: An Indoor 3D Location Sensing Technology Based on RF Signal Strength," UW ESE Tech. Rep. 00-02-02, University of Washington, Department of Computer Science and Engineering, Seattle (2000 Feb.).

[101] L. M. Ni, L. Yunhao, C. L. Yiu, and A. P. Patil, "LANDMARC: Indoor Location Sensing Using Active RFID," in *Proc. 1st IEEE Int. Conf. on Pervasive Computing and Communications. (PerCom 2003)* (2003 Mar.), pp. 407–415.

[102] J. Warrior, E. McHenry, and K. McGee, "They Know Where You Are," *IEEE Spectrum,* vol. 40, pp. 20–25 (2003 July).

[103] M. Tikander, "Internet Link List on Localization and Headset Technology for MARA," www.acoustics.hut.fi/links/ (2003).

[104] K. S. J. Nord and P. Parnes, "An Architecture of Location Aware Applications," in *Proc. Hawaii Int. Conf. on System Sciences* (Big Island, Hawaii, 2002 Jan. 7–10).

[105] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, "Earcons and Icons: Their Structure and Common Design Principles," *Human–Computer Interaction,* vol. 4, no. 1, pp. 11–44 (1989).

[106] S. Helle, G. Leplâtre, J. Marila, and P. Laine, "Menu Sonification in a Mobile Phone—A Prototype Study," in *Proc. Int. Conf. on Auditory Display* (Espoo, Finland, 2001 July), pp. 255–260.

[107] M. Cohen and L. F. Ludwig, "Multidimensional Audio Window Management," *Int. J. Man–Machine Studies,* vol. 34, pp. 319–336 (1991).

[108] W. L. Martens, "Psychophysical Calibration for Controlling the Range of a Virtual Sound Source: Multidimensional Complexity in Spatial Auditory Display," in *Proc. Int. Conf. on Auditory Display* (Espoo, Finland, 2001 July), pp. 197–207.

[109] D. S. Brungart, "Near-Field Virtual Audio Displays," *Presence: Teleoperators & Virtual Environ.,* vol. 11, pp. 93–106 (2002 Feb.).

[110] K. Crispien and H. Petrie, "Providing Access to Graphical-Based User Interfaces (GUIs) for Blind People: Using a Multimedia System Based on Spatial Audio Representation," presented at the 95th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 41, pp. 1060, 1061 (1993 Dec.), preprint 3738.

[111] E. Mynatt and W. K. Edwards, "Mapping GUIs to Auditory Interfaces," in *Proc. ACM Symp. on User Interface Software and Technology (UIST'92)* (Monterey, CA, 1992 Nov.), pp. 61–70.

[112] E. D. Mynatt and G. Weber, "Nonvisual Presentation of Graphical User Interfaces: Contrasting Two Approaches," in *Proc. 1994 ACM Conf. on Human Factors in Computing Systems (CHI'94)* (Boston, MA, 1994 Apr.), pp. 166–172.

[113] A. Savadis, C. Stephanidis, A. Korte, K. Crispien, and K. Fellbaum, "A Generic Direct-Manipulation 3D-Auditory Environment for Hierarchical Navigation in Non-Visual Interaction," in *Proc. Assets'96* (Vancouver, B.C., Canada, 1996 Apr.), pp. 117–123.

[114] G. Lorho, J. Hiipakka, and J. Marila, "Structured Menu Presentation Using Spatial Sound Separation," in *Proc. Mobile HCI 2002* (Pisa, Italy, 2002 Sept.), pp. 419–424.

[115] J. Hiipakka and G. Lorho, "A Spatial Audio User Interface for Generating Music Playlists," in *Proc. Int. Conf. on Auditory Display* (Boston, MA, 2003 July).

[116] O. Kirkeby, "A Balanced Stereo Widening Network for Headphones," in *Proc. AES 22nd Int. Conf. on Virtual, Synthetic and Entertainment Audio* (Espoo, Finland, 2002 June), pp. 117–120.

THE AUTHORS

A. Härmä

J. Jakka

M. Tikander

M. Karjalainen

T. Lokki

J. Hiipakka

G. Lorho

Aki Härmä was born in Oulu, Finland, in 1969. He received master's and doctor's degrees in electrical engineering from the Helsinki University of Technology, Espoo, Finland, in 1997 and 2001, respectively.

In 2000–2001 he was a consultant at Lucent Bell Laboratories and the Media Signal Processing Research Department of Agere Systems, Murray Hill, NJ. In 2001 he returned to the Laboratory of Acoustics and Audio Signal Processing at the Helsinki University of Technology. In 2004 May he joined the Digital Signal Processing Group of Philips Research, Eindhoven, The Netherlands. His research interests are mainly in audio coding, spatial sound, and acoustic signal processing.

●

Julia Jakka was born in Vihti, Finland, in 1978. She is studying for a master's degree in electrical engineering and telecommunications at Helsinki University of Technology, Espoo, Finland, while working as a research assistant at the Laboratory of Acoustics and Audio Signal Processing.

●

Miikka Tikander was born in Riihimäki, Finland, in 1973. He graduated in 2003 from the Helsinki University of Technology, Espoo, Finland, with a M.Sc. degree in electrical and communication engineering, his major subject being acoustics and audio signal processing.

Since 2001 he has been a research assistant and, after graduation, a researcher in the Laboratory of Acoustics and Audio Signal Processing at the Helsinki University of Technology. Currently he is working on a Ph.D. degree. His research topics include signal processing and transducer technologies related to augmented reality audio.

Since 2001 Dr. Tikander has been acting as secretary of the Acoustical Society of Finland.

●

Matti Karjalainen was born in Hankasalmi, Finland, in 1946. He received M.Sc. and Dr. Tech. degrees in electrical engineering from the Tampere University of Technology in 1970 and 1978, respectively. His doctoral thesis dealt with speech synthesis by rule in Finnish.

In 1980 he joined the Helsinki University of Technology as an associate professor and since 1986 he has been a full professor in acoustics and audio signal processing. His interest is in audio signal processing, such as DSP for sound reproduction, perceptually based signal processing, as well as music DSP and sound synthesis. His research activities furthermore cover speech synthesis, analysis, and recognition, perceptual auditory modeling, spatial hearing, DSP hardware, software, and programming environments, as well as various branches of acoustics, including musical acoustics and modeling of musical instruments. He has written 330 scientific and engineering articles and was involved in the organization of several professional conferences and workshops. He was the papers chair of the AES 16th International Conference on Spatial Sound Reproduction.

Dr. Karjalainen is an AES fellow and a member of the Institute of Electrical and Electronics Engineers, the Acoustical Society of America, the European Acoustics Association, the International Computer Music Association, the European Speech Communication Association, and several Finnish scientific and engineering societies.

●

Tapio Lokki was born in Helsinki, Finland, in 1971. He studied acoustics, audio signal processing, and computer science at Helsinki University of Technology and received an M.Sc. degree in electrical engineering in 1997 and a D.Sc. (Tech.) degree in computer science and engineering in 2002.

At present he is a teaching researcher with the Telecommunications Software and Multimedia Laboratory at the Helsinki University of Technology. His research activities include 3-D sound, room acoustics, virtual acoustic environments, auralization, and virtual reality.

Dr. Lokki is a member of the Audio Engineering Society, the IEEE Computer Society, and the Acoustical Society of Finland.

●

Jarmo Hiipakka was born in Kokkola, Finland, in 1973. He studied information technology as well as acoustics and audio signal processing at the Helsinki University of Technology, Espoo, Finland, and received an M.Sc. degree in 1999.

From 1996 to 2000 Mr. Hiipakka worked as a research scientist at the Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, and Telecommunications Software and Multimedia Laboratory. Currently he is with the Nokia Research Center, Audio-Visual Systems Laboratory, Helsinki. His research interests include interactive and 3-D audio applications, auditory user interfaces, and audio for virtual and augmented reality systems.

●

Gaëtan Lorho was born in Vannes, France, in 1972. He received a master's degree in fundamental physics from the University of Paris VII, France, in 1996, and a master's degree in sound and vibration studies from the Institute of Sound and Vibration Research at the University of Southampton, UK, in 1998.

Since 1999 he has been working as a research engineer at the Nokia Research Center in Helsinki, Finland. His main research interests are in subjective evaluation of audio quality, spatial sound reproduction, and audio user interfaces.