Robert Albrecht

Messaging in Mobile Augmented Reality Audio

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 31.1.2011

Thesis supervisor and instructor:

Adjunct professor Tapio Lokki, D.Sc. (Tech.)

Aalto University
School of Electrical Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Robert Albrecht		
Title: Messaging in Mobile Augmented Reality Audio		
Date: 31.1.2011	Language: English	Number of pages: 7+53
Department of Signal Processing and Acoustics		
Professorship: Acoustics and Audio Signal Processing		Code: S-89
Supervisor and instructor: Adjunct professor Tapio Lokki, D.Sc. (Tech.)		

Asynchronous multi-user communication is typically done using text. In the context of mobile use text input can, however, be slow and cumbersome, and attention on the display of the device is required both when writing and reading messages. A messaging application was developed to test the concept of sharing short messages between members of groups using recorded speech rather than text. These messages can be listened to as they arrive, or browsed through and listened to later. The application is intended to be used on a mobile augmented reality audio platform, allowing almost undisturbed perception of and interaction with the surrounding environment while communicating using audio messages.

A small group of users tested the application on desktop and laptop computers. The users found one of the biggest advantages over text-based communication to be the additional information associated with a spoken message, being much more expressive than the same written message. Compared with text chats, the users thought it was difficult to quickly browse through old messages and confusing to participate in several discussions at the same time.

Keywords: Augmented reality audio, messaging, binaural recording,
social media

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Robert Albrecht

Työn nimi: Viestintä lisätyssä äänitodellisuudessa

Päivämäärä: 31.1.2011	Kieli: Englanti	Sivumäärä: 7+53

Signaalinkäsittelyn ja akustiikan laitos

Professuuri: Akustiikka ja äänenkäsittely	Koodi: S-89

Valvoja ja ohjaaja: Dosentti Tapio Lokki

Monen käyttäjän välinen asynkroninen viestintä tapahtuu tyypillisesti tekstiä käyttäen. Mobiileissa käyttötilanteissa tekstinsyöttö voi kuitenkin olla hidasta ja vaivalloista. Sekä viestien kirjoittaminen että lukeminen vaatii huomion keskittämistä laitteen näyttöön. Tässä työssä kehitettiin viestintäsovellus, jossa tekstin sijaan käytetään puhetta lyhyiden viestien jakamiseen ryhmien jäsenten välillä. Näitä viestejä voidaan kuunnella heti niiden saapuessa tai niitä voi selata ja kuunnella myöhemmin. Sovellusta on tarkoitettu käytettävän mobiilin lisätyn äänitodellisuuden alustan kanssa, mikä mahdollistaa lähes häiriintymättömän ympäristön havaitsemisen samalla kun kommunikoi ääniviestien avulla.

Pieni ryhmä käyttäjiä testasi sovellusta pöytätietokoneilla ja kannettavilla tietokoneilla. Yksi isoimmista eduista tekstipohjaiseen viestintään verrattuna todettiin olevan puheen mukana välittyvä ylimääräinen tieto verrattuna samaan kirjoitettuun viestiin, puheviestinnän ollessa paljon ilmekkäämpää. Huonoja puolia verrattuna tekstipohjaiseen viestintään olivat hankaluus selata vanhojen viestien läpi sekä vaikeus osallistua useampaan keskusteluun samaan aikaan.

Avainsanat: Lisätty äänitodellisuus, viestintä, binauraalinen nauhoitus,
sosiaalinen media

Acknowledgements

This Master's thesis has been written for the Department of Signal Processing and Acoustics at Aalto University. The work for the thesis was done at the Department of Media Technology and was funded by Nokia Research Center.

I want to thank Tapio Lokki for giving me the opportunity to work on this project and providing inspiration and help when needed. Thanks also go to all my co-workers at the department.

I would like to thank the people at Nokia Research Center, especially Sampo Vesa, who came with ideas and feedback throughout the project.

I would like to thank my family, flatmates and friends. Special thanks go to Saurabh, Riku, Janne, Viljami and Klaus, who helped me out in different ways during the process.

Otaniemi, 31.1.2011

Robert Albrecht

Contents

Symbols and abbreviations

Symbols

δ	elevation angle
φ	azimuthal angle
τ	time difference
c	speed of sound
D	diameter
f_n	nth resonance frequency
L	length
r	distance, radius

Abbreviations

API	Application Programming Interface
AR	Augmented Reality
ARA	Augmented Reality Audio
FFT	Fast Fourier Transform
GPS	Global Positioning System
HRIR	Head-Related Impulse Response
HRTF	Head-Related Transfer Function
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ILD	Interaural Level Difference
IM	Instant Messaging
IRC	Internet Relay Chat
ITD	Interaural Time Difference
KAMARA	Killer Applications of Mobile Augmented Reality Audio
KEMAR	Knowles Electronics Mannequin for Acoustics Research
MARA	Mobile Augmented Reality Audio
SMS	Short Message Service
SSH	Secure Shell
TRS	Tip, Ring, Sleeve
TTS	Text To Speech
USB	Universal Serial Bus
VR	Virtual Reality
XMPP	Extensible Messaging and Presence Protocol

1 Introduction

Augmented reality (AR) adds information to our perception of the surrounding environment, augmenting it with virtual objects. Most research on augmented reality focuses on augmenting our visual perception of the world. Augmented reality audio (ARA), on the other hand, augments our auditory perception of the surrounding environment with virtual sounds. The virtual sound sources can be placed in different directions or locations through the use of spatial sound rendering.

The work for this thesis has been done as part of the KAMARA 2010 project (Killer Applications of Mobile Augmented Reality Audio). The project is one in a series of projects done since 2001 as a collaboration between Nokia Research Center and the Department of Signal Processing and Acoustics and the Department of Media Technology at Aalto University. The research has focused on applications using a headset consisting of insert headphones and binaural microphones (the so called ARA headset) together with a wearable terminal unit with network connection and in some cases position-aware techniques. The microphones are used to make the headset acoustically transparent, by passing the microphone signal to the headphones. On the way, the signal is passed through the ARA mixer. The mixer performs equalization to account for the changes in the ear canal resonances when the ear canal is occluded by the headphone earpiece, and mixes the microphone signal with the audio with which the surrounding acoustic environment is augmented.

The traditional use of mobile phones is synchronous voice communication, with the addition of the Short Message Service (SMS) allowing asynchronous communication with brief text messages. Voicemail services add support for asynchronous voice communication. In asynchronous multi-user messaging applications, however, communication is normally done using text. Examples of these are social networking applications, such as Facebook [1], Myspace [2] and Twitter [3], which are available both for desktop and mobile platforms. Using the keypads, touch screens or small keyboards on mobile devices for text input can, however, be both cumbersome and slow.

1.1 Aim of the thesis

This thesis looks at the alternative of using audio rather than text input for asynchronous multi-user messaging. One of the main research questions is how such a messaging application can be implemented on a mobile augmented reality audio (MARA) platform and how the application can utilize binaural audio and spatial sound rendering. Another question is how users would react to and comment on this kind of application. To be able to gather answers to this question, a messaging application called Mobile Augmented Messaging is implemented as a proof of concept. Of particular interest is the comparison of asynchronous multi-user voice communication with asynchronous and synchronous text communication, and synchronous voice communication.

1.2 Organization of the thesis

In Section 2, the fundamentals of spatial hearing are presented to give background information on auditory localization cues and how these can be used to render spatial audio. In Section 3, the concept of augmented reality audio is presented in more detail, along with headphone acoustics. The ARA headset and mixer are also presented in this section. Section 4 discusses different types of messaging applications, giving examples of these and their usage in a mobile context. In Section 5, the concept of the Mobile Augmented Messaging application is presented together with the details of the implementation. Section 6 covers the evaluation of the application and discusses the results. Section 7 summarizes the work and discusses possibilities for future work on the subject.

2 Spatial hearing

Several different types of spatial information are normally associated with an auditory event. These include the perceived location of the sound source, which partly is based on the perceived direction of arrival and distance. The sound that arrives at the ears will also give the listener hints about other properties of the propagation path from the sound source to the ears. The listener may, e.g., be able to estimate the size and shape of a room, as well as wall materials.

Figure 1 shows a spherical coordinate system relative to the head of a listener. This coordinate system can be used to define the perceived as well as actual location of a sound source.

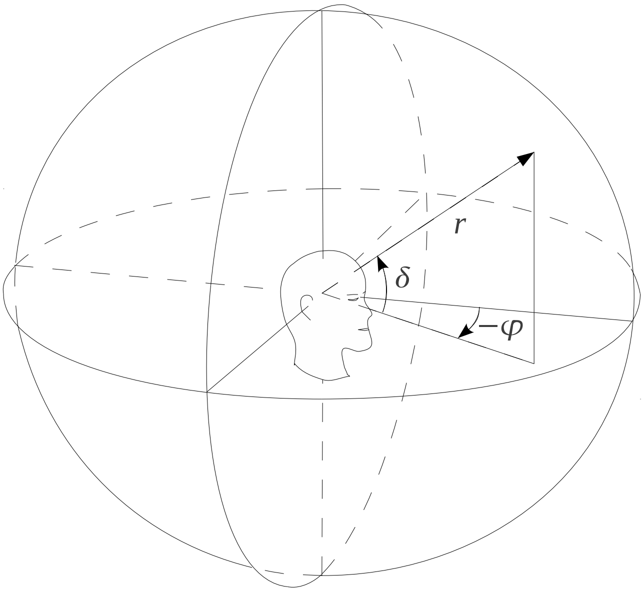


Figure 1: A spherical coordinate system relative to the head, for defining the direction (azimuth φ and elevation angle δ) and distance (r) of a sound source. After [4].

Sound source localization is based on different cues, provided either by the head and body of the listener or by the surroundings. For lateral localization, the two most important cues are the interaural time difference (ITD) and interaural level difference (ILD) cues [4]. Vertical localization is primarily based on spectral cues [5].

2.1 Interaural time difference

For a sound source direction with an azimuth φ other than 0° or 180°, one of the ears of the listener will be closer to the source than the other. The difference in distance can be approximated with the head being a sphere and sound diffracting along the surface of this sphere, as illustrated in Figure 2. The corresponding time difference will then be

$$\tau = \frac{D}{2c}(\varphi + \sin\varphi), \tag{1}$$

where D is the diameter of the head, c is the speed of sound and φ is the azimuth of the sound source in radians [4]. τ is called the interaural time difference (ITD). Here it is assumed that the distance from the sound source to the listener is large enough so that the sound wave can be approximated by a plane wave.

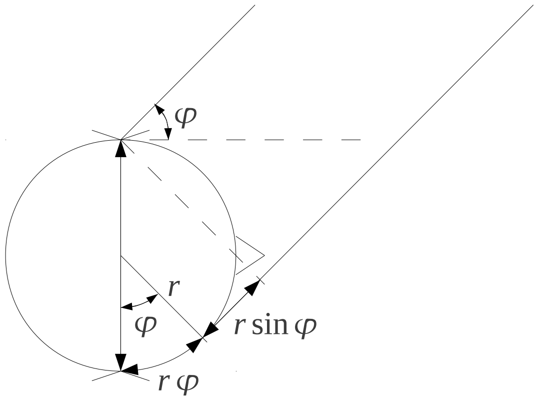


Figure 2: Illustration of the path difference from a sound source to the left and right ear. r is the radius of the head and φ is the azimuth of the sound source in radians.

When using Equation (1), the elevation angle δ is assumed to be 0°. If the angle increases (to the positive or negative), the ITD will be smaller than the one given by this formula. When the sound source is located straight above the head ($\delta = 90°$), the time difference will be zero. To calculate the interaural time difference for elevation angles δ other than 0°, the factor $\cos\delta$ can be added to Equation (1) [6].

For clicks or other sounds with a rapidly varying time envelope, the interaural time difference itself is used for lateral localization, but for steady sounds, the associated phase difference is used. When the frequency of a steady sound grows high enough, this phase difference becomes ambiguous. Experiments have shown that the ITD is used as a lateral localization cue for frequencies up to about 1000 Hz [7].

2.2 Interaural level difference

With one ear turned towards the sound source, and the other away from it (i.e. φ is not 0° or 180°), the head will be shadowing the one ear more than the other, causing a difference in sound pressure level between the ears. This interaural level difference (ILD) is smaller for lower-frequency sounds, as these diffract around the head, and is roughly proportional to the sine of the azimuthal angle [5]. This can be seen in Figure 3, which shows interaural level differences for sound sources at different azimuthal angles in the horizontal plane. The ILD is used as a lateral localization cue for frequencies of about 4000 Hz and above [7].

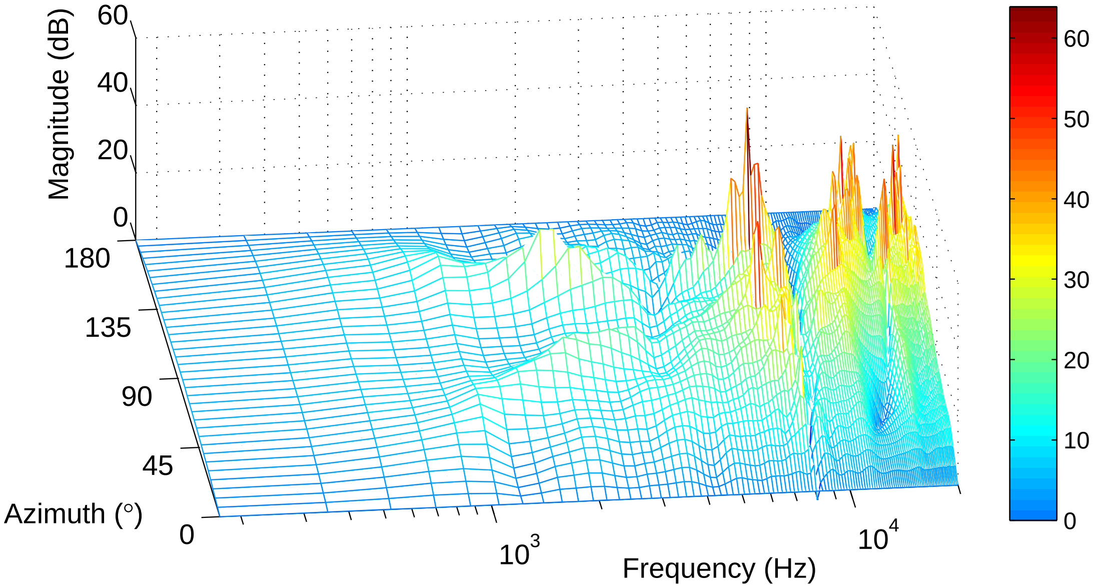


Figure 3: Interaural level differences for different azimuths in the horizontal plane. Based on data from [8].

2.3 Other lateral localization cues

As mentioned above, the use of interaural time differences is limited to frequencies below 1000 Hz and the use of interaural level differences to frequencies above 4000 Hz. The accuracy of lateral localization declines between 1000 and 4000 Hz [7].

The interaural time and level differences alone do not provide enough information to unambiguously determine the direction of a sound. Assuming a spherical head model, these cues suggest directions lying on the surface of a cone, the so called "cone of confusion" (Figure 4). All sound arriving along the surface of this cone produces the same time and level difference cues. Assuming the sound arrives from somewhere in the horizontal plane ($\delta = 0°$), both cues suggests two possible values for the azimuth. The forward-facing shape of the pinna aids in resolving this so called front-back confusion, by reducing the high-frequency magnitude response of sounds arriving from the back compared with those arriving from the front [9]. The spectrum-shaping function of the pinna, also used for vertical localization, is studied further in Section 2.4.

Another cue that can help in sound source localization is provided by head movement. With the head held still, interaural time and level differences place the sound source on a cone of confusion. When the head is rotated, these cues will, depending on source location and the type of rotation, place the source on a different cone of confusion. The source is thus located on the line where these two cones meet. Normally, head movement seems to provide little improvement to localization accuracy. However, head movement helps in disambiguating potential confusions, when other cues fail to provide enough information [5].

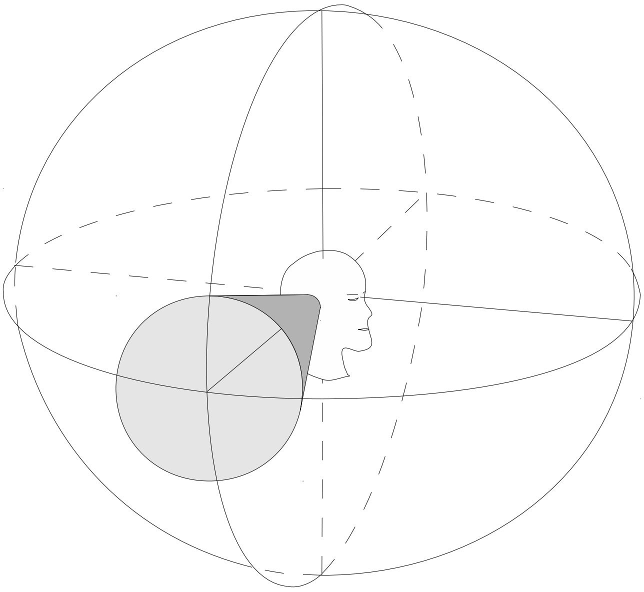


Figure 4: The cone of confusion. The interaural time and level difference cues alone are not enough to unambiguously determine the direction of a sound source. These cues only suggest that the sound source lies somewhere on the surface of a cone.

That visual cues affect the localization of sound is something familiar to most of us. When watching a television screen, a voice heard from the loudspeakers will be localized where the supposed speaker is seen on the screen, even if localization based on sound alone would place the sound source at a noticeable distance from this location. Whether visual or auditory cues dominate when these contradict each other depends not only on the distance between the locations these suggest, but also, for instance, on the context [10]. Without cues provided by head movement, the lack of visual information accompanying an auditory event will likely result in localization behind the head, even if acoustical cues suggest a location in front of the head [9]. This is due to the assumption that what cannot be seen is likely to be behind the head.

2.4 Vertical localization

For sound originating from anywhere in the median plane, where the distance to both ears is equal, there will be no interaural time or level differences to provide cues for localizing the exact position of the sound source in this plane (assuming that the head is symmetrical). Neither do interaural differences provide information about the exact location of a sound source on a cone of confusion (of which the median plane is a special case). The cues used for vertical localization must thus be of a different nature.

The major cues used for vertical localization seem to be so called spectral cues [5]. These are mainly caused by the geometry of the pinna, which shapes the spectrum of sound based on the direction of arrival. This shaping seems to be caused by multiple mechanisms. When the direct sound sums with sound reflected from the walls of the pinna cavities, the resulting signal will have peaks and notches corresponding to the delay between the summed signals. The pinna also has several resonant modes that provide spectral cues. These have resonant frequencies and Q factors independent of the direction of the incident wave, but the magnitude of excitation of these resonances depends strongly on the direction [11].

For accurate vertical localization, a stimulus with broad bandwidth and energy at high frequencies is required. Knowledge about the source spectrum is also of importance. If the spectrum of the source is not familiar to the listener, he or she will be unable to tell if a particular spectral feature is produced by interaction with the pinna or if it is present in the source spectrum. However, movement of the source will result in changing spectral characteristics, which the brain can separate from the constant characteristics of the source spectrum. It is likely that this will make spectral cues more easily detectable [9].

2.5 Localization blur

Localization blur is the smallest change in some attribute or attributes of a sound event that results in a change in localization of the associated auditory event. Alternatively, the underlying change may not be in an attribute of the sound event itself, but, for instance, in the position of the listener. For horizontal angular displacement of a sound source located in front of the listener ($\varphi = 0°$ and $\delta = 0°$), localization blur has a lower limit of about 1°, slightly depending on the type of signal [10]. Localization blur increases as the azimuth approaches 90° or $-90°$. This is consistent with the fact that changes in the interaural level and time differences are largest when the azimuth is close to 0° or 180°, and smallest when the azimuth is close to 90° or $-90°$(assuming a spherical head model).

For vertical angular displacement of a sound source in the median plane, localization blur has a strong dependence on the signal. In front of the listener, localization blur is about 17° for continuous speech by an unfamiliar speaker and about 9° for a familiar speaker [10]. For white noise, the localization blur is approximately 4°. Among five tested directions ($\varphi = 0°, \delta = 0°; \varphi = 0°, \delta = 36°; \delta = 90°; \varphi = 180°, \delta = 36°; \varphi = 180°, \delta = 0°$) for continuous speech by a familiar speaker, the maximum localization blur of 22° was found at the source direction with azimuth $\varphi = 180°$ and elevation angle $\delta = 36°$.

2.6 Perception of distance

Perception of distance of a sound source in the far field of the listener is primarily based on two factors: sound pressure level and echoes plus reverberation. In anechoic conditions, reliable estimation of distance is an almost impossible task [4]. In such conditions, estimation is based on the sound level, which may or may not correlate with the actual distance. In the case of a familiar signal source, such as a person talking, the distance can be estimated with some accuracy.

If echoes and reverberation are present, they provide the listener with information about the size of the space and the distance to nearby surfaces. As the distance to the sound source in a reverberant environment increases, the ratio of direct to reverberant sound decreases. Combined with the listener's experience of different sound environments, these hints together help the listener make an estimation of the distance to the sound source. Changes to the frequency spectrum caused by air absorption or reflections also provide hints about the source location, but require knowledge of the source spectrum [12].

For sound originating from sources in the near field of the listener, additional distance cues are available. For sources near the interaural axis (the azimuth φ is close to 90° or $-90°$), the ILD can increase up to 20–30 dB when the source distance is reduced from 1 to 0.12 m [13]. This increase is due to an increased head-shadowing effect for sources near the head, as well as an increased difference in distance attenuation [14]. Unlike head shadowing, which increases the high-frequency ILD, distance attenuation affects the ILD across all frequencies, resulting in a near-field ILD that can exceed 15 dB at 500 Hz, compared with the maximum far-field ILD of 5–6 dB at the same frequency. As the azimuth moves closer to 0° or 180°, the increase in ILD with reduced source distance gets smaller. Unlike the ILD, the ITD increases only slightly with decreasing distance in the near field.

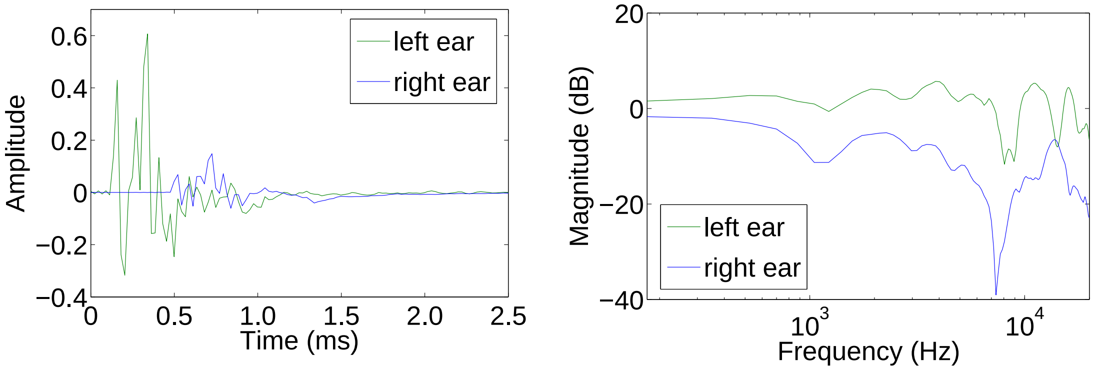
2.7 Head-related transfer functions

A head-related transfer function (HRTF) describes the transmission of sound from a sound source in a free field to a point at the entrance of or somewhere inside the ear canal. A HRTF thus describes the effects that the head (and possibly torso) have on sound transmission from the sound source to a listener. These effects can be determined by measuring the impulse response (called a head-related impulse response, HRIR) from a loudspeaker to a microphone in the ear canal of either a test subject or a dummy head in an anechoic chamber. A head-related transfer function is the Fourier transform of a head-related impulse response. HRTFs are not only useful when studying the different cues available for sound source localization, but a pair of HRTFs measured for both ears can also be applied to a mono recording to make it sound like it comes from a particular point in space.

Because the diameter of the ear canal is small in comparison with the wavelengths of sound, only a longitudinal plane wave can propagate in it [4]. This means that the transfer function from the entrance of the ear canal (and as far as 0.6 cm outside of it) to the eardrum is independent of the direction sound arrives from. The same localization cues will thus be present in HRTFs measured with microphones placed anywhere inside the ear canal.

A measured head-related transfer function does not only contain the desired response, but also the response of the measurement system, i.e. loudspeaker and microphone. By measuring the transfer function of the measurement system itself, an inverse filter can be created to remove the effects that the nonideal measurement system has on the measured data. Another way to remove these effects is to perform either free-field or diffuse-field equalization [15]. In free-field equalization, the measured HRTF is deconvolved by a reference HRTF measured in the same ear from a certain direction (typically $\varphi = 0°$ and $\delta = 0°$). Diffuse-field equalization is done by deconvolving the measured HRTF by a diffuse-field reference HRTF. This reference HRTF is calculated as an average HRTF over all measured directions (typically calculated as the root-mean-square value of the transfer function magnitude, i.e., the power average), estimating a diffuse sound field. Deconvolution by a free-field or diffuse-field reference HRTF will remove any factors that do not depend on the direction of the sound source. These include not only the effects of the measurement system, but also, e.g., the ear canal resonances, which are present in the response if the measurement microphone is placed inside the ear canal.

Figure 5(a) shows diffuse-field equalized HRIRs measured using a dummy head with a sound source azimuth of 45° and elevation angle of 0°. An interaural time difference of about 0.36 ms can be seen in the figure. Figure 5(b) shows the magnitude of the corresponding HRTFs. The figure displays an interaural level difference, which is smaller for low frequencies, as well as spectral cues present at higher frequencies. Figure 6 shows the magnitude of HRTFs from the same measurements for all azimuthal angles (with 5° spacing) in the horizontal plane.



(a) HRIRs (b) HRTFs

Figure 5: HRIRs and magnitude of HRTFs measured using a dummy head with a sound-source azimuth of 45° and elevation angle of 0°. Based on diffuse-field-equalized data from [8].

Diffuse-field equalized HRTFs can be used to produce synthetic binaural signals which have a timbre compatible with conventional microphone recordings and with conventional stereo reproduction on loudspeakers [16]. The signals are also compatible with reproduction on diffuse-field-calibrated headphones (see Section 3.3.6).

2.8 Auralization

Auralization can be defined as "*...the process of rendering audible, by physical or mathematical modeling, the soundfield of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space* [17]." The two main steps involved in auralization are thus the simulation of the

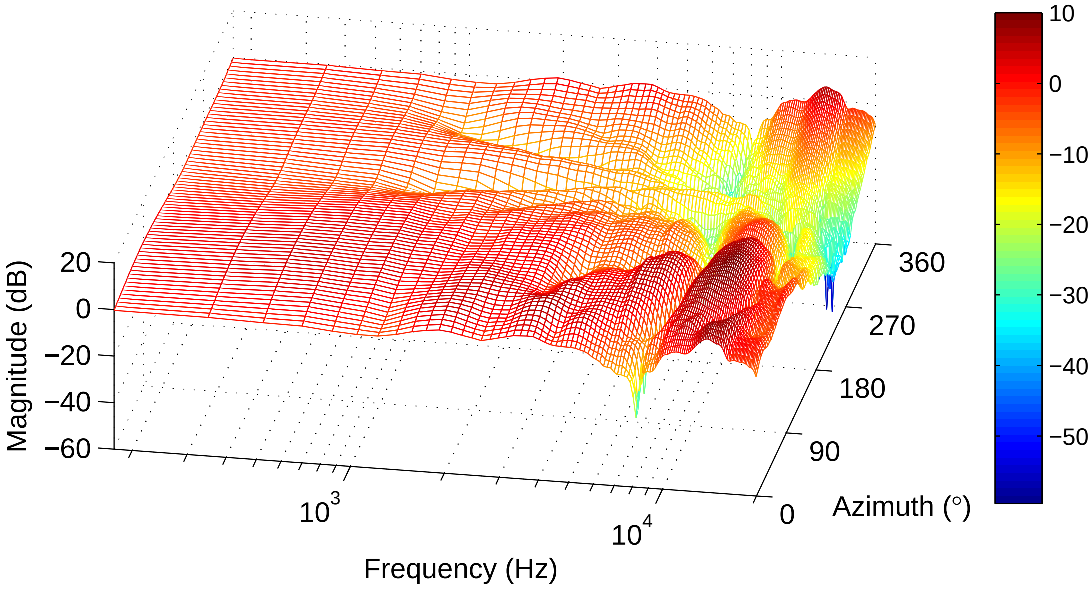


Figure 6: Magnitude of HRTFs measured using a dummy head with different azimuthal angles in the horizontal plane. Based on diffuse-field-equalized data from [8].

sound field in a space, and simulation of the effects of the listener's head (and possibly torso) on the sound depending on its direction of arrival. Simulation of the sound field might be as simple as adding a suitable amount of reverberation to achieve better externalization [18] (localization outside the head of the listener) and produce some sensation of space. On the other hand, it might involve detailed modelling of, e.g., a concert hall and the sound field produced by different sources placed in the hall.

To create a binaural listening experience, head-related transfer functions may be used to add a sense of direction to the sound. This can be done, e.g., by convoluting the audio data with the appropriate HRIRs for both ears. When the convoluted audio is played using headphones, the listener will hear the sound with spatial cues added, and it will thus be perceived as arriving from a certain direction.

If the same experience should be produced using loudspeakers, two problems arise. Firstly, assuming a setup with two loudspeakers, the output from the left loudspeaker will not only be heard by the left ear, but also by the right ear, and vice versa. This is referred to as crosstalk. Secondly, in addition to the spatial cues added artificially by using HRTFs, another set of spatial cues will be added to the sound that reaches the ear canals. These spatial cues will depend on the placement of the loudspeakers relative to the head of the listener, as well as the orientation of the listener's head. To avoid these problems, crosstalk cancellation using compensating filters should be applied to the signals sent to the loudspeakers.

When using loudspeakers for auralization, reflections and reverberation of the surrounding space will have a detrimental impact on both the quality of localization cues and perception of the simulated space. Using headphones, the sound reproduction chain is almost completely isolated from the effects of the surrounding space. However, using headphones the virtual sound field will follow the movements of the listener's head, having a detrimental impact on the perception of space. To avoid this, tracking of the position of the listener's head, and appropriate compensation according to this, should be performed.

Another important contribution from head tracking is improved localization. When moving virtual sound sources according to the movements of the listener's head, both in-the-head localization and front-back confusion are significantly reduced [19, 20]. This is especially beneficial when using generic HRTFs, where a large degree of front-back confusion normally is present.

Ideally, individual HRTFs should be used for auralization. HRTFs recorded with microphones placed in the listener's own ear canals provide the listener with the same spatial cues as he or she would naturally encounter. However, recording individual HRTFs is not very feasible when it comes to applications intended for a large crowd. In this case, generic HRTFs (measured either using a dummy head or a test subject) can be used instead. Alternatively, sets of HRTFs measured using different test subjects could be used, preferably accompanied by a convenient method for selecting the most suitable HRTF set (for an example of this, see [21]).

Horizontal localization using generic HRTFs is generally quite accurate, because of the robust interaural difference cues available [22]. However, because of the large variations in pinna geometry, spectral cues will rarely be a good match to those the listener is accustomed to, resulting in front-back and up-down confusion.

3 Augmented reality audio

The concept of augmented reality is different from that of virtual reality (VR), where a completely virtual environment, e.g., visual or auditory, is created, replacing the real environment. In augmented reality the sensations provided by the real environment are not replaced, but augmented with virtual objects. By one definition [23], an augmented reality system has the following three characteristics:

- Combines real and virtual
- Interactive in real time
- Registered in 3D

In augmented reality audio the natural auditory environment is augmented with virtual sound sources. The added virtual sounds in ARA applications can, if wanted, be made practically indistinguishable from similar real-world sounds, thus blending perfectly with the natural sound environment. On the other hand, the virtual sounds can deliberately be made to stand out from the natural environment, making it easy for the listener to distinguish between virtual and real sounds.

Direction or location is an important aspect in augmented reality audio applications. The virtual sounds are ideally rendered so that they are heard to arrive from a specific direction or location in the real world. This task might be quite difficult, especially when it comes to mobile augmented reality audio, requiring accurate tracking of head orientation and position, and adding the necessary localization cues to the sound. Spatial sound rendering can be used simply and intuitively to provide the listener with information about the location of a point of interest, but a specific direction might also convey other information associated with the sound being heard.

3.1 Mobile augmented reality audio

As with visually augmented reality, which can be realized using, for instance, see-through displays or video displays with virtual objects drawn on a camera image, augmented reality audio can be realized using different methods. Using loudspeakers, virtual sound sources can be added without disturbing or altering the listener's perception of the natural sound environment. However, this method has some major restrictions. Firstly, the ARA application is limited to the area where the loudspeakers are located, making this method unsuitable for mobile applications. Secondly, if each possible position for a virtual sound source is not represented by a separate loudspeaker, the area where the desired sound field can be rendered, referred to as the sweet spot, will be much smaller than the area where the loudspeakers are heard. This is the case both when a multi-loudspeaker surround system and a virtual surround system with two loudspeakers is used [24].

In mobile augmented reality audio applications, the sound reproduction system must be portable, allowing the listener to move freely. Different technologies can be used to achieve this. Conventional headphones cannot be used as such, because they occlude the ear canal opening, thus attenuating the sounds from the environment. Shoulder-mounted loudspeakers may be used, but these reduce the quality of spatial sound localization cues [25]. These cues are important if the added sounds should be perceived as arriving from a certain direction. Another problem with loudspeakers of this kind is that the sound will not only be heard by the wearer, but also by nearby people, possibly disturbing these.

An alternative to shoulder-mounted loudspeakers, which also leaves the ear-canal opening unoccluded, are bone-conduction headphones. The earpieces of the headphones are placed against the bone of the skull, transmitting sound via mechanical vibrations through the bone and tissue to the inner ear. One of the main problems with bone-conduction headphones is low interaural attenuation [26].

The problem with conventional headphones in ARA applications is that they, depending on their type, more or less seal off the natural sound environment. In augmented reality, the natural environment should ideally be perceived unaltered. One way to try to achieve this is to use microphones at each ear, which pick up the sound from the environment. This microphone signal is then mixed with an audio signal containing the desired virtual sounds, and passed to the ears through the headphones.

The microphones should be located close to the ear canal entrance or inside the ear canal, so that all the localization cues present in normal listening are picked up (see Section 2.7 for details). The transducers of the headphones should also be placed in such a way that the sound transmitted through these enters the ear canal with as little alteration of these localization cues as possible. This is important not only for transmitting sound from the environment, but also for the reproduction of virtual sounds to which localization cues have been added artificially.

Ideally, the natural sound environment reproduced through headphones would not differ from that heard without headphones, but in practice some differences will always be heard. To differentiate from the real acoustic environment, this representation of it is called the pseudoacoustic environment [27].

An advantage of using headphones that attenuate sounds from the environment together with microphones, is that the user can control the level of the pseudoacoustic sounds. When listening to music, for example, the pseudoacoustic environment can be mixed in at a low level, if at all.

3.2 Applications of augmented reality audio

Numerous different applications and application scenarios of augmented reality audio have been presented in previous studies [28, 29, 30, 31, 32]. One way to classify AR applications is based on how they mix the real and the virtual environment. One type of AR applications, which could be called pure AR applications, augments the real environment with virtual objects, creating a mixed virtual and real environment where the virtual objects presented depend on the real environment, and add information to the real environment. Another type of AR applications does not add information related to the real environment, but simply uses an AR platform to allow presentation of both the real environment and a virtual environment at the same time. One might argue that this kind of applications should not be called AR applications at all. This section presents some different ARA applications from both categories.

Speech communication

Speech communication will probably remain as one of the most important applications of mobile augmented reality [28]. In a communication situation between two persons, the voice of the remote talker could be placed in front of the local user and mixed with his or her pseudoacoustic environment. Alternatively, the local user could hear the binaural recording more or less exactly as it was made at the remote end. The user would thus be virtually transported to the location of the remote talker, a concept called telepresence. In telepresence, the recording at the remote end can also be done using a dummy head equipped with binaural microphones.

In a teleconference between multiple participants all at different remote locations, the voices of the other speakers can be placed in different directions around the listener. Speech intelligibility has been shown to improve when multiple simultaneous voices are spatially separated [33]. In another usage scenario, the local user wearing an ARA headset with binaural microphones participates in a meeting held between multiple participants all in the same room [32]. Among these participants is the remote user, also wearing an ARA headset. The local user will hear the conversation exactly as the remote user does. Problems arise when the remote user rotates his or her head, causing the voices the local user hears to rotate accordingly. The local user might also rotate his or her head, also causing the position of the voices to change. To avoid the possible deterioration of listening comfort and speaker separation associated with this, head tracking can be used to keep track of the head orientation of both users. Based on this information, the spatial localization cues present in the binaural audio can be altered, so that the voices are heard from the same positions, independent of any head movements of either user.

Museum guide

A good example of augmenting the real environment with audio is given by the virtual museum guide application implemented as part of the LISTEN project [34]. The museum visitor wears wireless headphones with motion tracking, only implicitly interacting with the museum guide system through movement and head orientation. Based on this information, a personalized audio presentation is created, presenting the exhibit the visitor is focusing his or her attention on, and making recommendations about other exhibits the visitor might be interested in. Visitor interests are modelled using meta-tags associated with the exhibits. As the visitor focuses his or her attention on a specific exhibit, the score values for all meta-tags associated with that exhibit are increased. The system recommends exhibits that have meta-tags which have received high scores, i.e. the user will probably find interesting.

Audio games

As with other augmented reality applications, augmented reality audio games consist of two different environments more or less mixed together. The first is the real-world environment, providing visual and auditory sensations as well as physical constraints. The second is the virtual audio environment, which provides the dynamic information needed for realizing the desired game play. An example of this is Eidola [35], an ARA game prototype. In this game, the player moves around in a room inhabited by invisible creatures. The player can defeat these creatures by locating them based on sounds they make, moving to their position, and pulling a virtual trigger. Movement of both the player and the creatures is restricted by different real-world objects placed in the room. The effects of room acoustics are simulated using a virtual model of the room and information about the player's position in the room.

Audio memo and audible stickers

The Audiomemo application [25] allows the user to do binaural recordings of the surrounding sound environment. With these recordings, metadata about time, location and orientation of the user is saved. The user can later look at a map of his or her movements and listen to a recording made at a specific location, experiencing the same sound environment as when it was recorded. The recordings may also be uploaded to a server, allowing the user to share these recordings with others.

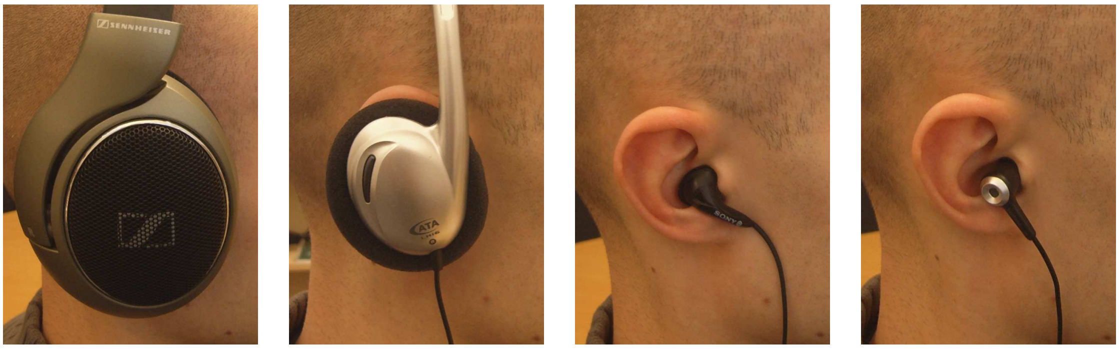
A somewhat similar application is the audible sticker application [36], which allows the user to leave audio messages at a specific location. These messages are played when the same user returns to this location or possibly when another user of the application arrives at this location.

3.3 Headphone acoustics

3.3.1 Headphone types

The ITU-T recommendation P.57 [37] defines five different headphone types, based on their placement in the ear. Circum-aural headphones enclose the pinna and sit on the surface of the head surrounding the ear (Figure 7(a)). Supra-aural headphones rest on the pinna without enclosing it (Figure 7(b)). Supra-concha headphones rest upon the ridges of the concha cavity while intra-concha headphones rest within the concha cavity (Figure 7(c)). Insert headphones partially or completely enter the ear canal (Figure 7(d)).

The recommendation also includes the definition of acoustically open and acoustically closed headphones. Acoustically open headphones intentionally provide an acoustic path between the ear canal and the external environment while acoustically closed headphones are designed to prevent any acoustic coupling between these.



(a) Circum-aural headphones. (b) Supra-aural headphones. (c) Intra-concha headphones. (d) Insert headphones.

Figure 7: Different headphone types.

3.3.2 Ear canal resonances

An open ear canal, with the inner end closed by the eardrum, acts as a quarter-wavelength resonator. The ear canal can be approximated by a cylindrical pipe open at one end and closed at the other, having resonances at

$$f_n = \frac{nc}{4\,(L + 0.61r)}, \tag{2}$$

where $n = 1, 3, 5, ...$, c is the speed of sound, L is the length and r is the radius of the pipe [7]. Because the pipe has one open end, the end correction $0.61r$ is added to the length of the pipe to obtain its acoustic length. With a length of 25 mm and a radius of 4 mm [38], the lowest resonance in a typical ear canal is located at about 3 kHz. The second-lowest resonance is thus located at about 9 kHz.

In sound transmission measurements from the open ear canal entrance to the eardrum performed on 12 subjects, the first peak in the amplitude response occured at frequencies between 3.0 and 5.5 kHz [38]. The ratio between the frequencies of the second and the first peak was in many cases close to 2.5:1, instead of the ratio 3:1 suggested by Equation (2).

When the ear canal entrance is closed, e.g., by a headphone earpiece, the quarter-wavelength resonance (and multiples of this) disappears and the ear canal acts as a half-wavelength resonator. Resonances are now located at

$$f_n = \frac{nc}{2L}, \tag{3}$$

where $n = 1, 2, 3, ...$ The lowest resonance is now located at about 7 kHz. As the earpiece is inserted deeper into the ear canal, the effective length of the canal becomes shorter, resulting in a higher resonance frequency.

To achieve natural reproduction of the surrounding sound environment in MARA applications using insert headphones, the half-wavelength resonance introduced when closing the ear canal should be filtered out [30]. Also, the quarter-wavelength resonance present in an open ear canal should be added if such a resonance is not already present in the magnitude response of the headphones.

3.3.3 Leakage

Headphones inserted into the ear canal provide attenuation of airborne sound. However, some leakage of sound is always present. The amount of leakage depends amongst other things on how tightly the earpieces of the headphones fit in the ear canal, and is largest at low frequencies.

This leakage of sound has two implications for mobile augmented reality audio. Because leakage is larger at low frequencies, the signal fed through the headphones should be equalized, so that low frequencies are not boosted when the leaked sound and the sound reproduced by the headphones are summed. The other implication is that the latency of the signal between the headphone microphone and driver should not be much longer than the time it takes the acoustic signal to travel this distance. If the time difference is significant, a comb-filter effect will be present in the resulting signal.

3.3.4 The pressure chamber principle

The sound pressure produced by a closed-back loudspeaker at low frequencies is proportional to the volume acceleration [39]. A closed headphone placed against the ear pumps on a cavity that is small in comparison with the wavelength at low frequencies, producing sound pressure proportional to the volume displacement. Constant excursion of the headphone driver thus gives constant pressure. A loudspeaker also has constant excursion below resonance, but the acceleration drop leads to a 12 dB drop in sound pressure level per octave. Any leaks present in the fitting of the headphones or in the construction of the headphones will cause a reduction in the sound pressure at low frequencies.

3.3.5 The occlusion effect

Vibration of the ear canal walls due to a bone-conducted signal produces acoustic energy in the ear canal. Normally, most of this energy escapes through the open ear canal entrance. However, if the canal entrance is closed, e.g., by a headphone, the energy is trapped inside the ear canal, and will be heard as amplified sound levels in comparison with the open-ear-canal case. This so called occlusion effect is pronounced at low frequencies. Due to this effect, sensitivity to bone conduction of sound below 2 kHz is raised by 10–25 dB when a headphone is worn over the ear [39].

The occlusion effect is familiar to hearing-aid users, who often complain about their own voice sounding hollow or having an echo [40]. Other sounds amplified by the occlusion effect include the sounds of chewing, breathing and swallowing.

There are several ways to reduce the occlusion effect. One option is to allow the bone-conducted energy to escape through vents in the headphones. The problems associated with this are the reduced low-frequency amplitude response of the headphones and the size of the vents needed. To have a significant effect, the vents should have a diameter of at least 2 mm [40]. Another possible problem specifically in MARA applications, is acoustic feedback from the headphone drivers to the microphones through the vents.

A second option to reduce the occlusion effect is deeper insertion of the headphone earpieces. Inserting ear molds into the bony part of the ear canal stops the soft tissue of the outer part of the canal from radiating sound into the inner canal end. Problems associated with this solution include discomfort and irritation inside the ear canal walls. A third option is to use active occlusion cancellation using a microphone inside the ear canal [41].

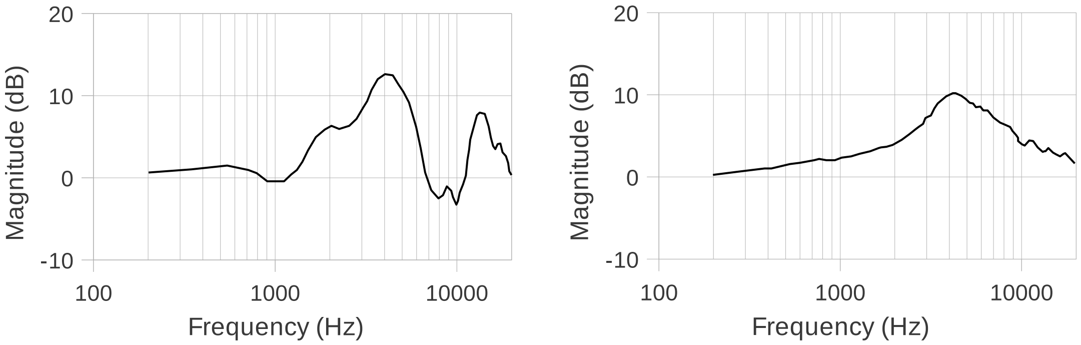
3.3.6 Headphone reproduction versus loudspeaker reproduction

Typically, headphones are used for reproduction of either music or radio programmes. This kind of material is normally mixed for reproduction using stereo loudspeakers in a not more than moderately damped room. When headphones are used instead of loudspeakers, they replace not only the loudspeakers but also the effects of the whole transmission path from the loudspeakers to the ears. Ideally, using appropriate signal processing, headphones could provide the listener with the same sound as reproduction through loudspeakers, but usually only some aspects of the loudspeaker reproduction chain are taken into account in a sound reproduction system using headphones.

A design goal for loudspeakers is commonly to produce a somewhat flat magnitude response in a free field. However, before the sound from the loudspeakers reaches the eardrums of the listener, many things happen. Direct sound from the left loudspeaker in a stereo setup reaches not only the left ear but also the right ear, and vice versa. In addition to the direct sound, reflections from different surfaces also reach the ears. The sound that actually enters the ear canal will have several different spatial localization cues added to it, as discussed in Section 2. Finally, the quarter-wavelength resonance of the ear canal will be present in the sound that reaches the eardrum (see Section 3.3.2).

In the case of headphone reproduction, there is no crosstalk, no reflections from the listening room, no spatial cues added, and the quarter-wavelength resonance of the unoccluded ear canal is replaced by the half-wavelength resonance of the occluded ear canal. Crosstalk, as well as simulated reflections and spatial cues, could be added to the electrical signal fed to the headphones, but generally the design criterion for headphones is that the magnitude response of the sound at the eardrum should be the same for headphone reproduction as for loudspeaker reproduction, giving the same timbre to the reproduced sound [42].

Two main equalization or calibration goals are used for the magnitude response of headphones. Free-field calibration simulates anechoic loudspeaker listening conditions, where only direct sound reaches the listener, normally from an azimuthal angle of 0°. Diffuse-field calibration assumes that reflections dominate over direct sound, and the headphone magnitude response is calibrated to match the magnitude response in a diffuse field, measured at the eardrum. The assumption of a mostly diffuse sound field is true if the distance from the loudspeakers to the listener is at least somewhat larger than the critical distance, where the level of the reverberant sound equals that of the direct sound. The critical distance depends on the directivity of the sound source and the absorption of sound in the room [43]. Figure 8 shows design goals for free-field and diffuse-field calibration of headphones.



(a) Free-field-calibration design goal.

(b) Diffuse-field-calibration design goal.

Figure 8: Design goals for headphones. After [42].

A subjective evaluation of headphone calibration goals showed preference for magnitude responses much closer to flat than to free-field-calibration and diffuse-field-calibration design goals [44]. The evaluation was performed using different target magnitude responses consisting of an otherwise flat response with a peak added with different amplitudes and at different frequencies near 3 kHz, where the quarter-wavelength resonance of the open ear canal is located. The tests showed that a peak of only 3 dB at 3 kHz was preferred when listening to music or speech samples.

3.4 Augmented reality audio headset

In previous KAMARA projects, different types of headsets have been used. These have been of insert type, quite effectively sealing the ear canal, and intra-concha type, loosely fit against the ear-canal opening. Some of them have been constructed from noise-cancelling headphones with built-in microphones (e.g. [30, 27]) while electret microphones have been added to others (e.g. [27]). The headsets used in the KAMARA 2010 project are modified Philips SHN2500 noise-cancelling headphones, also used in previous projects [30, 25, 31, 32].

These headphones are of insert type and have rubber sleeves around the drivers (see Figure 9), providing a tight fit into the ear canal and thus good attenuation of outside noise. Noise cancelling is done using microphones located at the end of the earpieces opposite to the drivers. The signal from the microphones is fed to a signal processing circuit, where a noise signal is extracted from it. The phase of the noise signal is then reversed and fed to the headphones. When this inverse-phase noise signal from the headphones sums with the already attenuated noise that has leaked through or past the headphones, this noise will be further attenuated.

For use in ARA applications, the headphones were modified by removing the noise-cancelling circuit and soldering two stereo-signal cables with 3.5 mm TRS

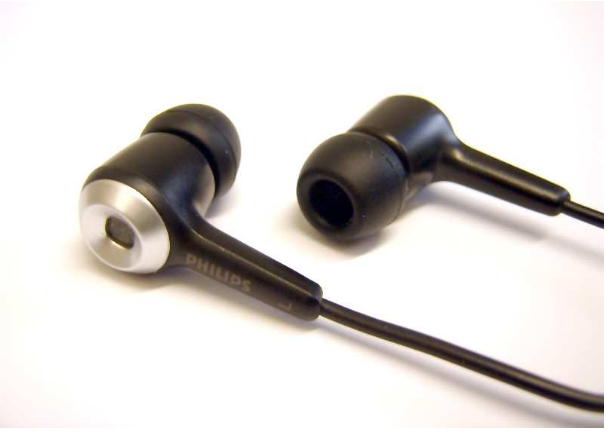


Figure 9: Earpieces of the Philips SHN2500 noise-cancelling headphones.

connectors for headphone signal input and microphone signal output in place instead of the original single connector for headphone signal input. To prevent short circuits and provide strain relief, the cable joint was protected by heat-shrink tubing with adhesive on the inside.

One concern with the headset is that when the earpieces are inserted into the ears, the microphones are located about 1 cm outside the ear canal entrance. At this distance, the spectral cues provided by the pinnae will exhibit some differences in comparison with the spectral cues available at the ear canal entrances. The earpieces themselves will also affect the sound field near the ear. If head tracking is used, cues provided by head movement should help with possible front-back confusion due to the deterioration of spectral cues.

Figure 10 shows the magnitude response of a Genelec 8030A loudspeaker, measured during an earlier KAMARA project [30]. The response was measured using one of the microphones of a Philips SHN2500 headset and a high-quality Brüel & Kjær 4191 free-field microphone as a reference. The responses are very similar up to about 4 kHz. Above 4 kHz there are some variations in the magnitude measured with the SHN2500 headset microphone compared with the reference microphone, but no serious flaws.

Figure 11 shows the magnitude response of the 8030A loudspeaker measured with the SHN2500 headset microphone rotated at different angles from the axis between the microphone and the loudspeaker. The responses show some directivity of the microphone at high frequencies.

Figure 12 shows the magnitude response of a Philips SHN2500 headset driver. The response was measured in three different conditions: in an ear canal simulator, in an impedance-matched tube and in a free field. The ear canal simulator is a 25 mm long plastic tube with a diameter of 9 mm and one end blocked by a hard wall. The response measurements made with the simulator show the half-wavelength resonance present in a closed ear canal as well as the amplification of bass frequencies due to the pressure chamber principle. In the matched-impedance tube, reflections and

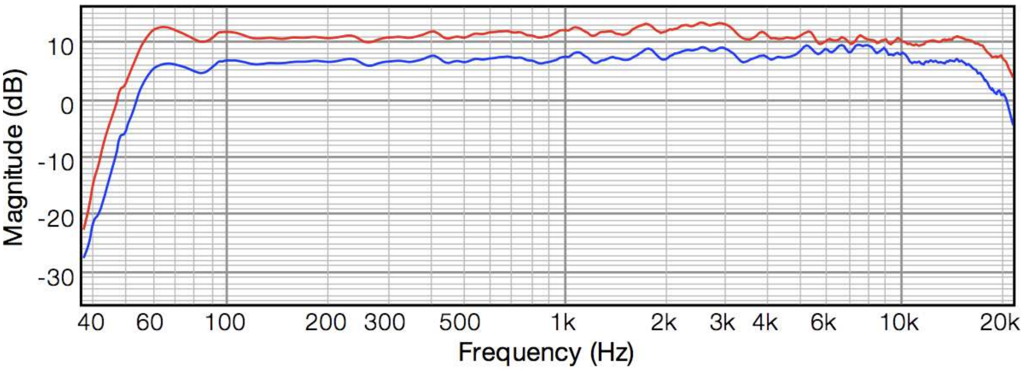


Figure 10: Magnitude response of a Genelec 8030A loudspeaker recorded with a Philips SHN2500 headset microphone (blue line) and a Brüel & Kjær 4191 free-field microphone (red line). From [30].

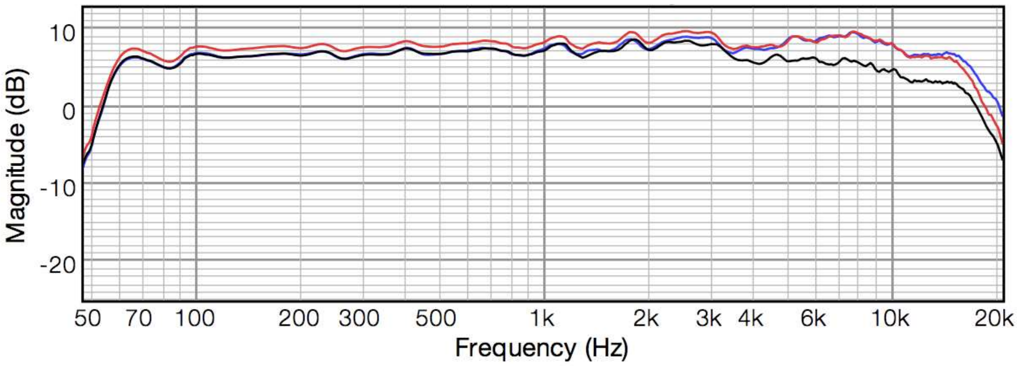


Figure 11: Magnitude response of a Genelec 8030A loudspeaker recorded with a Philips SHN2500 headset microphone in three different angles: blue is 0°, red is 45°, and black is 90°. From [30].

thus resonances are largely attenuated and the amplification of bass frequencies is much smaller. In the free-field response there are no resonances or any amplification of bass frequencies.

3.5 Augmented reality audio mixer

The mixer in the mobile augmented reality audio system performs several different tasks. As the name implies, the mixer performs mixing of the signal from the microphones with the signal containing the virtual sounds to be added. In addition, the mixer performs equalization of the microphone signal. The equalization consists of removing the half-wavelength resonance which is introduced with the occlusion of the ear canal, adding the quarter-wavelength resonance that is present in the open ear canal, and high-pass filtering to reduce the level of low frequencies which otherwise would be too high because of the leakage of low-frequency sound into

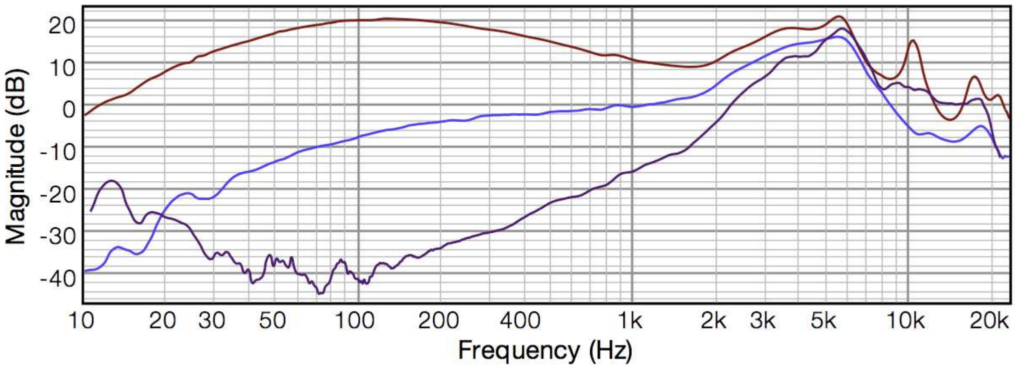


Figure 12: Magnitude response of a Philips SHN2500 headset driver measured in three different conditions: in an ear canal simulator (red line), in a matched-impedance tube (blue line), and in a free field (purple line). From [30].

the ear canal. High-pass filtering is also beneficial because insert headphones often have pronounced low-frequency magnitude response due to the pressure chamber principle, depending on the fitting of the earpiece to the ear canal [30].

The leakage of sound through and past the headphones also introduces the requirement that the latency of the mixer must be low. If the latency of the equalization circuit is much longer than the time it takes sound to travel the distance from the microphone to the driver elements of the headphones, a comb-filter effect will be heard as the leaking sound is summed with the almost identical but delayed pseudoacoustic sound.

3.5.1 Mixer design

For the KAMARA 2010 project, a novel mixer, based on the mixer developed in an earlier KAMARA project [30], was designed and ordered from an electronics prototyping company. Two essential requirements were made for the novel mixer. It should be smaller in size than the original mixer and thus easier to carry around. It should also connect to mobile devices and computers via USB, to enable input of stereo microphone signals to devices which only have a mono microphone input jack. The second requirement led to a design where the equalization and mixing circuit of the mixer were attached to a small USB sound card. This solution also allows the circuit to use the 5 V voltage supplied through the USB connector instead of being battery powered as the original mixer.

The novel mixer was intended to be tested together with the Mobile Augmented Messaging application (see Section 5). However, due to delays in the prototyping and manufacturing of the mixer, only one early prototype was available when the evaluation of the application was performed. Evaluation of the novel mixer was thus left outside the scope of this thesis. Below, the design of the equalization circuit of the original mixer is presented. The same circuit is used in the novel mixer.

To keep the latency low, equalization is performed using an analog electronic circuit. A first-order high-pass filter is used to attenuate low frequencies. The cut-off frequency of the filter can be varied between 6 and 720 Hz using a trimmer potentiometer.

To compensate for the missing quarter-wavelength resonance that is present in the unoccluded ear canal, and to remove the half-wavelength resonance that is added when the canal is occluded, two active peak/notch filters are used. Potentiometer adjustments can be made to select whether the filter produces a peak or a notch in the magnitude response and to select the Q factor and the center frequency of the filter. The center frequency can be adjusted from 0.7 to 3.2 kHz for the quarter-wavelength resonance filter, and between 1.8 and 8.5 kHz for the half-wavelength resonance filter.

3.5.2 Measurements for determining equalization parameters

To find suitable equalization parameters for the original mixer, transfer function measurements from a loudspeaker to a microphone inside the ear canal were performed [30]. In the first measurement, the ear canal was unoccluded and in the second measurement sound was routed through the ARA headset. The source loudspeaker was placed 2.5 m in front of the test subject. Equalization parameters were selected based on the comparison of these two measurements. Measurements were performed with four test subjects and a non-individual target equalization curve was calculated based on the measurements with all four test subjects.

Figure 13 shows the transfer functions measured with one of the test subjects. When comparing the natural open-ear-canal case (black line) with the unequalized pseudoacoustic representation (grey line), the pronounced bass response of the pseudoacoustic case can be seen. The quarter-wavelength resonance of the open ear canal around 2 kHz is missing from the pseudoacoustic representation. Instead, a half-wavelength resonance at about 7 kHz is present in the pseudoacoustic response. In this measurement, with the headset microphone turned at an angle of a little more than 90° from the axis between the loudspeaker and the listener, the directivity of the microphone can be seen as high-frequency attenuation. In the case of a diffuse sound field, the attenuation would be much smaller.

Individual target equalization curves are obtained as the difference between the two measured transfer functions. The equalization parameters are adjusted by hand to get an equalization magnitude response matching the target response as well as possible. Figure 14 shows the generic equalization curve calculated as an average of the individual equalization curves of all four test subjects. The curve was measured from the mixer after the equalization parameters had been adjusted to fit the target equalization curve. Figure 15 shows the magnitude response of the open-ear-canal case from Figure 13 compared with the response measured with an individually equalized ARA headset in use.

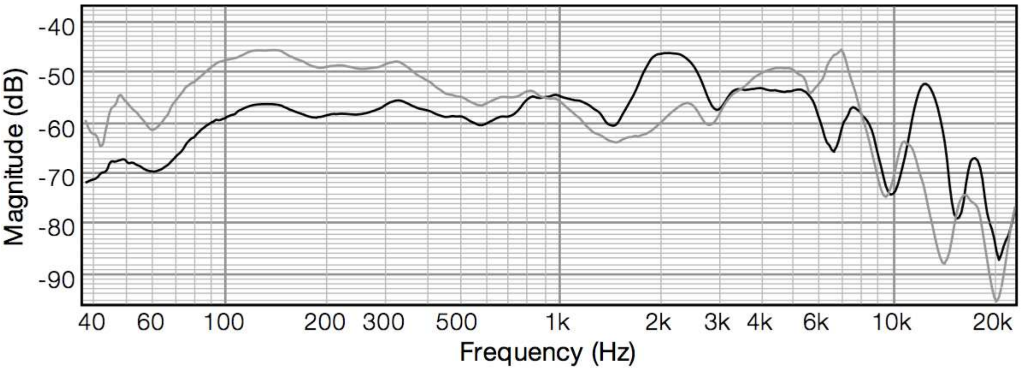


Figure 13: An example of the transfer function between the source loudspeaker and the microphone inside the ear canal. The black line shows the magnitude response with the ear canal open, and the grey line using the ARA headset without equalization. From [30].

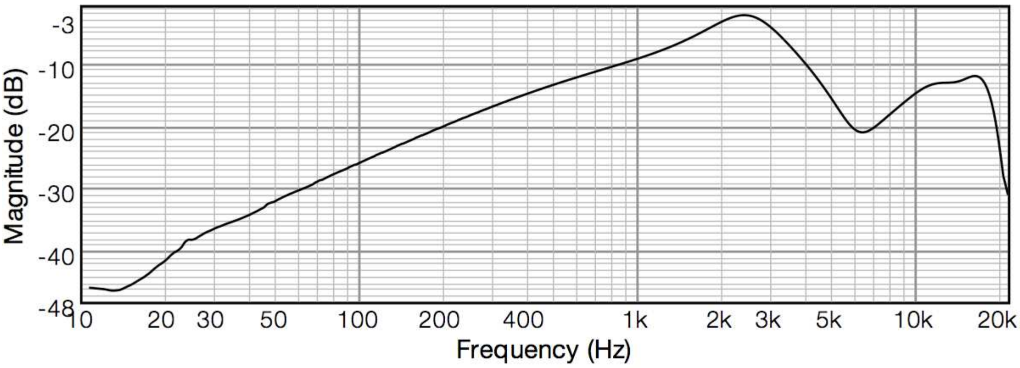


Figure 14: Generic equalization curve measured from the mixer. From [30].

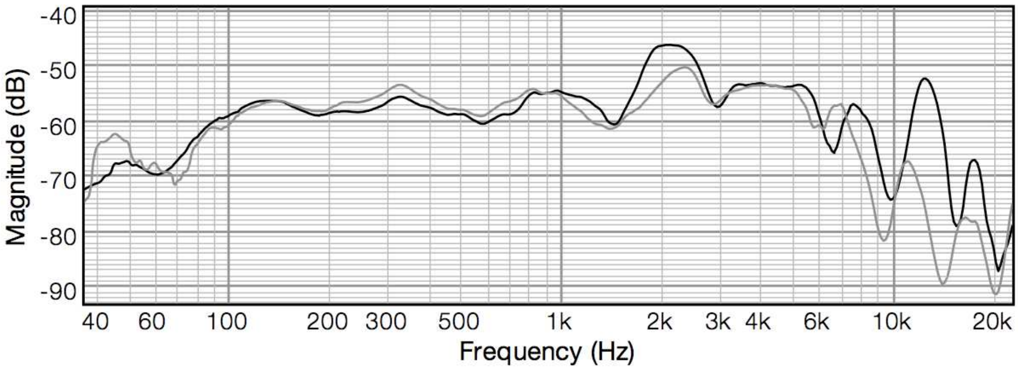


Figure 15: An example of the transfer function between the source loudspeaker and the microphone inside the ear canal. The black line shows the magnitude response with the ear canal open, and the grey line using the ARA headset with individual equalization. From [30].

4 Messaging applications

Messaging applications can be classified in different ways. They can, for instance, be divided into those intended for synchronous communication and those intended for asynchronous communication. Good examples of synchronous communication are phone calls and video calls. When talking to another person on the phone, the other person will hear what is said almost instantly, and is expected to respond to this right away. A discussion forum, on the other hand, provides an example of asynchronous communication. Messages left to the forum can be browsed and responded to when desired, and an immediate response is not expected.

The difference between synchronous and asynchronous communication is, however, not always clear. A text chat might be used for having an intensive discussion, with replies to messages given almost instantly. On the other hand, sometimes it might take hours or even days before someone reads a message and responds to it. In this case, the application allows for synchronous communication, but the actual communication done between the users is asynchronous.

Based on the amount of delay between one user inputting a message and another user receiving it, a separation can be made between real-time and non-real-time communication. In real-time communication, the delay is short enough to be completely or almost unnoticeable. Synchronous communication typically happens in real time, while asynchronous communication allows for longer delays. A normal phone call is an example of synchronous real-time communication. However, if the person being called does not answer, the caller might leave a message to an answering machine, making the communication asynchronous and non real time. A long delay introduced into a phone call would make it an example of non-real-time, synchronous communication.

Messaging applications can also be classified based on the audience of messages. For example, instant messaging (IM) applications are used for private conversations between a small number of people. The audience is thus predefined and limited in number. At the other extreme, a message sent using Twitter is normally available for anyone to read and may be indexed by Internet search engines. The audience might also be limited, but not predefined. For example, messages sent to a discussion forum could be shown only to registered users, but anyone might be allowed to register and thus read the messages at a later point in time.

4.1 Examples of messaging applications

Facebook

Facebook is a social networking website that allows several different types of communication, both synchronous and asynchronous. A user can communicate with his or her friends by posting messages either on the user's own wall or a friend's wall (the name of the space intended for these messages). Posts made to the user's own wall are often referred to as status updates, as these usually say, for instance, what the user is doing or has on his or her mind at the moment. These messages are text, but may include a photo, a video or a link to a web page or other web content.

The messages will be shown on the profile page of the user to whose wall they were posted, but they might also show up in other users' news feeds. In the news feed, a user is shown the activity of his or her friends, including posted messages, profile changes and upcoming events. Users can comment on the messages their friends post.

Facebook also has two methods for private communication. Private messages can be exchanged asynchronously between two or several users and a user can choose to receive an e-mail notification when receiving such messages. A user may also start a text chat with another user who is online. The Facebook chat supports XMPP (Extensible Messaging and Presence Protocol), which allows instant messaging clients to interoperate with the service.

Twitter

Twitter is a social networking and microblogging service, with which users can send and read messages called tweets. Tweets are text messages with a maximum length of 140 characters. These messages can be sent and received through the Twitter website, external applications, or SMS messages. Tweets are by default publicly visible on the user's profile page. A user can subscribe to other users' tweets, which is referred to as following the other user. These tweets are displayed on the follower's main Twitter page. Tweets can be grouped by adding hashtags specifying keywords or topics. A hashtag is added by prefixing a word in the message with a hash sign ("#"). To refer to or reply to other users, the at sign ("@") can be used to prefix a username.

Yammer

Yammer is a social networking service used for private communication within organizations. Users can join a company network by signing up using their company e-mail address. Yammer supports, among other things, posting of status updates, threaded conversations and private direct messaging between two or several users. Communication is done with text, with the option to attach images and other files. Messages and other content can be tagged by topic. Communication may be restricted to private or public groups within a company network, but can also take place between members of communities, which are larger networks not restricted to a single organization.

Internet Relay Chat

Internet Relay Chat (IRC) can be defined as a text-based teleconferencing system [45]. It uses a client-server model, with networks formed by one server or multiple servers connected in a tree structure, and clients each connected to one of these. In IRC, communication can be done either in named groups called channels, or privately between two clients. IRC offers extensive options for administrating channels. A channel operator can kick clients from the channel or completely ban a client from joining a channel. Channels can be open for anyone to join or require an invitation or a password. Clients are distinguished by nicknames with a maximum length of nine characters.

The IRC protocol is designed for synchronous communication. Messages are not stored on the servers, but directly sent to appropriate clients, so users will not be able to browse messages sent to a channel before the client connected to a server. Having a client running constantly allows a user to browse messages received when he or she has been away and thus communicate in a more asynchronous fashion. Client software, available for several different operating systems, may implement many de facto features not defined by the IRC protocol, such as highlighting of messages prefixed with the user's nickname which thus are addressed to the user.

Skype

Skype [46] is an application that can be used for making voice and video calls as well as for instant messaging. Both voice calls and instant messaging can be done between two or several users. Instant messages can be stored and browsed later, and a user can also edit his or her own recently sent messages.

Voice calls can be made not only to other Skype users. For a fee, users can call landline telephones and mobile phones. Also for a fee, users can get a local phone number in certain countries, making it possible to receive calls from telephones. Being able to receive voicemail messages is also subject to a fee.

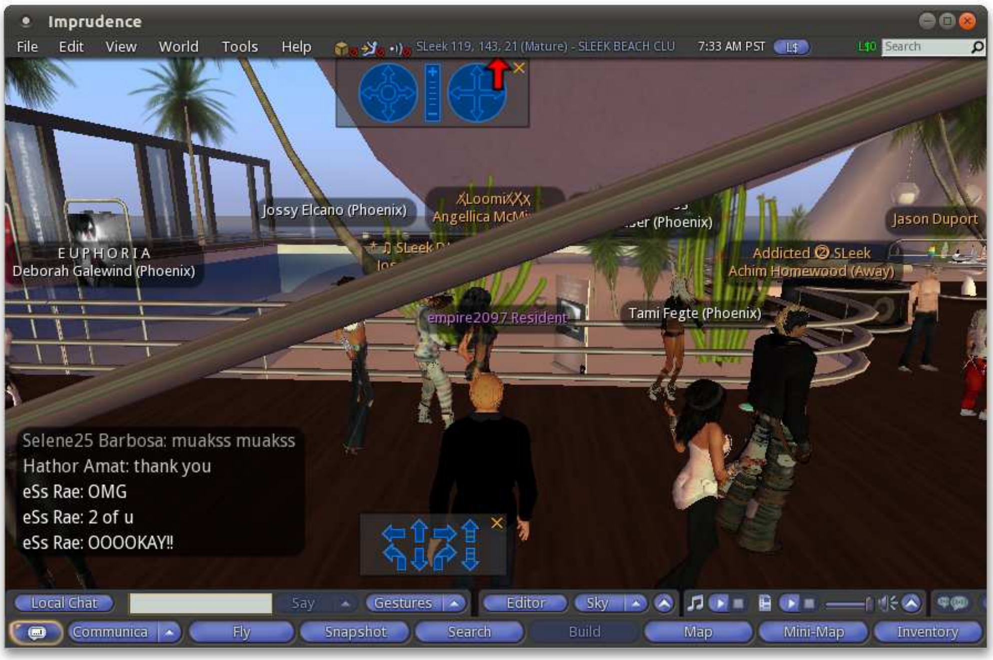
Second Life

Second Life [47] is an online 3D virtual environment where users can interact with the virtual world and other users through avatars, the users' alter egos in the virtual world. Users can, among other things, explore the world, socialize, and create and trade virtual property and services. Second Life has been used, e.g., for education and research (for examples of this, see [48]).

Second Life offers both text and voice chat in real time. The voice chat uses spatial audio rendering, placing the sound in the direction of the avatar speaking, and attenuating the sound as the distance to the avatar increases. Text chatting can also be done locally, between all avatars within a certain distance of each other. Alternatively, instant messages can be sent to individual avatars or groups of avatars regardless of their location in the virtual world. Figure 16 shows a view from a beach club in the Second Life virtual world, with the local text chat displayed at the bottom left.

4.2 Messaging applications in a mobile context

Facebook has a website optimized for mobile devices (see Figure 17(a) for an example of a profile page viewed through this website), and for several mobile platforms there are also applications that provide a tailored interface to a range of Facebook features. Some of these applications offer, e.g., the possibility to synchronize Facebook friends with the contact list of a mobile phone. These applications can add profile pictures and status updates from Facebook to the contact list, among other things.



Selene25 Barbosa: muakss muakss
Hathor Amat: thank you
eSs Rae: OMG
eSs Rae: 2 of u
eSs Rae: OOOOKAY!!

Figure 16: A view from a beach club in the virtual world Second Life, shown using the Imprudence client software [49]. The local chat is displayed at the bottom left, showing messages from avatars within a certain distance.

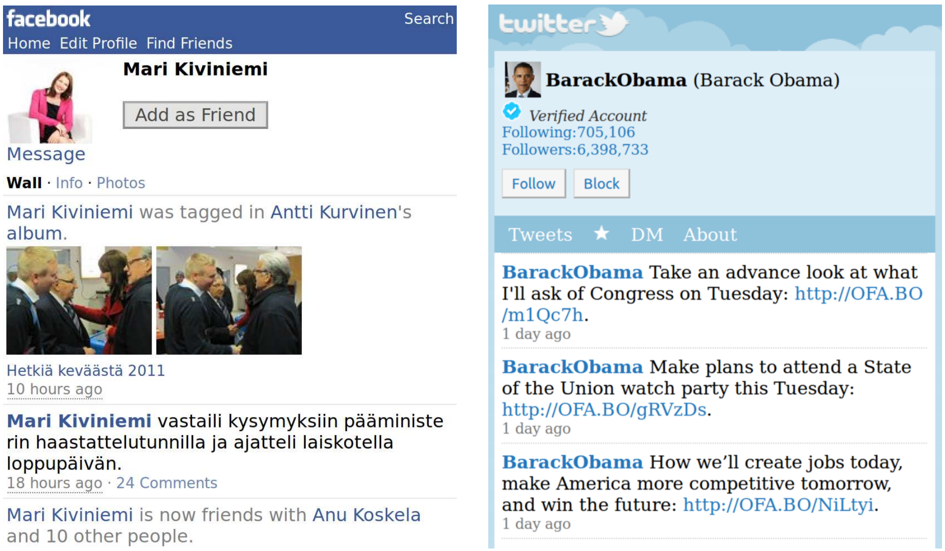
Like Facebook, Twitter can be used on mobile devices through applications available for several mobile platforms or through a website optimized for such devices (see Figure 17(b)). Messages can also be sent and received as SMS messages.

In addition to applications available for a number of mobile platforms, Yammer offers several different methods for sending and receiving messages. Messaging can be done using instant messaging clients, e-mail and SMS.

IRC can be used on mobile devices through client software available for several different platforms. Alternatively, an IRC client can run constantly on an SSH server and be used, when desired, by connecting to this server with an SSH client. This does not only allow for a more asynchronous style of communication, but also makes it possible to use the same IRC client from both mobile devices and desktop computers.

Skype is available for several different mobile phones and can thus, in some cases, be used as a replacement for making normal mobile phone calls, especially with the additional services of receiving calls from and being able to call regular telephones. Figure 18 shows a Skype contact list as displayed on a mobile phone running the S60 platform.

Second Life client applications showing the virtual world in 3D are only available for desktop and laptop computers, and not for mobile devices. However, there are applications for mobile platforms supporting the local text chat and instant messaging features of Second Life.



facebook
Search
Home Edit Profile Find Friends

Mari Kiviniemi

Add as Friend

Message

Wall · Info · Photos

Mari Kiviniemi was tagged in Antti Kurvinen's album.

Hetkiä keväästä 2011
10 hours ago

Mari Kiviniemi vastaili kysymyksiin pääministerin haastattelutunnilla ja ajatteli laiskotella loppupäivän.
18 hours ago · 24 Comments

Mari Kiviniemi is now friends with Anu Koskela and 10 other people.

twitter

BarackObama (Barack Obama)

Verified Account
Following:705,106
Followers:6,398,733

Follow Block

Tweets ★ DM About

BarackObama Take an advance look at what I'll ask of Congress on Tuesday: http://OFA.BO/m1Qc7h.
1 day ago

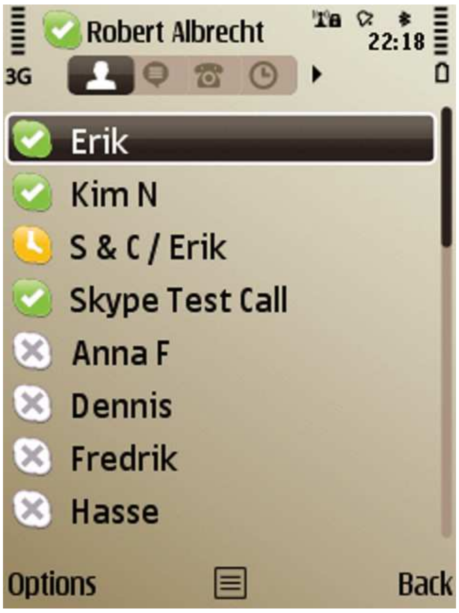
BarackObama Make plans to attend a State of the Union watch party this Tuesday: http://OFA.BO/gRVzDs.
1 day ago

BarackObama How we'll create jobs today, make America more competitive tomorrow, and win the future: http://OFA.BO/NiLtyi.
1 day ago

(a) Facebook profile page.

(b) Twitter profile page.

Figure 17: Examples of profile pages viewed through Facebook's and Twitter's websites for mobile devices.



Robert Albrecht
22:18
3G

- Erik
- Kim N
- S & C / Erik
- Skype Test Call
- Anna F
- Dennis
- Fredrik
- Hasse

Options Back

Figure 18: A Skype contact list as displayed on a mobile phone running the S60 platform. Contacts with a grey icon to the left of their name are not logged in. A green or yellow icon means the contact is logged in, but the yellow icon indicates that the contact is "away", i.e., has not been using the computer or mobile phone for a while.

5 The Mobile Augmented Messaging application

5.1 Description of the application

Asynchronous messaging is typically done using text. Many social networking applications, such as Facebook, Twitter and Yammer, provide examples of this. When inputting text, the user's attention is drawn to the display and text input interface of the device he or she is using, and especially on mobile devices text input can be cumbersome and slow. When other users read these messages, they have to focus their attention on the display of the device they are using.

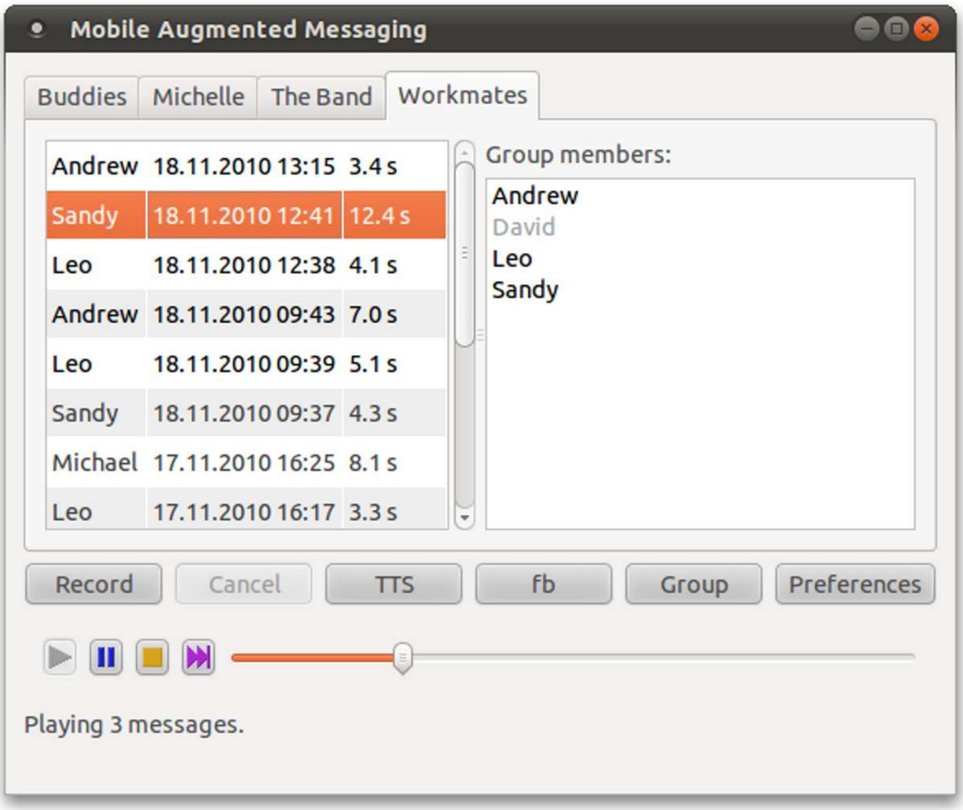
Sharing content using audio instead of text would mean only minimal attention on the display and keyboard, keypad or touch screen of the device is required, leaving the user's sight and hands free for other tasks. Ideally, the user interface could be made completely audio based using speech recognition and synthesis. Especially on mobile devices, message input using audio would be significantly faster than inputting the same message as text. Using the ARA headset and mixer allows playback of audio while perceiving the surrounding sound environment unattenuated. Recording of audio can also be done using the microphones of the headset.

To test the concept of sharing short audio messages on a mobile platform, the Mobile Augmented Messaging application was developed. The application allows non-real-time, asynchronous communication using audio messages, which are shared between members of groups. Audio messages are recorded using the microphones of the ARA headset, and are then stored on a database server. For situations where the recording of audio messages is not possible or desirable, messages can be input as text and converted to audio using text-to-speech synthesis. Users can listen to new messages automatically when they arrive or browse and listen to old messages.

Figure 19 shows the graphical user interface of the Mobile Augmented Messaging application. Groups are shown as tabs at the top of the application window. Users can create new groups and invite other users to join these. Messages are shown ordered by the time they were recorded, showing the length of the message in seconds. Members of the selected group are shown to the right of the messages, with users offline in grey colour. In addition to sharing audio messages with other users of the application, users can make binaural recordings and post them on their Facebook wall, allowing other Facebook users to listen to these. At the bottom of the window, there are playback controls and a status bar. From the Preferences window, users can select what should be done when new messages arrive. The alternatives are to play them when they arrive, to play a notification sound or to do nothing. Users can also select what kind of spatial sound rendering should be performed for messages.

5.1.1 Spatial separation of sound sources

As the application is not intended for real-time communication, there is no requirement for multiple messages to be played simultaneously. To allow for better perception and understanding of the messages, they are therefore always played one at a time.



Andrew	18.11.2010 13:15	3.4 s
Sandy	18.11.2010 12:41	12.4 s
Leo	18.11.2010 12:38	4.1 s
Andrew	18.11.2010 09:43	7.0 s
Leo	18.11.2010 09:39	5.1 s
Sandy	18.11.2010 09:37	4.3 s
Michael	17.11.2010 16:25	8.1 s
Leo	17.11.2010 16:17	3.3 s

Group members:
Andrew
David
Leo
Sandy

Playing 3 messages.

Figure 19: The graphical user interface of the Mobile Augmented Messaging application.

In addition to the separation in time for the playback of individual messages, some kind of separation of messages from different groups is desirable. If a speaker belongs to several groups which the listener also is a member of, the listener needs means of knowing which group the speaker's message is directed to. The required information can of course be provided by the graphical user interface of the application, but the user will not need to look at the display if this information also is provided by auditory cues.

One option for presenting an auditory cue of the group a message is directed to, is simply to have some short audio or speech signal associated with the group played before the message. For this application, however, it was decided that messages directed to different groups would be spatially separated.

Spatial separation can be performed based on both the direction and the distance of the sound source. In addition to any informational value associated with the location of a sound source, the spatial separation of speech sources improves speech intelligibility in the case of multiple simultaneous speakers. This has been shown to be true in the case of speaker separation based on different azimuthal angles [33], as well as distance in the near field [50]. These additional benefits would be of use, e.g., in a real-time multi-talker chat application, but are not applicable to the

Mobile Augmented Messaging application, where multiple messages are not played simultaneously.

In this application messages directed to different groups are separated by different azimuthal angles. Only angles between 90° and -90° in front of the listener are used, to avoid problems with front-back confusion, which are likely to occur when using non-individual HRTFs. All messages arrive from an elevation angle of 0°, avoiding possible up-down confusion. Figure 20 shows the spatial distribution of messages from the different groups in Figure 19. If the user only participates in one messaging group, the messages from members of this group will instead be separated along the 180° arc mentioned.

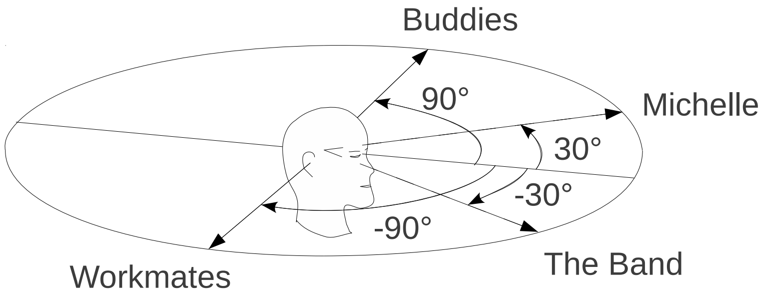


Figure 20: Angular distribution for spatialization of messages from the different groups in Figure 19. The groups are evenly distributed in alphabetical order from 90° to -90° in front of the listener.

5.1.2 Sharing binaural recordings

The possibility to make binaural recordings and post links to these on Facebook was added to the application, to allow sharing of binaural recordings with a larger audience. Other Facebook users can download and listen to these recordings, having almost the same auditory experience as the user who made the recording. The Mobile Augmented Messaging user can thus share, for instance, sounds from the nature, from a concert or from a sports event to his or her friends on Facebook. The links are accompanied by a short message about the recording, and a description which can, e.g., tell other users that this is a binaural recording that should be listened to using headphones. In addition to making new recordings and posting these, the Mobile Augmented Messaging user can browse through previously posted recordings and remove these, if wanted.

5.2 Implementation of the application

An important aspect when choosing the technologies used for the implementation of the application was portability. To be able to gain popularity, a messaging application like this one should work on different hardware and software platforms, although this would not be necessary for the mere testing of the concept of the application.

One possibility would be to do the implementation as a web application using HTML and Javascript on the client side, making it usable in any ordinary web browser. The user interface of the application would be presented in the user's web browser while the signal processing and most of the application logic would be performed on the server side. However, a major problem with this implementation would be that audio messages would have to be recorded using a separate program, and then uploaded to the server using the web browser, making the recording of messages cumbersome. For this reason, the idea of a web-browser-based application was dropped.

The application is implemented using a client-server approach. The server side of the application is mainly handled by a MySQL [51] database server used for the storage and retrieval of audio messages and associated metadata, as well as information about users and groups. The client application communicates directly with the database server. The signal processing as well as most of the logic involved is thus performed by the client application.

The client application is written in the C++ programming language and uses the cross-platform Qt application and user interface framework [52]. Qt provides components, e.g., for building the graphical user interface and for handling network and database access. The PortAudio library [53] is used for audio recording and playback. Audio processing is done using floating-point operations to keep the implementation simple. To allow for better performance on mobile devices with limited processing power, fixed-point operations should preferably be used. However, the processing delay is not a significant factor for this application, because messaging is not done in real time. Most database operations, as well as audio processing, are done in threads separate from the one that handles the graphical user interface. These operations will therefore not block the user interface.

For testing purposes, the application has been compiled for both Ubuntu Linux and Microsoft Windows operating systems. The components used should allow porting the application to other platforms such as Mac OS X. The Qt framework is also supported on the Symbian platform and mobile Linux platforms such as MeeGo, but porting the application to these would require, e.g., redesigning the graphical user interface for the small displays of mobile devices.

5.2.1 Recording speech using binaural microphones

When recording speech using binaural microphones, the sound will be coloured in comparison with the speech recorded with a microphone located in front of the speaker. A compensation filter should ideally be used to remove this colouration and make the recorded speech sound more natural. The filter can be calculated based on transfer functions from the mouth to the ears and from the mouth to a location in the far field in front of the speaker. However, no compensation filter is used in the current implementation of the application.

For the Mobile Augmented Messaging application, speech is meant to be the primary content of the messages. It would thus be beneficial to extract the speech of the user from background noise and ambient sounds. However, the ambient signal might also provide information relevant to the message. Ideally, the speech signal could be extracted and spatialized, so that it is heard to arrive from a certain direction, while the ambience of the recording should be heard as is.

In the Mobile Augmented Messaging application, no speech extraction is performed. Instead, the stereo channels from the binaural recording are summed to produce one mono signal. This summation will amplify the user's own speech and other sound from sources located in the median plane up to 6 dB. Sound from other sources will be amplified or attenuated, depending on the frequency and the interaural time difference. The ambient sounds are thus not removed, although partially attenuated, and are spatialized together with the speech of the user. The summation of the binaural signals is essentially beamforming (see, e.g., [54]). This could be performed separately and adaptively in different frequency regions, cancelling sources in other directions [27].

5.2.2 Speech synthesis

Text-to-speech synthesis is done using the eSpeak speech synthesizer [55] and performed on the server side. Including eSpeak text-to-speech synthesis in the client application would have introduced some restrictions to the portability of the application. eSpeak offers a large selection of different languages and dialects. It also offers the possibility to modify synthesis parameters, such as speed, volume, pitch and pitch range. The possibility to modify these parameters could allow users to customize the synthesized voice they use, making it possible for other users to recognize the sender of a synthesized message based on the voice. Modification of voice parameters was, however, not included in this implementation of the application.

5.2.3 Spatial sound rendering

For spatial sound rendering, head-related impulse responses measured at the Massachusetts Institute of Technology Media Lab using a KEMAR (Knowles Electronics Mannequin for Acoustics Research) dummy head and torso are used [8]. These impulse responses have been obtained using maximum-length sequences (see, e.g., [56]) at a 44.1 kHz sampling rate. The loudspeaker used in the measurements was positioned at a distance of 1.4 m from the dummy head, at elevation angles from -40° to 90°. The impulse responses used for this application have been diffuse-field equalized.

The convolution of audio messages and the appropriate head-related impulse response is done by multiplication in the frequency domain, to speed up the process compared to time-domain convolution. The discrete Fourier transforms of the impulse response and the audio signal are calculated using a fast Fourier transform (FFT) and the convolution is done using the overlap-add method (see, e.g., [57]). The resulting signal with the HRTF applied is then obtained using the inverse FFT.

Two alternatives to using HRTFs for spatial sound rendering were also added to test the possible advantages of using HRTFs over much simpler methods in this kind of application. These alternatives simply produce a level difference between the left and right channels, and thus require much less processing power than using HRTFs does. The first alternative is linear amplitude panning, where the amplitude of the signal passed to each ear depends linearly on the azimuth of the sound source. When the sound source azimuth is 0°, both ears receive the same signal. When the azimuth increases towards 90°, the amplitude of the left channel linearly increases and the amplitude of the right channel linearly decreases, so that an interaural level difference of 20 dB is produced when the azimuth is 90°. When the azimuth is $-90°$, the amplitude of the right channel will instead be 20 dB larger than the amplitude of the left channel.

The second alternative uses interaural level differences extracted from the HRTFs and averaged over the linear frequency scale, thus emphasizing high frequencies where larger level differences are encountered. These averaged level differences are then applied equally over the whole frequency range of the message signal. This approach should provide better mapping between the desired and the perceived azimuth of the sound source than the linear amplitude panning does. Figure 21 shows the amplification of the signal passed to the left ear, both using linear amplitude panning (red line) and average ILDs extracted from the HRTFs (blue line). The amplification of the signal passed to the right ear is the same as for the left ear but mirrored about 180° azimuth. Figure 22 shows the interaural level differences produced by these amplitude-panning methods.

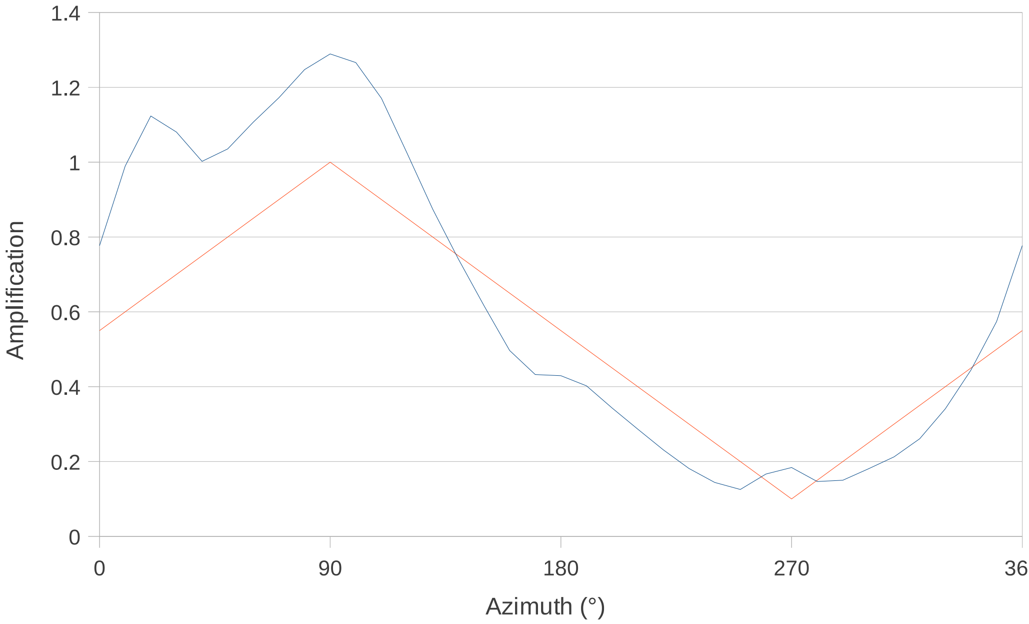


Figure 21: Amplification of the signal passed to the left ear for different sound-source azimuths in the horizontal plane. The red line represents linear amplitude panning, while the blue line represents average amplification levels extracted from HRTFs.

Users have the option to add reverberation to the messages played. This could

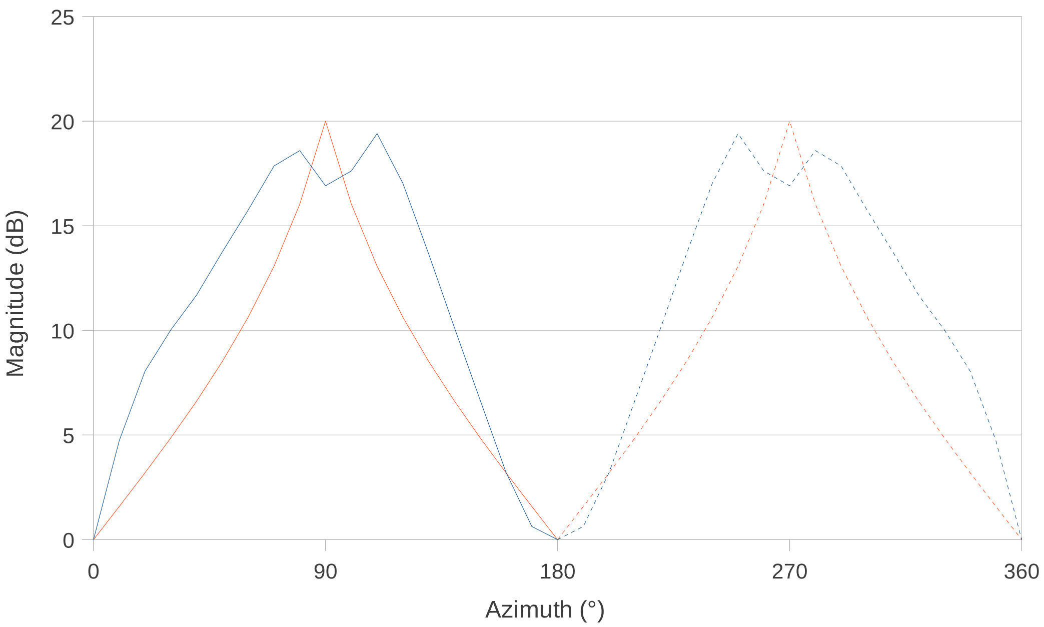


Figure 22: Interaural level differences produced by the two amplitude-panning methods for different sound-source azimuths in the horizontal plane. The red line represents linear amplitude panning, while the blue line represents average interaural level differences extracted from HRTFs. For the solid line, the level of sound arriving to the left ear is higher. For the dashed line, the level of sound arriving to the right ear is higher.

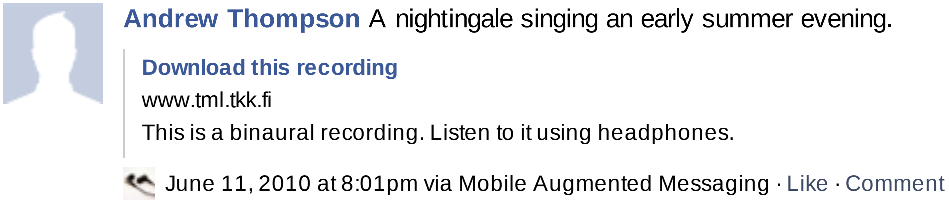
possibly aid in externalization, especially in the case of synthesized speech, where no natural reverberation is present. Reverberation is added by convoluting the audio message and an artificial reverberation impulse response created using the GVerb plug-in included with the Audacity audio editor [58].

In addition to listening to spatialized messages, users also have the option to listen to the original binaural recordings without any spatialization. Users will thus hear the original acoustic environment as the user who recorded it did.

5.2.4 Sharing binaural recordings

Facebook provides the Graph API for searching, accessing and modifying content [59]. An application registered on Facebook must use an embedded web browser to load the Facebook login page where the user is authenticated and is asked to authorize the application to access and modify content. If authorization is given, the application receives an access token, which is used for accessing and modifying content using HTTP GET, POST and DELETE requests.

The binaural recordings made using the Mobile Augmented Messaging application are made available for download using the Apache web server software and a PHP script which fetches the audio data from the MySQL database. A link to the recording's download address is posted on the user's wall on Facebook. A message accompanying the link, as well as a description saying that the link points to a binaural recording that should be listened to using headphones, is also included. Figure 23 shows an example of a link to a binaural recording posted on Facebook.



Andrew Thompson A nightingale singing an early summer evening.

Download this recording
www.tml.tkk.fi
This is a binaural recording. Listen to it using headphones.

June 11, 2010 at 8:01pm via Mobile Augmented Messaging · Like · Comment

Figure 23: Example of a link to a binaural recording posted on Facebook.

6 Evaluation of the Mobile Augmented Messaging application

6.1 Methods

Evaluation of the application was done with a group consisting of three test users. An additional test user was not able to participate due to compatibility issues between different operating system versions. All of the test users were Masters of Science in different fields of technology. The group consisted of former flat mates communicating with each other on almost daily basis using IRC, and was thus a suitable target group for the application.

The application should preferably have been evaluated using a mobile platform, but due to limited resources and technological limitations, mostly the lack of stereo microphone input support on some potential mobile devices, the application had to be tested on desktop or laptop computers. The advantage of faster and easier input of speech than text was thus not realized to the extent that it would have been on mobile devices. The limited mobility of the test platform also hindered the making of binaural recordings in different sound environments.

The application was intended to be tested using the ARA headset and the novel ARA mixer. However, due to delays in the delivery of the novel ARA mixer, the application had to be tested without the mixer, but the headset could still be used directly connected to the computer.

The test users were given the application executable and other necessary files, as well as a presentation of the usage and features of the application. The users were instructed to test the application for a period of about one week, using it, e.g., for the kind of communication they would normally do using IRC. The users were encouraged to try the different functionality and options of the application, and to write down comments during the test period. During the testing, messages were sporadically exchanges between the test users. At the end of the test period, a one-hour test session was organized, to evaluate the usability of the application for more intensive chatting.

After the test period, the test users were individually interviewed. No formal usability testing was done, as the focus of the evaluation was on gathering comments about the concept of the application rather than the usability of the implementation, although the test users were asked also for feedback on the usability of the implementation.

The interviews with the test users were done based on the following questions non-exclusively:

- What kind of communication was the application used for during the testing?
- Did you find the application to be of use?
- What kind of communication could the application be useful for?
- Did you prefer recording messages or writing them as text?

- Did you prefer listening to recorded or synthesized messages? How did these mix?
- What action for new messages did you prefer (playing them, playing a notification sound, or doing nothing)?
- How would you describe the difference between using HRTFs, ILDs and amplitude panning?
- What kind of effect did reverberation have?
- At what distance did you perceive the other speakers?
- What impression did the distribution of speakers or groups in different directions give you? Was it useful?
- Did you find the non-real-time delivery of messages a restriction?
- What were your experiences regarding the posting of binaural recordings on Facebook? Did your friends comment on your recordings and if so, how?
- Did you find any aspect of using the application difficult?
- How would you compare the usage of this application with that of IRC?
- Would you use this application if it were available for your mobile phone?
- Do you have any ideas for improvement of the application?
- Other comments?

6.2 Results

Below are the test users' comments on different aspects of the application. These comments are discussed further in Section 6.3.

What was the application used for during testing and what could it be used for?

During the testing, the application was used for free conversation on different subjects. The application itself was also a subject of discussion, and feedback about the application was given using the application itself. One of the users sang and played a song on the guitar for the other users. Testing the application on a desktop or laptop computer restricted the use of the application, as the users were tied to sitting in front of the computer when using the application.

To the question if the application was of any use during the testing, the users answered that it was as useful as IRC or instant messaging applications would have been for having the same discussions. One user suggested that the application could be useful for meetings, while another user suggested that the application could be used by band members to share ideas for new songs with each other.

Spatial separation of sound sources

One of the users did not think that spatially separating messages from different groups was very useful. The user did not either think that this kind of application was very well suited for having several conversations at the same time. When using the application on a mobile device, with messages arriving perhaps every ten minutes, it could, however, be useful to know which group a message was sent to without having to look at the display of the device. In meeting-type conversations, hearing the voices of different people from separate directions might make the conversation a bit more natural. The user also suggested that the voice of a person could be heard from the direction where that person is located. This information could then be used to locate that person, e.g., to have a discussion face to face.

Another user said that hearing messages sent to different groups from different directions was not very useful during testing, because the different groups consisted of the same people talking about the same subjects. The user thought that it might be useful when having several totally separate conversations, as a quick way to get information about the group a message was sent to.

The third user thought it was difficult to hear which direction a sound arrived from, even with only three groups spaced at 90° angles. The user pointed out that this would limit the number of groups that could be separated this way, and thought that it would be even more problematic in a noisy environment. Using another method to identify different groups, such as playing signature tunes, was suggested.

Auralization

One of the users did not find that any of the auralization options would have given the impression that the speakers were in the same room with the listener. He speculated that the quite high level of hiss present in the recordings could be one reason for this. When adding reverberation, it felt like the speakers were in a completely different place, sounding like a huge hall. The user thought adding reverberation could be of more use if the reverberation adapted to the environment the listener is in.

The user perceived the speakers as being pretty close to the listener. When using linear amplitude panning, there was no sense of distance, it only sounded like the sound came from either earpiece of the headphones. The sound sources were perceived most clearly to be in different directions when using linear amplitude panning, but the other alternatives were more pleasant to listen to.

The other two users did not find much difference between the different methods for adding sound source directionality. One of them thought that the amplitude panning method using ILDs extracted from HRTFs sounded a bit better than using HRTFs. Both users said that the other speakers sounded like they were pretty close to the listener, but that it did not sound like they were inside the listener's head. When listening to guitar play, it sounded a bit like the guitar would have been on the listener's lap. Both users also found the sound to be quite pleasant when adding reverberation. One of them pointed out that it did not sound like the speakers were in a huge hall.

One of the users reported having a severe hearing loss in one ear, making it impossible to use a telephone with that ear. In this application, however, it was possible to understand the messages, with a bit of concentration, as there always was some sound heard through both channels.

Speech synthesis

The quality of the synthesized speech was described as not being very natural or pleasant. Sometimes there was need for listening to a synthesized messages several times to understand the message. This was especially a problem when the message contained words in a language other than that used for the synthesis. One user thought it to be silly if one participant in a conversation is speaking while the other uses speech synthesis. This might give the impression that the other user might have something to hide. Another user thought there was a bit of a science fiction feeling when listening to synthesized speech, but the messages could be clearly understood, and he did not think that mixing real and synthesized speech caused any confusion.

One user thought that a lot of the usefulness of the application disappears if speech synthesis is used instead of real speech, as listening to a synthesized message does not give any additional information compared with reading the same text. Using speech synthesis was, however, seen as a good alternative for inputting messages in situations when it is not possible to record speech, e.g., in the presence of heavy background noise.

Arrival of new messages

All three users thought that playing messages automatically as they arrive was the best alternative when having an active conversation. One of the users thought that this also would be a good alternative when using a mobile device, while another user pointed out that when doing something important with the mobile device at the same time, it might be a good alternative not to play the messages automatically. When using the application on a desktop or laptop computer and not having an active conversation, it might be a better alternative to use a notification sound, because messages that are played automatically might go unnoticed if the user does not sit by the computer when they arrive.

Intensive chatting

During the intensive chat session, there were pauses in the conversation, just like in an IRC or IM conversation. Recording and listening to long messages produced equivalently long pauses in the conversation. One user suggested that it might be good to try to keep the messages a bit shorter when using this kind of application.

The pauses were, however, not experienced as particularly disturbing or problematic, when using the application for the same type of conversation as IRC typically is used for, and not assuming that the conversation should work the same way as when talking on the phone. The users could do other things while chatting, e.g., play the guitar or surf the Internet. Compared with IRC it was difficult to quickly get information about the contents of messages. It was considered easier to follow several different discussions at the same time using IRC.

During the chat session, it often happened that while recording the answer to some question, someone else commented on another message in between. Still, with the amount of messages shared during testing, there were no problems understanding which message a certain comment referred to. One user commented that in some cases this might actually be an advantage compared with real-time communication, where one person speaks at a time. In this application, every participant can say what they want at any time, without having to wait for their turn while someone else is speaking.

Advantages of speech communication compared with text communication

All users thought the application would be useful on mobile devices, because it does not require attention on the display and text input interface of the device. If the user keeps the headphones in his or her ears all the time, new messages can be heard as they arrive, without having to touch or look at the device. Other tasks, such as recording a message, require some amount of input using, for instance, a touch screen. The usage of voice commands using speech recognition was suggested, to allow interaction with the application while keeping the mobile device in the pocket.

All users pointed out the advantages of communicating using speech instead of text, because of the information provided in addition to the textual content of the messages. The tone used, e.g., tells the listener more about the actual intentions and sincerity of the speaker. Speech communication is much more expressive than text communication, where there is need, e.g., for interpreting smilies and there is a bigger risk of misunderstandings. Recording messages also makes it possible to record things other than speech, either in the background or as the main content of the message.

One of the users noted that in some cases speech communication can set a slower tempo on a conversation than text communication using IRC, because of the time required to listen to messages arriving. Using IRC, a lot of messages might arrive at the same time and it might be difficult to react to them all. With voice messages, however, users probably will not send new messages before having listened to all the messages that have arrived but have not been played yet.

Disadvantages of speech communication compared with text communication

IRC or other text chats were thought to be better for participating in multiple discussions at the same time. In text chats, it is easier to browse through old messages, and there is no confusion as to which group a message is sent to. The possibility to listen through all unread messages might, however, be useful when the user cannot focus his or her attention on the display of the device. One user suggested the option to select one group that the user is actively participating in at the moment. Only new messages from this group would be played automatically. Two of the users noticed that in the current implementation of the application, new messages that arrive while the user is recording a message are played while the message is still being recorded. New messages should preferably be played after the recording has been completed.

Two of the users described sending voice messages as maybe a bit frightening, as there is no option of correcting any mistakes while recording, like you would be able to correct typing errors while writing a message. When recording messages, there is usually some nervousness present, especially when the user is not that familiar with the application. The third user, however, saw the advantage of having the time to think what one wants to say and if one actually wants to say it while recording a message, compared, e.g., with a normal teleconference, where the other participants hear everything as it is being said.

One of the users noticed that any sounds in the surrounding environment are heard quite easily in the messages, and wondered if this might be a big problem in noisy environments. He also noted that it sometimes might be desirable to hear sounds other than the user's voice, and sometimes not. The same user also noticed quite a lot of hiss in the recordings, and wondered if this could be filtered out.

Browsing and playing old messages

As pointed out earlier, it is quite fast to browse through messages in a text chat. Browsing through audio messages, however, requires listening to at least a part of them to know their contents, and is thus much slower. The users thought that a textual summary of the message contents would be good, but doubted if anyone would take the time to type the text after recording a message. The current implementation of the application lacks the possibility to address a message to a certain user in a group, a feature available in IRC. One can of course say the name of the person the message is addressed to in the message itself, but some way to do this through the graphical user interface would allow a user to quickly see which messages are specifically meant for him or her.

Some kind of simple tags that could be added to messages were suggested. These could be used to describe both the contents in maybe a few words and to select one or several users a specific message is intended for. It was however questioned, if even using short tags would be convenient enough to be used widely. One user suggested using speech recognition to extract keywords from messages.

The option to fast forward through messages, jumping over silent parts of messages and speeding up the playback a bit, was suggested by one user. Another user wondered if messages could be automatically grouped in some way. For example, several messages sent around the same time could be grouped as one discussion, although this might not always be the case. A case of two users sending messages in turns could, however, quite likely be considered one discussion. In case a user wants to listen to a message already played again, it was suggested that messages could be kept in memory on the user's device, to avoid loading the same messages again and thus speed up the process.

Sharing binaural recordings

Two of the test users use Facebook, but neither one tried posting binaural recordings on Facebook. All users nevertheless thought the idea was interesting, as long as there is something interesting to share, like a recording from a concert, for instance. One of the users considered the idea of posting a recording of speech, and thought this would feel a bit silly. It was pointed out that almost all social media nowadays uses images and text, and the possibility to share audio has not really existed.

Would you use the application if it were available for your mobile phone?

All of the users would consider trying the application if it were available for their mobile phones. One user speculated that if one carries a mobile phone all the time but does not use it that much, one would not necessarily use this kind of application either. He thought that this application might not give good enough a reason to carry around headphones, if they are not needed for other purposes. People who normally carry a headset for hands-free phone calls might, however, be much more likely to use this kind of application.

One of the users said that he probably would use this kind of application, if it also were used by some of his friends. He speculated that it might be quite difficult for the application to gain popularity, especially as it does not necessarily contain anything revolutionary tempting people to use it. He thought that the application could be much more tempting if it were part of another application or provided additional functionality to another application. Text-to-speech synthesis of IRC or IM messages was suggested, allowing one to follow a discussion when riding a bike, for example.

6.3 Discussion

The spatial separation of messages from different groups was seen as useful in some cases, but this method is severely limited by problems in separating the different directions as the number of groups grows. With as little as three groups, one user reported problems separating the different directions. Spatial separation should thus not be used at least as the only method for identifying messages from different groups. Another problem with the tested method is that joining a new group might change the directions associated with any old groups. This may lead to confusion, if the user is accustomed to the directions of the old groups. Allowing the user to choose the directions of the groups him- or herself might help, but the arc the directions are placed on still gets more crowded for every group added.

Other methods which help to identify the group a message belongs to using sound could be added as alternatives to spatial separation, or perhaps completely replace this alternative. Playing either signature tunes or sounds the user can choose, or playing synthesized speech of the group's name before each message, could provide the means to identify the different groups without the risk of confusion. In some cases users might not consider any of these methods for identifying groups necessary.

The users did not have any strong preferences for any of the methods used for spatial sound rendering. Linear amplitude panning was deemed the least pleasant by one user, exhibiting lack of externalization. Another user described the amplitude panning method using ILDs extracted from HRTFs as slightly more pleasant than using HRTFs. Interestingly, the users did hear some differences between this amplitude panning method and linear amplitude panning. Proper listening tests would probably give better results regarding the differences between the different methods, but the comments the users gave call into question the necessity of using HRTFs rather than a much simpler amplitude-panning method for this kind of application.

The test users had mixed opinions about adding reverberation to messages. One user thought the amount of reverberation added made the messages sound like they were recorded in a huge hall, while the other users thought the amount of reverberation was pleasant. One can question the need for adding reverberation to messages which have been recorded in a reverberant environment, since these already have an amount of reverberation present. However, if more effective methods for separating the speech of the user from sounds from the environment, including echoes and reverberation, were employed, there would probably be need for adding artificial reverberation. As one user suggested, this reverberation could be adapted to the surrounding environment of the listener, so that the speakers would sound like they were in the same room with the listener.

One of the biggest problems with the application was to quickly get information about the contents of the messages. Using short text tags to describe message content could help, but requires that the users find the time to tag the messages. Another problem was some confusion present when participating in multiple discussions at the same time. A solution to this could be the possibility to select an active group, so that only messages from this group would be played automatically, as suggested by one user. Messages from users in this group could then be spatially separated, to provide a more natural conversation experience.

One thing to look into when further developing this application would be better separation of the user's speech from background noise. In some cases the user might still want to record something else than noise, so the user should be able to decide if background sounds should be filtered out or not. It would also be beneficial to employ filtering of hiss, as there were quite high levels of hiss present in the messages.

Based on the comments of the test users, one may speculate that Mobile Augmented Messaging probably will not become a killer application of mobile augmented reality audio. As one user pointed out, the application does not contain anything revolutionary enough to make it a "must have". Still, it might gain popularity among some people as an alternative to, e.g., IRC or instant messaging, especially among people who normally wear headsets for extended periods of time. The possibility to publish binaural recordings on Facebook is a feature that can be used without knowing other users of the application and could thus help attract new users to the application. The recordings posted on Facebook would also act as advertisement for the application.

In addition to general chatting, e.g., between friends, there are specific scenarios where the application could be of use. One example given by one of the test users was to use it for meetings. Although the non-real-time communication with associated delays between message recording and delivery does not allow for the natural flow of conversation possible in full-duplex teleconferences or when meeting face to face, it gives the participants the possibility to say what they want at any time, without having to wait for their turn. As messages are automatically stored on a database server, it is possible to listen to the conversation later. Methods for adding metadata to organize messages would of course be required for efficient browsing of old conversations.

The application could also be good for communication of push-to-talk type, maybe as an alternative to Push to talk over Cellular (see [60]). Push-to-talk communication might be interesting, e.g., to taxi companies that want to communicate with their employees frequently [61]. Sending audio messages to taxi drivers instead of communicating with text would allow the drivers to keep their eyes on the road and could thus improve traffic safety.

7 Summary and future work

7.1 Summary

Asynchronous multi-user communication is commonly done using text. Sharing status updates on Facebook or Twitter are examples of this type of communication. When using a mobile device with a small touch screen or keypad, text input may, however, be slow and cumbersome. One solution to provide easier and faster message input is to use audio instead of text. Using audio leaves the user's hands free for other tasks when recording messages, and does not require focusing attention on the display of the device either when recording or listening to messages.

Conventional headphones attenuate sounds from the environment to some degree, depending on their type. In some cases, this can be desirable, but if the headphones are used only for sporadic messaging, for instance, it is probably not. Binaural microphones can be used in a headset to pick up the sound at both ears. This microphone signal can then be mixed with other audio to be played, and passed through the headphones to the listener's ears, making the headset acoustically transparent. Using such a headset together with a mobile device thus allows for almost undisturbed perception of and interaction with the surrounding environment while communicating using audio messages.

The Mobile Augmented Messaging application was developed to test the concept of using audio rather than text input for asynchronous communication between members of groups. Messages containing speech or other sounds are recorded and stored on a database server. Other users can listen to these messages as they arrive, or browse and listen to old messages. This kind of application is ideally suited for a mobile platform with an acoustically transparent headset. The current evaluation of the application was, however, limited to three test users running the application on desktop and laptop computers. The users sent each other messages sporadically during a week, after which a one-hour intensive chat session was organized.

In the application, messages from different groups are heard to arrive from different azimuthal angles. This was found to be a convenient way to separate messages from different groups in some cases, but the ability to identify different angles was quickly reduced as the number of groups increased. The users did not have any strong preferences for either HRTFs or the amplitude panning methods used for spatial sound rendering. Amplitude panning methods, which are simpler to implement and require less processing power than using HRTFs, might thus be a better alternative for this kind of application.

There is an option to write messages as text and play these messages using text-to-speech synthesis. This was seen as a good alternative when recording messages is difficult or impossible, e.g., in noisy environments. The possibility to make binaural recordings of interesting sound environments and post these on Facebook was considered an interesting idea.

In addition to the more convenient sending and receiving of messages using speech rather than text, the test users pointed out the advantages of communicating using speech rather than text. Spoken messages are, e.g., much more expressive than written messages and the intentions of the speaker are more easily understood. When comparing the Mobile Augmented Messaging application with text chats, the users thought participating in several discussions at the same time somewhat difficult or confusing. Browsing through old messages was also seen as problematic, as the limited metadata available requires users to listen to the messages to know their contents.

Compared with synchronous real-time communication, asynchronous and non-real-time delivery of voice messages does not allow for a natural flow of conversation. This was, however, not seen as particularly problematic by the test users. In a teleconference, e.g., this might actually be an advantage, allowing all participants to say what they want without having to wait for their turn.

7.2 Future work

The next step in developing the Mobile Augmented Messaging application would be to test it on a mobile platform, ideally on mobile phones. On such a platform, the possibilities and advantages of audio-based messaging over text-based messaging would become clearer than when testing the application on desktop or laptop computers. To enable almost hands- and vision-free usage of the application, an audio-based user interface using speech recognition could be added.

The group of users testing the current implementation of the application was quite small, but still provided valuable comments and ideas. On some questions, e.g., adding reverberation to messages, there were mixed opinions. Controlled listening tests with a larger number of participants would give a better idea of how reverberation and spatial audio could and should be used in this kind of application. Testing the application with a larger number of users would also be beneficial, but requires that the users have something to talk about.

An important area for further development of the application would be to provide means for easier browsing of messages. The possibility to add short text tags to messages and to select specific users that a message is intended to would help users to find relevant messages, but this would require some extra effort from the users. Preferably, speech recognition could be used to extract keywords from the messages without requiring input from the users.

8 References

[1] Facebook. `http://www.facebook.com/`. Last accessed 3 November 2010.

[2] Myspace. `http://www.myspace.com/`. Last accessed 11 January 2011.

[3] Twitter. `http://twitter.com/`. Last accessed 10 December 2010.

[4] Matti Karjalainen. Kommunikaatioakustiikka. Report 7, Helsinki University of Technology, Department of Signal Processing and Acoustics, 2008.

[5] John C. Middlebrooks and David M. Green. Sound localization by human listeners. *Annual Review of Psychology*, 42(1):135–159, 1991.

[6] Lauri Savioja, Jyri Huopaniemi, Tapio Lokki, and Riitta Väänänen. Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, 47(9):675–705, 1999.

[7] Thomas D. Rossing, F. Richard Moore, and Paul A. Wheeler. *The Science of Sound*. Addison Wesley, third edition, 2002.

[8] Bill Gardner and Keith Martin. HRTF measurements of a KEMAR dummy-head microphone. `http://sound.media.mit.edu/resources/KEMAR.html`, 1994. Last accessed 15 July 2010.

[9] Francis Rumsey. *Spatial Audio*. Focal Press, 2001.

[10] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, second edition, 1997.

[11] Patrick Satarzadeh, V. Ralph Algazi, and Richard O. Duda. Physical and filter pinna models based on anthropometry. In *Audio Engineering Society Convention 122*, May 2007. Paper no. 7098.

[12] Søren H. Nielsen. Auditory distance perception in different rooms. *Journal of the Audio Engineering Society*, 41(10):755–770, 1993.

[13] Douglas S. Brungart and William M. Rabinowitz. Auditory localization of nearby sources. Head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3):1465–1479, 1999.

[14] Douglas S. Brungart. Near-field virtual audio displays. *Presence: Teleoperators and Virtual Environments*, 11(1):93–106, 2002.

[15] Jyri Huopaniemi and Matti Karjalainen. Review of digital filter design and implementation methods for 3-D sound. In *Audio Engineering Society Convention 102*, March 1997. Paper no. 4461.

[16] Jean-Marc Jot, Véronique Larcher, and Olivier Warusfel. Digital signal processing issues in the context of binaural and transaural stereophony. In *Audio Engineering Society Convention 98*, February 1995. Paper no. 3980.

[17] Mendel Kleiner, Bengt-Inge Dalenbäck, and Peter Svensson. Auralization – an overview. *Journal of the Audio Engineering Society*, 41(11):861–875, 1993.

[18] Naraji Sakamoto, Toshiyuki Gotoh, and Yoichi Kimura. On "out-of-head localization" in headphone listening. *Journal of the Audio Engineering Society*, 24(9):710–716, 1976.

[19] Kiyofumi Inanaga, Yuji Yamada, and Hiroshi Koizumi. Headphone system with out-of-head localization applying dynamic HRTF (Head-Related Transfer Function). In *Audio Engineering Society Convention 98*, February 1995. Paper no. 4011.

[20] Philip Mackensen. *Auditive Localization. Head movements, an additional cue in Localization.* PhD thesis, Technische Universität Berlin, 2004.

[21] Bernhard U. Seeber and Hugo Fastl. Subjective selection of non-individual head-related transfer functions. In *Proceedings of the 9th International Conference on Auditory Display*, pages 259–262, July 2003.

[22] Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123, 1993.

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

[24] Alan Kraemer. Two speakers are better than 5.1. *IEEE Spectrum*, 38(5):71–74, 2001.

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

[26] Miikka Tikander. Development and evaluation of augmented reality audio systems. Report 13, Helsinki University of Technology, Department of Signal Processing and Acoustics, 2009.

[27] Aki Härmä, Julia Jakka, Miikka Tikander, Matti Karjalainen, Tapio Lokki, Jarmo Hiipakka, and Gaëtan Lorho. Augmented reality audio for mobile and wearable appliances. *Journal of the Audio Engineering Society*, 52(6):618–639, 2004.

[28] Aki Härmä, Julia Jakka, Miikka Tikander, Matti Karjalainen, Tapio Lokki, Heli Nironen, and Sampo Vesa. Techniques and applications of wearable augmented reality audio. In *Audio Engineering Society Convention 114*, March 2003. Paper no. 5768.

[29] Tapio Lokki, Heli Nironen, Sampo Vesa, Lauri Savioja, Aki Härmä, and Matti Karjalainen. Application scenarios of wearable and mobile augmented reality audio. In *Audio Engineering Society Convention 116*, May 2004. Paper no. 6026.

[30] Ville Riikonen. User-related acoustics in a two-way augmented reality audio system. Master's thesis, Helsinki University of Technology, 2008.

[31] Jussi Rämö. Evaluation of an augmented reality audio headset and mixer. Master's thesis, Helsinki University of Technology, 2009.

[32] Hannes Gamper. Audio augmented reality in telecommunication. Master's thesis, Graz University of Technology, 2010.

[33] W. Todd Nelson, Robert S. Bolia, Mark A. Ericson, and Richard L. McKinley. Spatial audio displays for speech communications: A comparison of free field and virtual acoustic environments. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting*, pages 1202–1205, September–October 1999.

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

[35] Nikolaos Moustakas, Andreas Floros, and Nikolaos Kanellopoulos. Eidola: An interactive augmented reality audio-game prototype. In *Audio Engineering Society Convention 127*, October 2009. Paper no. 7872.

[36] Laura Seppänen. Development of an audible sticker application and a video-based tracking system. Master's thesis, Helsinki University of Technology, 2008.

[37] Telecommunication Standardization Sector of the International Telecommunication Union. Recommendation ITU-T P.57: Artificial ears, 2009.

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

[39] Carl Poldy. Headphones. In John Borwick, editor, *Loudspeaker and Headphone Handbook*. Focal Press, third edition, 2001.

[40] H. Gustav Mueller. CIC hearing aids: What is their impact on the occlusion effect? *The Hearing Journal*, 47(11):29–30,32,34–35, 1994.

[41] Jorge Mejia, Harvey Dillon, and Michael Fisher. Active cancellation of occlusion: An electronic vent for hearing aids and hearing protectors. *The Journal of the Acoustical Society of America*, 124(1):235–240, 2008.

[42] Henrik Møller, Clemen Boje Jensen, Dorte Hammershøi, and Michael Friis Sørensen. Design criteria for headphones. *Journal of the Audio Engineering Society*, 43(4):218–232, 1995.

[43] David Howard and Jamie Angus. *Acoustics and Psychoacoustics*. Focal Press, fourth edition, 2009.

[44] Gaëtan Lorho. Subjective evaluation of headphone target frequency responses. In *Audio Engineering Society Convention 126*, May 2009. Paper no. 7770.

[45] Jarkko Oikarinen and Darren Reed. Internet relay chat protocol. `http://tools.ietf.org/html/rfc1459`, May 1993. Internet Engineering Task Force Request for Comments 1459. Last accessed 25 January 2011.

[46] Free Skype calls and cheap calls to phones – Skype. `http://www.skype.com/`. Last accessed 26 January 2011.

[47] Virtual worlds, avatars, free 3D chat, online meetings – Second Life official site. `http://secondlife.com/`. Last accessed 14 September 2010.

[48] Andrew Lang and Jean-Claude Bradley. Chemistry in Second Life. *Chemistry Central Journal*, 3:14, 2009.

[49] Kokua and Imprudence blog. `http://blog.kokuaviewer.org/`. Last accessed 26 January 2011.

[50] Douglas S. Brungart and Brian D. Simpson. Distance-based speech segregation in near-field virtual audio displays. In *Proceedings of the 7th International Conference on Auditory Display*, pages 169–174, July–August 2001.

[51] MySQL: The world's most popular open source database. `http://www.mysql.com/`. Last accessed 4 August 2010.

[52] Qt: Cross-platform application and UI framework. `http://qt.nokia.com/`. Last accessed 4 August 2010.

[53] PortAudio: Portable cross-platform audio API. `http://www.portaudio.com/`. Last accessed 4 August 2010.

[54] Barry D. Van Veen and Kevin M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, 1988.

[55] eSpeak: Speech synthesizer. `http://espeak.sourceforge.net/`. Last accessed 9 September 2010.

[56] Douglas D. Rife and John Vanderkooy. Transfer-function measurement with maximum-length sequences. *Journal of the Audio Engineering Society*, 37(6):419–444, 1989.

[57] Sanjit K. Mitra. *Digital Signal Processing: A Computer-Based Approach.* McGraw-Hill, third edition, 2006.

[58] Audacity: Free audio editor and recorder. `http://audacity.sourceforge.net/`. Last accessed 13 December 2010.

[59] Graph API reference – Facebook developers. `http://developers.facebook.com/docs/reference/api/`. Last accessed 18 January 2011.

[60] Open Mobile Alliance. Push to talk over cellular (PoC) – architecture. Candidate version 2.0. `http://www.openmobilealliance.org/technical/release_program/docs/poc/v2_0-20080421-c/oma-ad-poc-v2_0-20080226-c.pdf`, February 2008. Last accessed 18 January 2011.

[61] Timo Uhlmann. Push to talk. In *Multimediale Datenübertragung – Hauptseminar im Sommersemester 2008*, pages 57–64. Universität Ulm, Fakultät für Ingenieurwissenschaften und Informatik, Institut für Verteilte Systeme, 2008.