Title: Auditory-induced presence in mediated environments and related technology

Authors: Pontus Larsson\(^1\), Aleksander Väjamäe\(^2\), Daniel Västfjäll\(^1,3\) and Mendel Kleiner\(^1\)

RUNNING HEAD: AUDITORY-INDUCED PRESENCE

Affiliations

\(^1\)Department of Applied Acoustics, Chalmers University of Technology, Sweden
\(\text{pontus.larsson@ta.chalmers.se, daniel@ta.chalmers.se, mk@ta.chalmers.se}\)
\(^2\)Department of Signals and Systems, Chalmers University of Technology, Sweden
\(\text{aleksander.valjamae@s2.chalmers.se}\)
\(^3\)Department of Psychology, Göteborg University, Göteborg, Sweden

Draft version 050523

Chapter submitted to Handbook of Presence

Keywords: auditory presence, virtual acoustics, audio technology, sound representation, wave field synthesis, binaural synthesis, auralization, head-related transfer functions, ambisonics, bone conducted sound, auditory telepresence

1 Auditory induced presence .............................................................................................................. 2
1.1 Auditory presence ...................................................................................................................... 2
1.2 Presence and a dual-process model .......................................................................................... 3

2 Presence in auditory-visual virtual environments: previous findings .............................................. 5

3 Effects of levels of sound representation and technology on presence .............................................. 7
3.1 Multichannel reproduction ........................................................................................................ 7
3.2 Headphone reproduction ........................................................................................................... 8
3.3 Bone conducted sound and augmented reality applications ...................................................... 8
3.4 Auditory telepresence applications .......................................................................................... 9
3.5 Virtual acoustics synthesis and optimization ............................................................................. 9

4 Presence oriented design in spatial sound reproduction ................................................................. 10

5 General conclusions ....................................................................................................................... 13

6 Acknowledgements ....................................................................................................................... 13

7 References ...................................................................................................................................... 13

IN REVIEW, PLEASE DO NOT CITE WITHOUT THE AUTHOR'S PERMISSION
If realistic auditory displays can be developed, they may be a cheap solution to the problem of presence. However, additional research is needed to determine in what detail and with what fidelity the auditory background needs to be synthesized to create a realistic virtual sound field and an adequate sense of presence.

Gilkey and Weisenberger, 1995

Although there is an abundance of studies indicating correlations between various immersion components and presence, little of this research has been devoted to investigating the sound-related determinants of presence. Many presence researchers do however stress the importance of including also non-visual displays (Ijsselsteijn 2004, Slater 2003, Pope & Chalmers 1999, Gilkey & Weisenberger 1995) after all, most people experience the world not only through their eyes. Moreover, vision is clearly limited in the way that only a fraction of an environment is perceivable at a time, a fact which is believed to decrease overall realism (Pope & Chalmers, 1999). With audition, we can sense both direct sound and reflections from all directions in space without turning our heads, which in turn enables us to get an impression of geometry and size of the (virtual) environment (Pope & Chalmers, 1999). In fact, it has even been proposed that auditory input is crucial to achieving a full sense of presence, given that auditory perception is not commonly “turned off” in the same way as we regularly block visual percepts by shutting our eyes (Gilkey & Weisenberger, 1995). As Gilkey and Weisenberger suggest, auditory displays, when properly designed, are likely to be a cost-efficient solution to high presence virtual displays.

This chapter and two other companion chapters (Väljamäe et al., 2005a, Väljamäe et al., 2005b) in this book represent our vision on why sound is important in the mediated environments. For convenience of the reader we have divided our discussion into three parts where the first part is dedicated to sound perception issues (Väljamäe et al., 2005a), the second part gives an overview the audio-visual interaction and synergy effects (Väljamäe et al., 2005b), and third, present part is concerned with auditory-induced presence. The figure below represents the common idea of these three parts: how to optimize current immersion technologies in order to maximize user’s overall presence in rendered VEs.

In the present chapter we discuss auditory psychological and technological factors involved in forming presence experiences. In the figure below, this is represented by the leftmost block (VE rendering) and the shaded box in the right part of the figure (immersion response – auditory induced presence). It is hoped that by shedding light on the relationship between external auditory cues, the sensory processing, and related potential presence responses, the role of audition in VE’s will be clarified.

Figure 1. Optimization of VE rendering: Identification of auditory cues maximizing presence response.

1 Auditory induced presence

1.1 Auditory presence

The term presence has been used in many different contexts and there is still need for the clarification of this term (Slater, 2003). We would like to examine the question of auditory-induced presence from Slater’s perspective of perceptually-induced presence with sensory modalities contributing to overall presence responses. So, rather that talking about modality specific presence categories such as visual or auditory-induced presence,
we suggest that the term “auditory presence” should be used in the situations where sound is likely to have a profound positive or negative impact on the overall presence response (i.e. auditory-induced presence). For example, in a situation when a person is listening to music or a radio drama with either open or closed eyes, we can talk about different levels of the auditory information’s impact on the overall presence (Slater, 2003). The impact of a certain sensory modality on end-user presence might also be determined by the cognitive style of this person, i.e. to which extent the visual modality is dominating over other senses (Slater et al., 1994; on cognitive styles see Bertini et al., 1986).

In order to define situations where one can refer to auditory presence we should re-iterate some of the conclusions from the companion chapters on the sound perception (Väljamäe et al., 2005a) and cross-modal effects (Väljamäe et al., 2005b). First, the visual system provides us with only a fraction of an environment perceived a time (assuming also proper light conditions) while with audition we can sense both direct sound and reflections from all directions in space without turning our heads, which in turn enables us to get an impression of geometry and size of the (virtual) environment (Pope & Chalmers, 1999). Second, temporal resolution is much higher in the auditory domain and it is likely that rhythm information across all modalities is encoded and memorized based on “auditory code” (Gutman et al. 2005, see companion chapter by Väljamäe et al. (2005b) for further details). These properties of the auditory system made it ideally suitable for taking a role of the warning system guiding our visual attention as proposed by Popper and Fay (1995). Research on attention bring further evidence for this statement showing that auditory attention cues enhance the visual localization but the reverse is not true (Spence & Driver, 1997). Attributing warning system functions to the auditory system might be one of the factors explaining the strong tendency for association of acoustic attributes of heard sounds to more ecologically meaningful objects and events (see Väljamäe et al. (2005a) and references therein). One could see the auditory domain as a “medium of suggestion” (Cavalcanti, 1939) and a further discussion on affective and mood induction properties of sound in Väljamäe et al., (2005a) gives support to this notion.

The specific factors of auditory perception described above can have a direct implication for understanding situations which can be referred to as auditory presence. We can see this relation by examining the so called “dual-process model”, which can represent the interaction between low- and high-order auditory processing (Chaiken & Trope, 1999; Västfjäll & Jekosch, 2005) or similar processes of multimodal perception in mediated environments.

1.2 Presence and a dual-process model

A central assumption of many theories of evaluation is that perception and judgment (of for instance sound objects) are produced by a complex interaction of “automatic” and “controlled” processes (Västfjäll & Jekosch, 2005). Controlled responses are effortful, flexible, intentional and conscious, while automatic responses are effortless, direct, impossible to steer, and unconscious (Lieberman et al., 2002). We argue that it is beneficial to adopt such a “dual-processing model” also with respect to sound or multimodal perception. More specifically, such a view might help to understand how high-order factors (i.e. cognitive factors) influence low-order perception. Following the terminology used by (Lieberman et al., 2002) we suggest that both reflexive and reflective processes serve as input to the human perceptual system. We refer to these systems as the X-system (the x in reflexive) and the C-system (the c in reflective). These systems have different functions, are instantiated in different parts of the brain, and are associated with different experiences. We will in the following briefly outline the main characteristics of each of these systems.

The X-system. The X-system is a parallel-processing system that monitors changes in the environment. In our definition, the X-system is a “rapid pre-processor” of any type of perceptual stimulation. The main function of the X-system is to continuously monitor sensory input and detect changes in the input (LeDoux, 1999). The X-system is thus concerned with receiving information (low-level sensory perception) and sort that according to different automatized criteria. These criteria are often linked to an initial gross categorization that is important for survival. For that reason, the initial input must be fast and cannot afford to activate thinking and serial processing (reasoning). In earlier research, we have identified three such general criteria: Cognition-driven (e.g. receiving and sorting certain sensory information), experience-driven (e.g. identification based on previous exposure), and affect-driven (physiological responses activating different degrees of mobilization of the individual). If a change in the input stimulation causes a change in any (or several) of these criteria, the main function of the X-system is to inform the C-system (thus alarming the individual that this is a significant piece of information that needs to be attended: the alarm function of the X-system).

The C-system. The C-system is a serial processing system that translates the input from the X-system to interpretable experience that guides action, behavior, and communication. In our definition, information from the
X-system is interpreted into a subjective reaction/inference (“I feel angry”, “I like this”, “the sound makes me feel upset”, “it is a truck”, “it is loud” etc). We term this transformation of information into conscious awareness or salience (“object-of-perception”). Note that at the object-of-perception stage, the C-system only enabled the listener to make a reason-based inference about incoming information that results in a subjective experience/reflection. 2) Second, the object-of-perception can serve as a further input to a deliberative, comparison between current perception (“object of perception”) and the ideal perception (object of expectation). It is only at this stage the final perceptual judgment will be evident.

The joint effect of X and C processing. The main aspect of the X- and C-systems is that they are working in parallel and that both are needed to produce a global evaluation. Figure 2 is a schematic description of how two systems interact and guide judgment, action, and behavior.

From the perspective of auditory presence, the X and C framework suggests that auditory information that doesn’t belong to a virtual scene will activate the alarm function of the X-system with consequences for the experience of presence. Direct evidence for this comes from a study on “breaks in presence” (Slater & Steed, 2000; Brogni et al., 2004). Brogni et al. reported that participants that were more aware of backgrounds sounds belonging to the research laboratory (rather than the rendered VE) reported more breaks in presence.

Another type of event (e2) may cause a different reaction and corresponding presence response. If the X-system reacts well above a threshold of change detection an affectively-driven response may cause a direct link between the x-system and behavior (arrow D). This is particularly true for auditory presence taking to account sound’s affective and suggestive features (see companion chapter by Väljamäe et al., 2005a). In this case, the C-system may come into play later in order to access to what extent the initial, direct response was valid (e.g. initial startle response to a loud noise, followed by a visual search. For example, if the sound evoked visual search ends up in identifying the loudspeaker the C-system will override the response; if the visual search identifies a lion the C-
system may amplify the direct response (arrow D). If one can talk about presence in cinema, such an experience is more likely a case of auditory presence rather than visually induced presence, since visually induced presence can be rapidly overridden by awareness of cinema screen, surroundings and other people. The effect of suggestive and affective sound is masterfully used in the beginning of Michel Moore’s film “Fahrenheit 9/11” (Moore, 2004), where the dramatic episode of terrorists’ attack on New York was rendered in the radiodrama fashion, using spatial sound only.

A third type of interaction between the two systems is when the C-system accesses the X-system (e and line E). This may for instance be situations where aspects of the environment is not matching the expectations (i.e. enter a virtual room that looks very spacious, but sounds very small). For example, Larsson et al. (2004) showed that ratings of presence were significantly higher when the soundscape contained realistic object (buses driving, dogs barking) then when the objects were perceived to be “out-of-place” or unrealistic (artificial sounds).

2 Presence in auditory-visual virtual environments: previous findings

Sound has received comparably little attention in presence research. To further examine the role of audition in VE’s, Gilkey and Weisenberger (1995) compared the sensation of sudden deafness with the sensation of using no-sound VEs. When analyzing the work of Ramsdell (Ramsdell 1978, in Gilkey and Weisenberger 1995), who interviewed veterans returning from World War II with profound hearing loss, they found terms and expressions similar to those used when describing sensations of VE’s.

One of the most striking features of Ramsdell’s article was that the deaf observers felt as if the world was “dead”, lacking movement, and that “the world had taken on a strange and unreal quality” (Gilkey & Weisenberger, p. 358). Such sensations were accounted for the lack of “the auditory background of everyday life”, sounds of clocks ticking, footsteps, running water and other sounds that we do not typically maintain conscious awareness of and that are not required for adequate communication and alertness of warning events (Ramsdell, 1978; Gilkey & Weisenberger, 1995). Drawing on these tenets, Murray et al. (2000) carried out a series of experiments where participants were fitted with earplugs and instructed to carry out some everyday tasks during twenty minutes. Afterwards, the participants were requested to account for their experience and to complete a questionnaire comprising presence-related items. Overall, support was found for the notion of background sounds being important for the sensation of “being part of the environment”, termed by Murray et al. (2000) as environmentally anchored presence. Paradoxically, the study also suggested that due to the lack of auditory cues, the participants also had an increased attention to what was happening, they felt as being more aware of the situation than normally. Finally, the use of the earplugs also resulted in a situation where participants had a heightened awareness of self, in that they could better hear their own bodily sounds and which in turn contributed to the sensation of unconnectedness to the surround (i.e. less environmentally anchored presence). In addition to stressing the importance of the auditory background in VE, Murray et al.’s study shows that the auditory self-representation can be detrimental to an overall sense of presence. This suggest that the use of closed- or insert headphones, which in principle lead to a heightened auditory self-awareness in a similar manner as earplugs do, would not be appropriate for presenting VEs.

In a similar vein, Pörschmann (2001) considered the importance of adequate representation of one’s own voice in VEs. The VE system described by Pörschmann has the possibility of providing a natural sounding feedback of the user’s voice through headphones by compensating for the insertion loss of the headphones. Furthermore, the voice feedback as well as other sound sources may be auralized with room reflections. In his experiment, Pörschmann showed that participants’ ratings of presence increased significantly with the addition of room reflections and the compensation of the headphones’ insertion loss, while the latter had the strongest influence.

Regarding spatial properties of sound and presence in VEs, Hendrix and Barfield (1996) showed that sound compared to no-sound increased presence ratings but also that spatialized sound was favored in terms of presence compared to non-spatialized sound. The effect was however not as strong as first suspected. Hendrix and Barfield suggested that their use of non-individualized HRTFs may have prevented the audio from being externalized and thus less presence enhancing than expected. Some support for this explanation can be found in the study by Väljamäe et al. (2004), where indications of individualized HRTFs having a positive influence on the sensation of presence in auditory VE’s were found.

Proper externalization can however also be obtained by adding room acoustic cues (Begault et al. 2001). In a study by Larsson et al. (2005a), anechoic representations of auditory-only VEs were contrasted with VEs containing room acoustic cues. A significant increase in presence ratings was obtained for the room acoustic
cues conditions in this study, which was explained by the increased externalization. However, even though room acoustic cues are included, the reproduction technique seem to influences the sense of presence. In a between-groups study by Larsson et al. (2005b), stereo sound (with room acoustic cues) combined with a visual VE was contrasted to binaural sound (also with room acoustic cues) combined with the same visual VE. Here, it was shown that although room acoustic cues were present for both conditions, the binaural simulation yielded significantly higher presence ratings.

Taking a broader perspective, Ozawa et al (2003a) performed experiments with the aim to characterize the influence of sound quality, sound information and sound localization on users’ self ratings of presence. The sounds used in their study were mainly binaurally recorded ecological sounds, i.e. footsteps, vehicles, doors etc. In their study, Ozawa et. al., found that especially two factors, obtained through factor analysis of ratings of thirty-three sound quality items, had high positive correlation with sensed presence: “sound information” and “sound localization”. This implies that there are two important considerations when designing sound for VEs, one being that sounds should be informative and enable listeners to imagine the original (or intended) scene naturally, and the other being that sound sources should be well localizable by listeners.

As suggested in the research by Ozawa et al. (2003a), the spatial dimension of sound is not the only auditory determinant of presence. In support of this, Freeman et. al. (2001) found no significant effect of adding three extra channels of sound in their experiment using a audiovisual rally car sequence, which was partly explained by the fact that the program material did not may not capitalize on the spatial auditory cues provided by the additional channels. On the other hand, they found that enhancing the bass content and sound pressure level (SPL) increased presence ratings. In a similar vein, Ozawa and Miyasaka (2004) found that presence ratings increased with reproduced SPL for conditions without any visual stimulus, but that sensed presence in general was highest for realistic SPLs when visual stimulus was presented simultaneously with auditory stimuli. In audiovisual conditions with a “inside moving car” stimulus however, presence was highest for the highest SPL, which was explained by the fact that increased SPL may have compensated for the lack of vibrotactile stimulation. Study by (Sanders & Scorgie, 2002) compared conditions in virtual reality with no sound, surround sound, headphone reproduction1 and headphone plus low frequency sound reproduction via subwoofer. Both questionnaires and psychophysiological measures (temperature, galvanic skin response) were used to access affective responses and presence. All sound conditions significantly increased presence, but only surround sound resulted in significant changes in physiological response followed by a marginal trend for headphones plus subwoofer condition. Interestingly, questionnaires did not show such discrimination between sound reproduction techniques confirming unreliability of this measure when used alone.

Another, related line of research has been concerned with the design of the sound itself and its relation to presence (Serafin & Serafin 2004, Chueng & Mardsen 2002). Taking the approach of ecological perception, Chueng and Marsden (2002) proposed that expectation and discrimination are two possibly presence-related factors; expectation being the extent to which a person expects to hear a specific sound in a particular place and discrimination being the extent to which a sound will help to uniquely identify a particular place. The result from their studies suggested that what people expect to hear in certain real life situation can be significantly different from what they actually hear. Furthermore, when a certain type of expectation was generated by a visual stimulus, sound stimuli meeting this expectation induced a higher sense of presence as compared to when sound stimuli mismatched with expectations were presented along with the visual stimulus. These findings are especially interesting for the design of computationally efficient VEs, since they suggest that only those sounds that people expect to hear in a certain environment need to be rendered.

An often reoccurring theme in research related to presence induced by the auditory modality is that of congruency between the auditory and the visual display (Pörschmann 2001, Freeman 2001, Ozawa 2003b, Chueng & Mardsen 2002, Serafin & Serafin 2004). Consistency may be expressed in terms of the similarity between rendered visual and auditory spatial qualities (Pörschmann 2001), the methods of presentation of these qualities (Freeman et al. 2001), the degree of auditory-visual co-occurrence of events (Ozawa 2003b, Serafin 2004) and the expectation of auditory events given by the visual stimulus (Chueng & Mardsen, 2002). Ozawa et. al. (2003) conducted a study in which participants assessed their sense of presence obtained with binaural recordings and recorded video sequences presented on a 50-inch display. The results showed an interesting auditory visual integration effect; when the sound was matched with a visual sequence where the sound source was actually visible presence ratings were higher.

1 Though not clearly stated in the work, it is likely that stereo or 5 channel surround sound was converted to stereo sound was used
Larsson et al. (2005c) suggested that proper relations between auditory and visual spaciousness is needed to achieve a high sense of presence. In their experiment, a visual model was combined with two different acoustic models; one corresponding to the visual model and one of approximately half the size of the visual model. The models were represented by means of a CAVE-like virtual display and a multichannel sound system, and used in an experiment where participants rated their experience in terms of presence after performing a simple task in the VE. Although some indications were found supporting that the auditory-visualy matched condition was rated as being the most presence inducing one, the results were not as strong as predicted. An explanation to these findings, suggested by Larsson et al., was that, as visual distances and sizes often are underestimated in VEs (Knapp & Loomis, 2004. Thompson, 2004), it was likely that neither the “proper” sized acoustic model, nor the “wrong” sized acoustic model corresponded to the visual model from a perceptual point of view. Thus, a better understanding of how visual “spaciousness” or room size is perceived would be needed to perform further studies on this topic.

It has been suggested that also the degree of consistency within modalities would affect presence (Slater 2003). In the auditory domain, an example of an inconsistent stimulus could be a combination of sounds normally associated with different contexts (e.g. typical outdoor sounds combined with indoor sounds). Another type of within-modality inconsistency could be produced by spatializing a sound with a motion trajectory not related to that particular sound. In our example with the passing car above, this situation could occur e.g. if the sound of a car driving at slow speed is convolved with an HRTF trajectory corresponding to a high-speed passage. However, although these types of inconsistencies intuitively seem to be detrimental for the presence sensation, there is, to the best of the author's knowledge, little evidence to support this notion.

In sum, we see that a general understanding of various auditory display factors' contribution to the sense of presence begin to emerge. It is however clear that the findings presented above need further corroboration with different content and methodologies.

3 Effects of levels of sound representation and technology on presence

Since the invention of the phonograph, made by Thomas A. Edison in 1877, sound recording and reproduction techniques have been continuously evolving. Preserving the spatial characteristics of the recorded sound environment has always been an important topic of research, a work which started already at 1930s with the first stereo systems made commercially available in 1960s. The aim of spatial sound rendering is to create an impression of a sound environment surrounding a listener in the 3D space; thus simulating auditory reality. This goal has been assigned many different terms including auralization, spatialized sound/3-D sound, and virtual acoustics, which have been used interchangeably in the literature to refer to the creation of virtual listening experiences. For example, the term auralization was coined by Kleiner et. al. (1993) and is defined as “... the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such way as to simulate the binaural listening experience at a given position in the modeled space”.

One common criterion for technological systems delivering spatial audio is the level of perceptual accuracy of the rendered auditory environment, which may be very different depending on application needs (e.g. VR simulator or videoconferencing). Apart from the qualitative measure of perceptual accuracy, spatial audio systems can be divided into head-related and soundfield-related methods. Interested readers can find detailed technical information on these two approaches in books by Begault (1994) and Rumsey (2001) respectively (see also recent review by Shilling & Shinn-Cunningham, 2002). The following sections will briefly describe current state-of-the-art technology for reproduction of spatial audio and give different methods of analysis from the perspective of presence delivery.

3.1 Multichannel reproduction

Soundfield-related or multichannel audio reproduction systems can give a natural spatial impression over a certain listening area – the so called sweet spot area. The size of this sweet spot area mainly depends on the number of audio channels used. At present time, 5-channel, often referred to as surround sound systems, have become a part of many audio-visual technologies and standards in digital broadcasting and in the cinema domain. The next generations of multichannel audio rendering systems are likely to have a larger number of channels, as it used in the technologies providing better spatial rendering such as 10.2-channel (Zimmermann et al., 2004), Vector Based Amplitude Panning - VBAP (Pulkki, 1997), Ambisonic (Gerzon, 1985), or Wave Field Synthesis - WFS (Berkhout, 1988, Horbach et al., 2002). WFS can create a correct spatial impression over an entire listening area using large loudspeaker arrays surrounding it (typically >100 channels). However, direct
recording/transmission of spatial audio using WFS principles is difficult and requires novel multichannel audio compression methods (see Väljamäe, 2003 for review). Currently the WFS concept has been coupled with object-based rendering principles, where the desired soundfield is synthesized at the receiver side from separate signal inputs representing the sound objects and data representing room acoustics (Horbach & Boone, 1999).

3.2 Headphone reproduction

Head-related audio reproduction systems, also referenced as binaural or 3D-audio systems, are based on special pre-filtering of sound signals imitating mainly the outer ears (the pinnae) effects. Pre-measured catalogues of Head-Related Transfer Functions (HRTFs) are used for binaural sound synthesis, where a non-spatialized (“dry”) sound is convolved with transfer functions corresponding to the desired spatial position of the source. Currently most of 3D rendering systems use generic HRTFs catalogues due to the lengthy procedure of recording listeners’ own HRTF, however individualized catalogues are proven to enhance presence (Väljamäe et al., 2004).

When generic HRTFs are used the most common problem is in-head localization (IHL), where sound sources are not externalized but are rather perceived as being inside the listener’s head (Blauert, 1997). Another known artefact is a high rate of reversals in perception of spatial positions of the virtual sources, where binaural localization cues are ambiguous (cone of confusion), e.g. front-back confusion (Begault, 1994). Errors in elevation judgments can be also observed for stimuli processed with non-individualized HRTFs (Wenzel et al., 1993). These problems are believed to be reduced when head-tracking and individualized HRTFs are used (Blauert, 1997). At the current time it is popular to use anthropometric data (pinnae and head measurements) for choosing “personalized” HRTFs from a database containing HRTFs catalogues from several individuals (Zotkin, 2004). However, as auditory system express profound plasticity in spatial localization domain, a person can adapt to localization with some generic HRTFs catalogue. One could see this processes as re-learning to hear with modified pinnae as was shown in (Hofman et al., 1998). This natural ability to adapt for new HRTFs catalogues might be used when specifically modified, “supernormal” as termed by Durlach et al. (1993), transfer functions are introduced in order to enhance localization performance, e.g. reduce the front-back confusions (Gupta et al., 2002).

Generally, binaural systems are used for sound reproduction over headphones, which this option very attractive for wearable augmented reality applications (see the excellent review by Härmä et al., 2004). However, binaural sound can be also reproduced by a pair of loudspeakers if additional processing is applied (cross-talk cancellation technique), which sometimes is used in teleconferencing (Evans et al., 1997). 3D-audio systems are designed for creating a spatial audio impression for a single listener, which can be disadvantageous for applications where several users are sharing same auralized space and use it for communication, such as in CAVE simulations.

3.3 Bone conducted sound and augmented reality applications

Mixed and Augmented Reality (MR and AR) technologies aim for enhancement of the surrounding environment for assistance, entertainment, or artistic purposes. Auditory AR, where virtual sound objects are inserted into real sound environment using binaural sound synthesis methods (Härmä et al., 2004), is a first step towards future multi-modal MR and AR applications. In a proposed solution for auditory space augmentation (Härmä et al., 2004), surrounding sounds are picked up by microphones inserted in headphones so that both real and virtual parts undergo signal processing. However, an alternative way of auditory AR could be thought of, where unaltered sounds from the normal sound conduction path are combined with bone conducted sound.

Bone conducted (BC) sound is elicited by human head vibrations which are transmitted to the cochlea through the skull bones (Tomndorf, 1972; for recent on BC review see Stenfelt & Good, 2005). Most people experience BC sound in everyday life; approximately 50% of the sound energy when hearing one’s own voice is transmitted through bone conduction (Pörschmann, 2000). It is important to note that spatial sound reproduction is also feasible via BC sound if bilateral stimulation via two head vibrators is applied (Snik et al. 2004).

---

2 Reflection from head, shoulders and torso also influence the HRTF shape (Begault, 1994)
Currently, the largest area of BC sound applications is hearing acuity testing and hearing aids, where the transducer is either implanted or pressed against the skull. Recently, BC sound has proven interesting for communication systems as it facilitates the use of open ear canals. Such headphone free communication not only allows for perception of original surrounding environment but also ideally suits for MR and AR applications. We envision several areas where BC conducted might be successfully applied (Väljamäe et al. 2005d). First, additional sound rendering with open ear canals is ideally suited for speech translation purposes, where original language can be accompanied with synchronized speech of the translator. Another option is a scenario of interactive augmented auditory reality, where a voice rendered via BC sound is telling a story and the real sonic environment plays the role of a background. As a user can have several sensors providing the information about environment like in the “sonic city” application by Gaye et al. (2003), the BC sound based narration can be dynamically changing to fit the user’s environment. Finally, a combination of loudspeaker reproduction with bone conducted sound can improve externalization of spatial sound image as BC sound localized as been close or inside the head provides a good reference point (Väljamäe et al. 2005d).

3.4 Auditory telepresence applications

As was described in section 3.1, spatial sound transmission for multichannel reproduction is a challenging task both for transmission bandwidth and for recording techniques. For example, there is no agreement of microphones setup for 5-channel sound recording and sound engineers often use their specific knows-know for creating an optimal spatial image (Rumsey, 2001). In the case of head-related reproduction systems one could use an artificial head with ear-inserted microphones positioned at the receiver position. In fact, the very first transmissions using this technique were conducted in Bells Labs as early as 1933 (Fletcher, 1993).

At the current time several techniques start to emerge for binaural auditory telepresence applications where the user, at a remote location, can operate the sound recording/transmission device (Algazi et al., 2004; Toshima et al. 2004). When envisioning future auditory telepresence applications based on the binaural sound reproduction, an interesting presence-related research question is a level of adaptation to audio-visual temporal or spatial disparity. For example in the situation, where a user records the binaural sound in a specific location and plays it back several minutes later while been in the same location, what impact such auditory augmented reality scenario can have on our perception of time? In the case of spatial disparity, people can exchange the binaural sound of each other (Härmä et al., 2004) or can have an aural tour while sitting in the bus for the sightseeing, sequentially visiting different visually distant locations equipped with binaural sound recording/transmission capabilities.

3.5 Virtual acoustics synthesis and optimization

The idea of capturing or modelling desired room acoustics data is very important for spatial audio in immersive VR applications, where adequate real-time changes of the auditory environment reflecting a listener or sound objects movement can be crucial for the sense of presence (Durlach & Mavor, 1995). With the development of software implementations of different geometrical acoustics algorithms, high accuracy (often off-line) computer-aided room acoustic prediction and auralization have become important tools for acousticians (CATT, ODEON). Acoustic prediction algorithms simulating early sound reflections (<100 ms) and diffuse reverberation field in real-time are often denoted as Virtual Acoustics (see for review Funkhouser et al., 2004). Two different strategies for this real-time auralization are usually employed 1) the physics-based approach, where geometrical acoustics techniques are adapted to meet low input-output latency requirements (e.g. Savioja et al., 1999) or 2) the perceptual approach, where attributes of created listener impression, such as room envelopment, are of main concern (e.g. Jot & Warusfel, 1995). A Second, problem is to how accomplish optimal listening conditions for multiple listeners as optimization today only can be performed for a single sweet spot area.

Room acoustic properties are very important for speech intelligibility and in real environments there is a often clear separation of acoustic optimization for speech or music. However, in VE’s one is not restricted by architectural constrains so communication-optimized and time-varying room acoustics can be a part of technologies for smart virtual or augmented surroundings (Nijholt et al., 2004). For example, similar to the “acoustic zoom-in” (where a directional microphone is used to capture a speaker voice in the surrounding noise) one might envision a sort of auditory zoom-in in the case of virtual auditory scene rendering. In the case of such auditory zoom-in, the target (e.g. speaker) can be amplified and allocated a specific frequency range assuring its intelligibility and spatial position. Knowledge on spatial auditory processing has been successfully implemented when using binaural synthesis in videoconferencing, where end-user could safely allocate the conversation partners in auditory space without loosing their speech intelligibility (Faller & Baumgarte, 2003). It should also be noted that the affective side of room acoustics (Howard & Angus, 2001; Västfjäll et al., 2002) is largely unexplored in the areas of teleconferencing and communication in VR. For example, sound aspects of
visual environment; transitions between different acoustic environments; situations where acoustic cues provide
movement in VE. Most critical situations include strong incongruence between room acoustics and presented
challenging task is to render perceptually relevant changes in the acoustic environment corresponding to user
Interaction - far field
While all spatial audio technologies can provide constant sound ambience, a more
critical situations include strong incongruence between room acoustics and presented
visual environment; transitions between different acoustic environments; situations where acoustic cues provide
spatial audio technologies and whole auditory scene rendering in real-time still requires dedicated
multi processor systems or distributed computing systems. The problem of efficiently simulating a large number
of sound sources (>60) in a scene together with the proper room acoustics simulation still remains the one big
unsolved problem for virtual acoustics applications (Tsingos et al., 2004). A common approach to adapt the
synthesis resources, is to employ automatic distance-culling, where more distant sound sources are not rendered.
However audible artifacts can occur when a large number of sound sources are close to the listener, especially
when these sources represent one large object, e.g. a car. Extending the ideology of the perceptual approach and
taking into account spatial masking effects of the auditory system (Faller & Baumgarte, 2003) only perceptually
relevant information can be rendered thus supporting a vast number of sources (Tsingos et al., 2004).

4 Presence oriented design in spatial sound reproduction

In this section we will restrict the discussion to the parameters of spatial sound which we think are likely to
contribute to auditory presence, although it is important to acknowledge that other parameters may also influence
overall presence. For example, it has been suggested that even non-spatialized mono sound consistent within an
expected scene should increase a sense of presence (Sanders & Scorgie, 2002; Lawson 1998). Furthermore, as
sound provides a dynamic cue, it may compensate for e.g. lack of movement in the visual domain (see Väljamäe
et al., 2005b for cross-modal compensation effects). Nonetheless, determining the parameters of spatial and
room acoustic simulation that are perceptually relevant for the sensation of presence in multi-modal VE an
important research question (Martens & Wosczynski, 2003). Following the approach of Martens and Wosczynski
(2003) and Serafin & Serafin (2004), we will below describe the parameters which might be important from the
perspective of auditory presence and which in turn can determine a type of technical implementation of virtual
acoustics.

**Spaciousness.** Spaciousness has since long been one of the defining perceptual attributes of concert halls and
other types of rooms for music. For VE applications involving simulations of enclosed spaces, the correct
representation of spaciousness is of course essential for the end-user’s experience. A “correct representation”
would mean that the virtual room’s reverberation envelops the listener in a similar manner as in real life.
Interesting to note is that adding reverberation also increases the user’s sense of externalization in binaural
systems, i.e. that the sound is perceived as coming from outside rather than from inside or close to the head
(Begault et al., 2001). As externalization in turn is believed to increase presence (Hendrix & Barfield, 1996),
including room reverberation seems important from the presence perspective. The potential benefits of adding
spacious room reverberation has unfortunately been largely overlooked in previous research on presence, apart
from the studies presented in Larsson et al. 2005a, Larsson et al. 2005b, and Larsson et al. 2005c. Recently
however, subjective evaluation methods of spatial audio have employed a new spaciousness attribute termed
“presence” similar to attributes used in VE presence research, defined as “…the sense of being inside an
(enclosed) space or scene” (Rumsey, 2001 and references therein), which promises future explorations of this
subject. This auditory presence attribute of spatial sound quality is also used for the recreated outdoor
environments and generally refers to the background ambient sound energy arriving from various directions. It
should be noted that low frequency sound strongly contributes to the ambience and in turn to presence as was
confirmed by psychophysiological measurements (Sanders & Scorgie, 2002).

**Interaction - near field.** Interaction is generally believed to be one of the key factors contributing to presence
(Regenbrecht & Schubert, 2002). Congruent cross-modal simulation of active exploration in close, peri-personal
space in VE imposes localization requirements for spatial sound rendering. Although cross-modal synergy
effects, described in companion chapter (Väljamäe, 2005b), might compensate for some spatial disparities across
sensory inputs, the intermodal time contiguity window requires their good synchronization. Confronted to
multichannel reproduction limitations, binaural reproduction provides a possibility to render accurate spatial
sound in users’ peripersonal space. Although HRTFs are typically recorded for sound reproduction in far-field
(>1-1.5 meters), interesting interpolation solution approximating the near-field HRTFs set has been recently
proposed (Duraiswaini et al., 2004).

**Interaction - far field.** While all spatial audio technologies can provide constant sound ambience, a more
challenging task is to render perceptually relevant changes in the acoustic environment corresponding to user
movement in VE. Most critical situations include strong incongruence between room acoustics and presented
visual environment; transitions between different acoustic environments; situations where acoustic cues provide
information about sound source or user position and/or physical properties of environment: being close to the wall, occlusion (a listener and a sound source are in different rooms), exclusion (only indirect sound is heard, e.g. through an open window), obstruction (an object is situated between a listener and a sound source).

**Scene consistency.** In our experiments on auditory-induced self-motion sensation (Larsson et al. 2004, Väljamäe et al. 2005c) we observed that inconsistencies in created auditory scene had a considerable impact on presence and self-motion ratings. These inconsistencies included both artifacts from using the generic HRTFs (wrong motion trajectories) and more high-order effects caused by ecological inconsistency of the sound environment, for example, strange combinations of naturalistic stimuli representing concrete sound objects (e.g. dog, bus) and artificial sound (modulated tone). Therefore one should be careful when creating a virtual sound environment where ecological consistency and efficient spatial sound localization have to be assured. It is likely that a combination of both ambient and clearly localizable sounds should result in auditory-induced presence. It should be also noted that virtual sound environment complexity might create a reverse effect, where too many sound effects will rather destroy a spatial sound image (see Murch (2000) for such effect in the cinema sound design).

When designing spatial sound rendering system for VR one primarily needs to consider: 1) if the system is intended for sound delivery for a single point of audition or if the listener should be able to change his listening position without applying headtracking and corresponding re-synthesis of sound scene (e.g. WFS with large listening area vs. Ambisonics with relatively sweet spot); 2) how large listening area is required; 3) complexity and cost of the system, 4) if there are any undesirable acoustic properties of the room in which the system should be used (e.g. room native reverberation, external noise, etc.); and 5) what type of visual rendering system (if any) is to be used with the sound system. A short comparison table of spatial audio systems regarding the first three points is given in (Funkhouser et al., 2004). An alternative comparison of audio systems for VR is presented in Table 1, where sound and visual rendering technologies are classified into wearable or stationary.
Table 1. Different combinations of visual and audio rendering technologies. Advantages and disadvantages for each type are marked as + (positive), - (negative), +/- (positive with minor implications), -/+ (negative with minor advantages).

<table>
<thead>
<tr>
<th>VISUAL RENDERING</th>
<th>AUDIO RENDERING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wearable (e.g. Head Mounted Display)</td>
<td>Wearable (headpones)</td>
</tr>
<tr>
<td>+ direct communication between users</td>
<td>-/+ indirect, microphone-based communication in the case of multiple users</td>
</tr>
<tr>
<td>+ shared VE: users with different visual viewpoints and points of audition</td>
<td>+ shared VE: users with different visual viewpoints and points of audition</td>
</tr>
<tr>
<td>+/- relatively high level of “mediation awareness” due to HMD</td>
<td>+/- relatively low level of “mediation awareness” due to headpones</td>
</tr>
<tr>
<td>- external sound (reverberation, noise) can influence the rendered audio scene, therefore specific acoustic requirements for the room are necessary</td>
<td>+ isolation from external sonic environment (but VR augmentation possibilities remain)</td>
</tr>
<tr>
<td>+ auditory mediation is completely invisible to users - large scale multichannel systems can be used without restrictions</td>
<td>+/- auditory mediation is invisible but headpones can contribute to “mediation awareness”</td>
</tr>
<tr>
<td>+/- area of physical user movement (actual locomotion) is limited by area surrounded loudspeakers (sweetspot area even smaller)</td>
<td>-/+ area of physical user movement (actual locomotion) is limited by the visual rendering of VE</td>
</tr>
<tr>
<td>Conclusion: not often used, but option of shared VE makes it very attractive, for example, for investigation in the field of social presence</td>
<td>Conclusion: very good combination –current state-of-the-art - if-appropriated binaural synthesis (headtracking, individualized HRTFs).</td>
</tr>
</tbody>
</table>

Stationary (projection screen(s)) | Stationary (loudspeakers) |

| + direct communication between users | +/- indirect, microphone-based communication in the case of multiple users |
| +/- shared VE: several users can share the same audio scene but one visual viewpoint | + shared VE: users with different visual viewpoints and points of audition |
| + low level of “mediation awareness” | +/- relatively low level of “mediation awareness” due to loudspeakers |
| - external sound (reverberation, noise) can influence the rendered audio scene, therefore specific acoustic requirements for the room are necessary | + isolation from external sonic environment (but VR augmentation possibilities remain) |
| - auditory mediation technology (loudspeakers) have to be hidden | +/- auditory mediation is invisible but headpones can contribute to “mediation awareness” |
| - area of physical user movement (actual locomotion) is limited both by visual and auditory rendering technologies | -/+ area of physical user movement (actual locomotion) is limited by the visual rendering of VE |
| Conclusion: this type is a classic example of current visually-oriented VE rendering, where typically several loudspeakers provide ambient sound background. | Conclusion: very good combination –current state-of-the-art - if-appropriated binaural synthesis (headtracking, individualized HRTFs). |

Generally, stationary visual setups impose restrictions on the number and configuration of loudspeaker used (e.g. it might be difficult to use large loudspeaker arrays), thus significantly reducing the spatial audio system’s capabilities. Headphone reproduction on the other hand clearly manifests a mediation device (i.e. the user knows for sure that the sound is delivered through the headphones). No direct comparison between effect of headphone and loudspeaker listening to spatial sound and presence has been carried out apart from Sanders and Scorgie (2002) study, who did not use proper binaural synthesis for headphone reproduction conditions and thus could not adequately access the importance of sound localization accuracy on auditory presence responses.
One would expect that the use of headphones (particularly the closed- or insert types), which in principle lead to a heightened auditory self-awareness in a similar manner as earplugs do, would have a negative influence on the sense of presence (Pörschmann, 2001, Murray et al., 2000, and see also Section 2). Binaural synthesis also can result in high auditory rendering update rates (this can be addressed by perceptually optimized auralization algorithms described in section 3.3) and lower sense of externalization. Positive features of headphone reproduction are that they may be used for wearable applications such as AR and that the acoustic properties of the listening room are less critical. Depending on the application, indirect, microphone-based communication between users sharing VE can be considered as advantageous or not.

5 General conclusions
From the previous sections one may draw the conclusion that the audio technology and design of auditory VE’s involves several degrees of freedom and that the system/design parameters in turn affect the sense of presence in a number of ways. A main challenge is therefore to seek the relation between system/design parameters and sense of presence, in order to provide better and more efficient VE’s, both in terms of end-user experience and computational costs. An ideal VE sound system should provide a maximum sense of presence at a minimal cost. While the present chapter provide some guidelines for development of such systems, exactly how the ideal system should be designed is as of yet far from fully understood and should be an issue for future research.

6 Acknowledgements
The work presented in this paper was supported by the EU FET Presence Research Initiative project POEMS (Perceptually Oriented Ego-Motion Simulation), IST-2001-39223 and the Swedish Science Council (VR). The second author of this paper would like to thank for the support from Alfred Ots’ scholarship foundation

7 References
Christensen, C. L. ODEON room acoustics software. [Computer software]. URL: http://www.dat.dtu.dk/odeon/


Stenfelt, S., & Goode, R.L. Bone conducted sound: physiological and clinical aspects. (Submitted 2005)


