I

Publication I

Pulkki, V. and Hirvonen, T., ”Localization of Virtual Sources in Multichannel Audio Reproduction”, IEEE Transactions on Speech and Audio Processing, Vol 13, No. 1, Jan 2005.

Localization of Virtual Sources in Multichannel Audio Reproduction

Ville Pulkki and Toni Hirvonen

***Abstract*—The localization of virtual sources generated with different two-dimensional (2-D) multichannel reproduction systems has been studied by means of auditory model simulations and listening tests. The reproduction was implemented with typical five- and eight-channel loudspeaker setups. The microphone systems used were first- and second-order Ambisonics as well as a spaced microphone technique. Pair-wise panning was also studied. The results show that the auditory model can be used in the prediction of perceived direction in multichannel sound reproduction near the median plane. Some systematic deviations between the model predictions and the listening test results were found farther from the median plane. The frequency-dependent capability to produce narrow-band virtual sources to targeted directions is reported for the studied systems.**

***Index Terms*—Audio systems, binaural auditory model, spatial sound reproduction quality.**

I. Introduction

THE TEMPORAL and spectral structure of a sound signal can be captured and reproduced accurately by using modern audio technology. In contrast, the reproduction of the spatial attributes of sound cannot be considered to be accurate in general. Here, spatial attributes denote the part of sound perception that depends on listening room acoustics and on the listening setup within one room. Some such attributes can be characterized as, e.g., direction and distance of sound source and strength of reverberation.

Two-channel stereophony [1] is the most commonly used spatial sound reproduction method. The listener perceives all auditory objects appearing on a line between the loudspeakers. The line can be thought of as an acoustical opening to the room where the recording was made. Using such a system a listener cannot have an equal perception of spatial sound as in the actual recording room. In the past ten years, a five-loudspeaker listening standard (5.1) [2] has become increasingly popular. The listener is surrounded by loudspeakers and more realistic spatial perceptions are assumed to be reproduced. There are also other standards for loudspeaker placement which utilize more loudspeakers around the listener. Some of these setups also have elevated loudspeakers.

However, there seems to be no decisive method to record spatial sound for multiloudspeaker systems with existing microphone types. Although a multitude of techniques have been suggested for specific loudspeaker systems, none of these techniques has been commonly recognized. Also, there is little knowledge on how different techniques reproduce different spatial attributes. For this reason, we decided to study how the directions of sound sources are reproduced using the combination of a specific microphone technique and a specific multichannel loudspeaker system. The localization of virtual sources produced with multichannel reproduction systems is evaluated using a binaural auditory model, and the results are verified through listening tests. In Section II the spatial hearing mechanisms of humans are discussed, and in Section III some multichannel sound reproduction techniques are covered. The binaural auditory model used in this study is described in Section IV. The model is applied to a number of multichannel reproduction systems in Section V. These simulation results are verified by means of listening tests, presented in Sections VI and VII. The validity of obtained simulation results is discussed in Section VIII and conclusions are drawn in Section IX.

II. Spatial Hearing

Spatial and directional hearing have been studied intensively (for overviews, see, e.g., [3]) or [4]. The duplex theory of sound localization states that the two main cues of sound source localization are the *interaural time difference* (ITD) and the *interaural level difference* (ILD) which are caused, respectively, by the wave propagation time difference (primarily below 1.5 kHz) and the shadowing effect by the head (primarily above 1.5 kHz). The auditory system decodes the cues in a frequency-dependent manner.

The main cues are used to resolve in which cone of confusion the sound source lies. A cone of confusion can be approximated by a cone having its axis of symmetry along a line passing through the listener's ears and having the apex at the center point between the ears. Direction perception within a cone of confusion is refined using other cues, such as spectral cues and the effect of head rotation to ITD and ILD. Spectral cues and head rotation are considered to mediate elevation and front-back information.

The precedence effect [3], [5] is an additional assisting mechanism of spatial hearing. It can be regarded as suppression of early delayed versions of the direct sound in source direction perception. This helps to perceive the sound source directions in reverberant conditions.

This study focuses on the perception of virtual sources. Both ITD and ILD of virtual sources may be inconsistent depending

Manuscript received September 1, 2003; revised July 9, 2004. This work was supported by the Academy of Finland under Grant 101339 and 201050. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Walter Kellermann.

The authors are with the Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Helsinki FIN-02015, Finland (e-mail: ville.pulkki@hut.fi).

Digital Object Identifier 10.1109/TSA.2004.838533

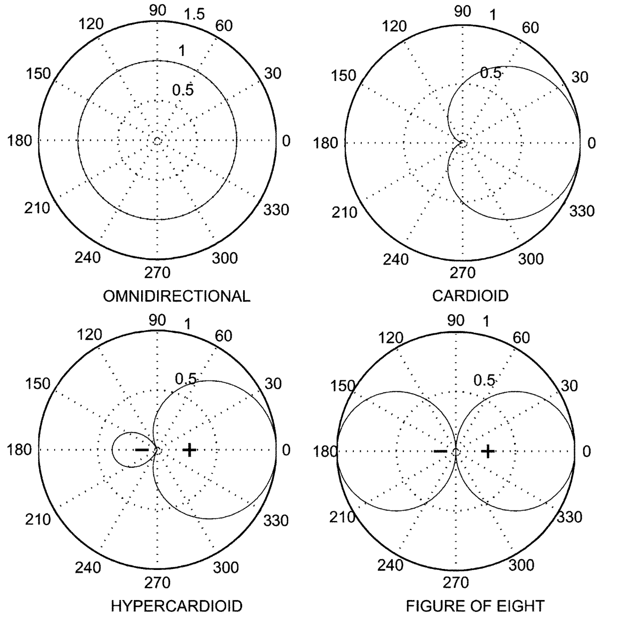

Fig. 1. Typical microphone polar patterns. The captured sound wave is weighted depending on polar pattern of the microphone, which is a function of direction.

on frequency. Here, a consistent cue denotes a cue that is produced by a real source in anechoic conditions. In order to investigate the cue relations when they suggest different directions, many experiments have employed conflicting ITDs and ILDs by using headphones. Some early studies on time-intensity trading emphasized the importance of the ITD cue, e.g., [6]. In the situation where two cues conflict, it has been shown that they interact in some degree. For example, an ITD cue suggesting direction slightly left and an ILD cue suggesting direction slightly right may produce perception of center direction [7]. However, the discrepant cues may produce two images. It has been shown that with sufficient training, listeners may perceive separate sound images based on both time and intensity disparities [7].

In modern studies it has been found that when both ITD and ILD are consistent, but indicate different real source directions, the low-frequency ITD cue dominates the localization [8]. In the case when one of the cues was set to be inconsistent, the consistent cue was more prominent [9]. The case in which both cues are set inconsistent has not been studied thoroughly. Amplitude-panned virtual sources produce ITD and ILD cues which are inconsistent depending on frequency [10]. In this particular case, it was found that the low-frequency ITD is the most salient cue if it is available. With high-frequency sounds the ILD cue was the most salient.

III. Multichannel Sound Reproduction Techniques

When recording sound, obviously some kind of microphone has to be used. The following sections discuss some commonly available microphone types. The vast frequency range perceived by humans makes it very difficult to produce microphones that would not only have high directional characteristics, but would also capture sound without prominent coloring. In practice, microphone polar patterns are of zeroth-order (omnidirectional), or of first-order (figure-of-eight, cardioid and hypercardioid), as shown in Fig. 1. An omnidirectional microphone captures sound from all directions with equal amplitude. The polar pattern $f(\theta)$ of a first-order microphone is defined as

$$f = p_1 + 2p_2\cos(\theta) \tag{1}$$

where θ is the space angle between the frontal axis of the microphone and the direction of the sound source. If $p_1 = 0$, and $p_2 \neq 0$, the polar pattern is a figure-of-eight. If $p_1 = 1$ and $p_2 = 1$, a hypercardioid is obtained, and if $p_1 = 1$ and $p_2 = 0.5$, the polar pattern is cardioid.

When sound is recorded for multichannel listening, several microphones are typically employed. Some common microphone layouts are as follows. A coincident technique, first used by Blumlein [1], refers to a microphone technique in which two or more directive microphones are placed as close as possible to each other. The resulting signals differ in amplitude. The phase difference between the microphones can be either 0° or 180°. A noncoincident technique, in turn refers to a setup in which the microphones are separated in space. This also produces time differences between loudspeaker signals. The directional patterns of the microphones may be of any form.

The microphone techniques can also be divided into methods where microphones are placed either close to the sound sources, or far-away from the sources. The latter technique is used to also capture the response of the room in which the sound sources lie. This method is commonly used in recording classical music. Typically, the sound sources are in front of the microphones and the response of the room comes from all directions. In proximity techniques, the sound signal is recorded so as to eliminate as much reverberated sound as possible. This monophonic signal is later applied to loudspeakers with an appropriate technique, such as amplitude panning (see Section III-D).

There are some standardized, or widely used multichannel loudspeaker setups. In the 1970s, a four-loudspeaker setup, which included loudspeakers in ±45° and ±135° directions, was introduced. However, it was never widely accepted. The most widely used multichannel loudspeaker system is the 5.1 loudspeaker configuration, which has loudspeakers in the directions ±110°, ±30°, and 0° [2]. It is widely used in cinemas and is gaining popularity in domestic use as well. Various setups having more than five loudspeakers have also been suggested, typically for cinema use. In computer music, a reproduction system which consists of six or eight loudspeakers evenly spaced around the listener, is often used.

When a sound source is reproduced to a listener with a microphone technique and a loudspeaker setup, the resulting sound image is referred to as a virtual source. With respect to the listener, a virtual source may appear as point-like or spread. If a realistic reproduction is desired, the perceived properties of the virtual source should be equal to the perception of the real source in the recording room. However, often realistic reproduction is not desired; e.g., virtual sources broader than in reality, may be reproduced.

Different microphone techniques have been developed to reproduce spatial sound over multiple loudspeakers [11]. Furthermore, there are different methods to spatialize monophonic sound signals for multichannel setups. Some of these techniques are presented below.

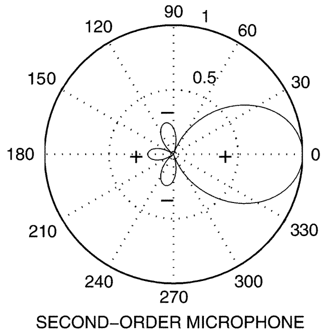

Fig. 2. Polar pattern of a second-order microphone.

A. Ambisonics

Ambisonics [12] is a microphone technique based on the use of the Soundfield microphone [13]. Typically, the output of the microphone consists of four audio signals recorded with different polar patterns that include an omnidirectional microphone and three figure-of-eight microphones placed along the three coordinate axes. In reproduction, the signals are matrixed so that the signal applied to each loudspeaker corresponds to a signal that could have been recorded with a hypercardioid or cardioid microphone facing the direction that corresponds to the direction of the loudspeaker in the listening room. Ambisonics is used often with four, six or eight loudspeakers symmetrically placed in the horizontal plane around the listener. This approach results in relatively broad polar patterns that create cross talk to loudspeaker signals. Basically, sound coming from one direction emanates from all loudspeakers in the listening phase. The directional quality of Ambisonics in a four-loudspeaker setup has been studied with broad-band speech [14].

A theory of second-order Ambisonics has been proposed [15]. The method is based on a hypothesized second-order microphone. The polar pattern of signals fed to loudspeakers would then have the form

$$f = \frac{1}{n}(p_1 + 2p_2\cos(\theta) + 2\mathrm{p}_3\cos(2\theta)). \quad (2)$$

One such polar pattern is plotted in Fig. 2. The pattern is considerably narrower than first-order patterns and results in less cross-talk between loudspeakers. Second-order Ambisonics has been researched mostly on a theoretical level [16]. Second- and higher-order microphone techniques have been used as panning methods by simulating corresponding microphones [15]. However, there have not been any psychoacoustical studies published on directional quality of virtual sources produced with second-order Ambisonics.

In principle, first- and second-order Ambisonics can be applied to any loudspeaker system. They are often used with a symmetric layout with four, six, or eight loudspeakers, but can also be applied to asymmetric layouts, e.g., to the 5.1 system.

B. Spaced Microphone Techniques

There exists a wide variety of spaced microphone systems for multichannel reproduction. Many of them have been designed for the 5.1 loudspeaker setup, as presented in [11]. In many cases, the microphones are in the configuration of a star, with each point facing approximately toward the corresponding loudspeaker direction. The distances between microphones vary from 10 cm to several meters. Different directional patterns of microphones can be used. There have not been any formal studies concerning the directional quality obtained with such systems. In stereophonic reproduction, it is known that the spaced microphone techniques produce a spread localization of virtual sources [17].

C. Wave Field Synthesis

When the number of microphones and loudspeakers is large, wave field synthesis [18] can be used. It reconstructs the whole sound field that appeared in the recording space in the listening room. Wave field synthesis is superior as a technique but the required loudspeaker systems are not often available. This method is not discussed any further in this paper.

D. Amplitude Panning

Amplitude panning is not a microphone technique but it is used frequently in sound reproduction. A monophonic sound signal is applied to loudspeakers with different amplitudes. The amplitudes are controlled by multiplying a sound signal with different gain factors. The listener perceives a virtual source direction which is dependent on the gain factors.

Ambisonics can also be treated as a special form of amplitude panning. This is because with it the sound is applied virtually to all loudspeakers with different gains, which may be positive or negative. Techniques where the sound emanates from all loudspeakers are also referred to as matrixing. An alternative approach is to use only a subset of loudspeakers for one virtual source. The pair-wise amplitude panning [19] method uses maximally two loudspeakers to produce one virtual source. The sound signal is applied to two loudspeakers between which the panning direction lies. If a virtual source is panned coincident with a loudspeaker, only that particular loudspeaker emits the sound signal.

Several panning laws have been suggested for pair-wise panning [10]. When loudspeakers are located symmetrically with respect to the listener, the tangent law [20], [21] most correctly estimates the virtual source direction [10]. The tangent law has been reformulated with vectors to a form which is called vector base amplitude panning (VBAP) and can be generalized also for three-dimensional (3-D) loudspeaker layouts [22]. The unit-length vectors $\mathbf{l}_m$ and $\mathbf{l}_n$ point from the listening position to the loudspeakers. The intended direction of the virtual source (panning direction) is presented with a unit-length vector $\mathbf{p}$. The gain factors of loudspeakers can be solved as

$$\mathbf{g} = \mathbf{p}^T\mathbf{L}_{mn}^{-1} \quad (3)$$

where $\mathbf{g} = [g_m g_n]^T$ and $\mathbf{L}_{mn} = [\mathbf{l}_m\mathbf{l}_n]$. The calculated factors can be used after suitable normalization, e.g., $||\mathbf{g}|| = 1$.

Pair-wise amplitude panning can be interpreted as an idealized coincident microphone technique. The polar patterns of the microphones corresponding to each loudspeaker can then be computed by using the selected panning law. In Fig. 3, the polar patterns are shown for 5.1 reproduction being computed with VBAP. It is quite clear that microphones having such polar patterns and no prominent coloration cannot be easily constructed.

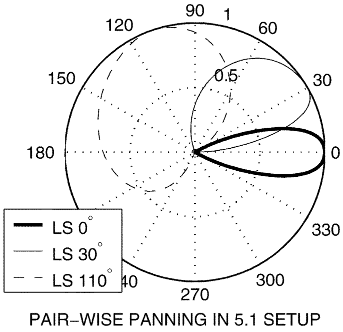

Fig. 3. Polar patterns of hypothetical microphones for 5.1 loudspeaker setup that would spatialize sound equally as it would occur when using pair-wise panning.

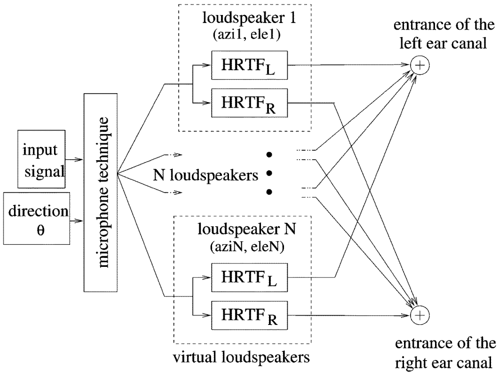

Fig. 4. Simulation of ear canal signals in arbitrary sound reproduction systems.

The directional quality of pair-wise panned virtual sources is relatively well-known. When a loudspeaker pair is symmetric with the median plane of the listener, the panning direction corresponds well to the perceived direction, i.e., the cones of confusions of the virtual source and the panning direction coincide. When a loudspeaker pair is located on either side of a listener, the perceived direction is biased toward the median plane. If direction ±90° of azimuth is inside a loudspeaker pair, there is a region around ±90° of azimuth where virtual sources cannot be positioned. This is because the cone of confusion of the virtual source can only lie between the cones of confusion of loudspeakers [10].

However, it is not known if the results obtained with pair-wise panning can be extrapolated to other amplitude panning methods in 2-D loudspeaker setups, such as Ambisonics or other matrixing techniques. In this paper, this topic is approached with simulations and listening tests.

IV. Modeling Virtual Source Perception

In the previous chapter, a variety of microphone techniques were described. To gain insight into spatial audio reproduction, it would be beneficial to compare different techniques. The most reliable method to accomplish this would be to conduct a large set of listening tests. Listening tests are, however, time-consuming and financially expensive. Computational simulation of virtual source perception is a faster method, although the model may not be valid in all cases. Nevertheless, the main cues for direction perception are relatively well-known, and have been used in the directional analysis of virtual sources before [10]. In this paper, a standard binaural model of directional hearing was applied to the analysis of virtual source directions. It was used to compute localization cues for the audio signals arriving at the ear canals.

Some simplifications however, must be tolerated. In this study, we have restricted our scope by eliminating the influence of the precedence effect as much as possible so that it would not have to be modeled. When the model omits the precedence effect it gives reliable results only if all incidents of a sound signal reach the ears within about a one ms time window. This can be achieved only in anechoic conditions, since in all rooms the reflections and reverberations violate the 1 ms window. Qualitatively, the results are also valid in moderately reverberant conditions. Furthermore, the microphones cannot be separated more than 35 cm in the analyzed setups, or the loudspeaker signals would violate the window.

The model of auditory localization used in this study consists of the following parts:

- simulation of microphone technique;
- simulation of ear canal signals during the listening phase;
- binaural model of neural decoding of directional cues;
- model of high-level perceptual processing.

Since the use of the model is described elsewhere [23], it is discussed here only briefly.

A. Simulation of Ear Canal Signals

Sound reproduction simulation as well as torso and ear filtering simulation in the model approximate the sound signals arriving at the listener's ear canals. A block diagram of the simulation is shown in Fig. 4. In this study, the audio signals applied to the loudspeakers are calculated by simulating a microphone technique. The microphones are considered to have an equal directional pattern at all frequencies, and to have flat frequency and phase responses. Ideal microphones are used in this study, since our primary interest is on how multiple microphones should be arranged to capture spatial sound for multichannel reproduction. Also, any comparison between panning methods and microphone techniques would be unequal otherwise. The effect of microphone nonidealities to directional perception is left for future studies. The signals arriving at the ear canals from each loudspeaker are computed using digital filters that implement the measured head-related transfer functions (HRTFs) of the corresponding direction. The arriving HRTF-filtered loudspeaker signals are added to form ear canal signals.

B. Binaural Model of Directional Cue Decoding

A schematic diagram for the binaural model of neural decoding for directional cues is presented in Fig. 5. The model takes the sound signal arriving at the ear canals as input and computes the decoded frequency-dependent ITD and ILD cues. It models the cochlea, the auditory nerve, and the binaural decoding. The cochlea, and auditory nerve models have been implemented based on the HUTear 2.0 software package [24]. The cochlear filtering of the inner ear has been modeled using a 42-band gammatone filter bank [25]. Center frequencies of the filter bank follow the ERB (equivalent rectangular bandwidth) scale [26]. Auditory nerve responses are modeled with half-wave rectification and low pass filtering. The impulse sharpening that occurs in the cochlear nucleus [27] is modeled roughly by raising the signal to a power of two.

The binaural computation consists of ITD and ILD decoding. The neural coincidence counting [27] that performs ITD decoding is modeled using the cross-correlation calculation as suggested by Jeffress [28]. The cross-correlations are calculated with a $[-1.1\text{–}1.1]$ ms time lag range at each ERB band. This produces a function for each frequency band that denotes how the ear signals coincided with different time lags. The time lag corresponding to the highest peak implies the ITD in each frequency band. Due to low-pass filtering of the auditory nerve, the ITD corresponds to carrier shifts at low frequencies and envelope shifts at high frequencies.

The loudnesses of each frequency band in each ear are calculated using Zwicker's formulae [29]. Due to its simplicity, this model is used instead of the more thorough model proposed by Moore [30]. The difference of loudness levels between the ears at each frequency band is treated as an ILD spectrum. The loudnesses are summed at each ear and each frequency band to form an estimate of the overall loudness of a sound source.

The sound sample used for simulation was 400 ms pink noise. The cross correlation computation for ITD and loudness computation for ILD were integrated over the sound sample. This implements a rectangular time window starting from 0 ms and ending at 400 ms. In the auditory system the corresponding time window is not rectangular. However, because we use a stationary signal, the shape of the window has no influence on the result.

C. Model of High-Level Perceptual Stages

Higher levels of human auditory processing produce direction perception as a fusion from ITD, ILD, and other cues. High-level perceptual mechanisms are generally regarded to be very complex. The authors are not aware of a physiologically-based computational model which would simulate such mechanisms of humans. However, the modeling of high-level perceptions would be beneficial since the ITD and ILD cues are measured in different scales, which means that they cannot be compared directly with each other. Additionally, ITDs or ILDs, cannot be compared between subjects due to the individuality of the cues. If a mapping from the cues to the spatial directions to which they correspond is formed, the cues can be compared in the above ways.

A straightforward method to form such a mapping is a functional model that consists of a database that holds the sound source ITDs and ILDs produced by a sound source at each direction for each individual (Fig. 6). An auditory cue value that has been measured from a virtual source is transformed into a direction angle value by a database search. Two subsequent values between which the cue lies are found. The resulting direction angle value is interpolated between these two values. The functional model computes frequency-dependent ITD angles (ITDA) and ILD angles (ILDA). These present the azimuth angles that the binaural properties of the measured virtual source suggested at each frequency band. Since this study considers only virtual sources on the horizontal plane, the database consists of ITD and ILD values of sound sources at azimuths $\theta = -90^\circ, -80^\circ, \ldots, 90^\circ$.

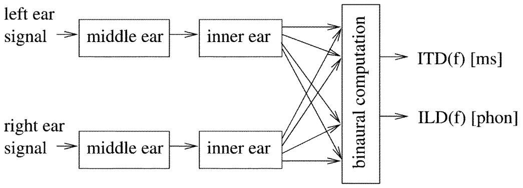

Fig. 5. Binaural model of directional cue decoding.

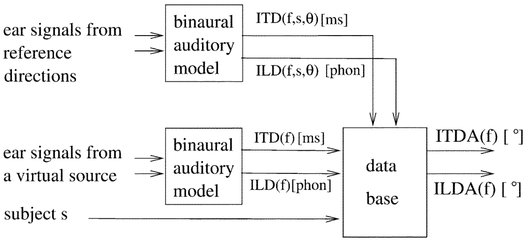

Fig. 6. Functional model of auditory localization.

The cues may behave in an inconsistent manner in some cases. Especially ILD behaves nonmonotonically, at frequencies approximately between 500 Hz and 4 kHz the absolute ILD value first increases and then decreases, when a distant sound source is moved from the median plane toward side [3]. Since an equal ILD value is produced with sound source in two directions, the ILD does not carry unequivocal information of source direction. When a nearby real source is moved similarly around the listener, nonmonotonic behavior vanishes, and larger ILD values occur [31]. Thus, at this frequency region, ILD carries mostly information about source distance.

Non-monotonic parts of the ILD curves are removed in the model described here, leaving the monotonic part around the 0° of azimuth. If a larger virtual source ILD value than is found on the ILD table emerges, the response is extrapolated from previous values in the table. However, the absolute value of ILDA cannot exceed 90°. This implies that the ILD database has to be evaluated to find possible regions where the cues do not carry directional information. The existence of these regions has to be taken into account in virtual source analysis.

The ITD values calculated for the database from HRTF-measurements might also be inconsistent, which would generate errors to ITDA estimation. To avoid this, the ITD databases were post-processed. If one value differed considerably from adjacent values, it was replaced with the mean of values produced by the same sound source at adjacent frequencies. In addition, the validity of computed ITDA values was checked and values that were clearly erroneous were removed. The virtual sources may generate large ITD values that do not correspond to any direction. If at any frequency band the value of a virtual source ITD cue is smaller or larger than any of the database ITD values at the corresponding frequency band, the ITDA is not calculated and is considered a missing value in the data analysis.

The model thus computes two estimates of perceived direction in each frequency band. In the case when the cues propose

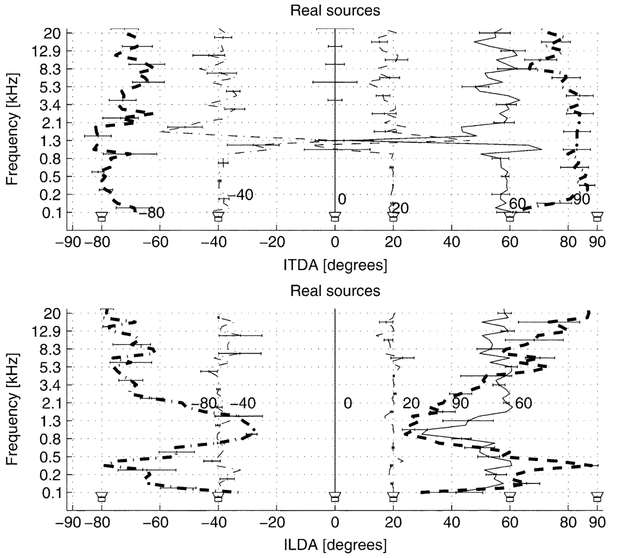

Fig. 7. ITDA and ILDA values measured with real sound sources. Whiskers denote 25% of standard deviation.

different directions, it is not known in advance what the listener will perceive. Depending on the frequency and the type of signal, there are three different mechanisms how perceived direction is formed, as reviewed in Section II.

- either ITD or ILD may dominate;
- a "traded" perception of direction between the directions proposed by the cues may occur;
- the listener may perceive two separate sound sources.

Later in this paper the auditory model output is compared to perceived directions. In the comparison we assume that the perceived direction will match with either ITDA or ILDA, or that perceived direction will lie between ITDA and ILDA.

D. Using an Auditory Model in Virtual Source Perception Simulation

The ITDA and ILDA angles were calculated for each simulated virtual source at 42 frequency channels. Each virtual source was simulated separately with ten individual HRTFs and symmetrically to both sides of the listeners. The resulting values obtained from left side HRTFs are turned to right side values by inverting the cue angle value sign. This results in 20 estimates of the direction that the virtual source produces at each frequency band. The mean value and standard deviation are calculated over individuals.

In the results, the means and standard deviations of cue angles with microphone systems and different sound source directions are plotted on the same figure. The polarity of ITDA and ILDA values are changed to negative in roughly half of the virtual source plots. This is done to maintain clarity in the figures.

To find possible regions where the cues do not carry directional information, as explained in Section. IV-C, the auditory model was tested by analyzing real sound sources in different directions around the listener. In the ideal case, estimates for directional perception that are constant with frequency should be achieved this way. The results are shown in Fig. 7. It can be seen that ITDA corresponds closely to the direction of sound source. There are some minor deviations at large sound source direction angles. The ILDA values behave consistently with directions below 50°. Even though ILDA generally deviates from the sound source direction with angles $>50°$, it is roughly correct only at frequencies higher than 4 kHz. The large deviations are caused by nonmonotonic ILD behavior with source direction [3].

This suggests that ITDA can be used in spatial sound analysis generally, whereas in ILDA analysis the fact that ILD does not have large values between 700 Hz and 4 kHz should be taken into account. The previous statement is valid in the case of distant sources, as ILD may get larger values when a source is near the head [31].

V. Simulation Results

A set of simulations was conducted. The loudspeaker systems used in tests were selected to be in the 5.1 setup and an eight-channel setup. The 5.1 setup was chosen because it is the most widely used multichannel setup. The eight-channel setup with loudspeakers in directions $[0°, 45°, \ldots, 315°]$ was used as to represent a slightly larger loudspeaker setup. Unlike the 5.1 system, the selected eight-channel setup also has loudspeakers in ±90° directions. This is beneficial when producing lateral virtual sources with pair-wise panning [32], [10]. However, it is not known how the perception of lateral sources with other reproduction methods is affected.

The microphone systems simulated were first- and second-order Ambisonics, a spaced microphone technique and pair-wise panning. Second-order Ambisonics was not used with the 5.1 setup since the utilized second-order polar pattern is too broad to be used in it. Also, the spaced microphone technique was not used with the eight-channel setup since such techniques have not been widely used with an eight-channel setup. The directions of simulated virtual sources were set to present worst cases in different setups, typically at the centre point between loudspeakers. In pair-wise panning, virtual sources were never simulated toward loudspeaker directions since in that case the sound would have emanated from only one loudspeaker. The results are shown for different systems separately.

A. First-Order Ambisonics

The results for first-order Ambisonics are shown in Fig. 8 for the 5.1 setup and for the eight-channel setup. The results for the 5.1 setup are considered first. The ITDA values at low frequencies are fairly consistent; however, they deviate from the target value prominently, especially with sound source directions far from the median plane. Also, there is a decreasing trend with increased frequency. The ITDA is inconsistent and compressed between $-30°$ and $30°$ at high frequencies. The ILDA is also generally inconsistent and deviates from the sound source direction prominently. The resulting stability of ITDA proposes that virtual sources will be localized relatively stably to one direction. However, the bias of the values toward the median plane predicts that consistent virtual sources are not produced in lateral directions. Also, especially with large sound source directions there should be a trend that the virtual source is localized nearer to the median plane at high frequencies.

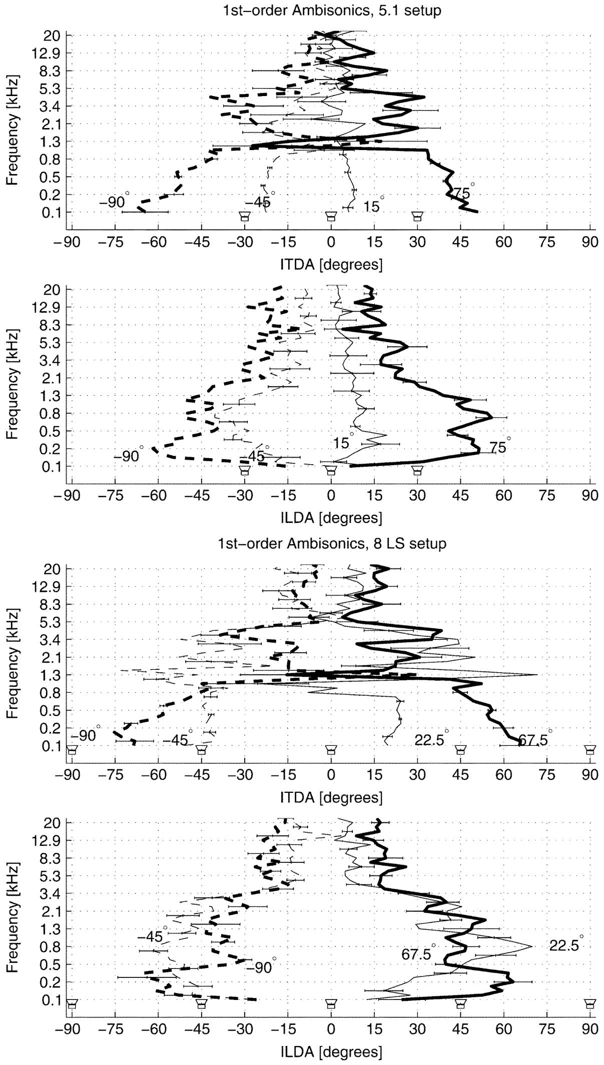

Fig. 8. ITDA and ILDA values simulated with first-order Ambisonics in the 5.1 and eight-channel loudspeaker setups with target sound sources in four directions. Whiskers denote 25% of standard deviation.

Simulation results for the eight-channel setup are considered next. The results are also presented in Fig. 8. When the results are compared with results from the 5.1 setup, it can be seen that the low-frequency ITD cues correspond better to target values. ITDA is accurate in the 22.5° case and is biased by only a few degrees in the $-45°$ case. Larger target direction values generate increasingly inaccurate ITDA values. They are highly dependent on frequency and have a high bias toward the median plane.

It seems that the ILD cues and the high-frequency ITD cues have not been notably improved by changing the loudspeaker setup. An interesting fact is that the ILDA values are quite large between 700 Hz and 2 kHz, which is not possible with distant real sources. Such large ILDA values are possible only with nearby real sources [31]. This may lead to near- or inside-the-head localization.

B. Second-Order Ambisonics

The simulation results for the eight-channel setup are shown in Fig. 9. The low-frequency ITDA indicates the sound source directions quite consistently and accurately. The cues at higher frequencies are inconsistent and biased prominently toward the median plane. The ILDA is roughly constant with frequency, although it does not coincide with sound source direction generally. The ILDA seems to be biased toward the median plane, especially at high frequencies. Both functions deviate between individuals.

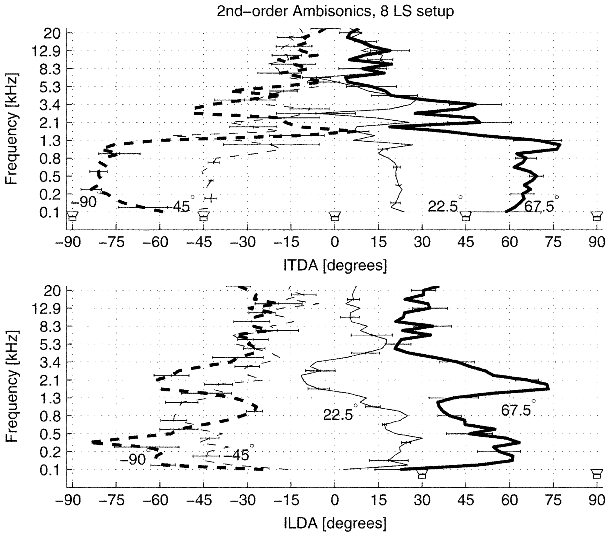

Fig. 9. ITDA and ILDA values simulated with second-order Ambisonics in the eight-channel setup with target sound sources in directions 22.5°, $-45°$, $67.5°$, and $-90°$. Whiskers denote 25% of standard deviation.

Altogether, this simulation suggests that second-order Ambisonics produces directional cues relatively accurately at low frequencies, whereas it fails to generate consistent cues at high frequencies. Differences between individuals also occur. When compared to 1st-order Ambisonics, it can be seen that ITDA curves are more accurate. This suggests that the directional quality is better with second-order Ambisonics than with first-order Ambisonics. However, the ILDA values are still unnaturally large between 700 Hz and 2 kHz.

C. Spaced Microphone System

In recording techniques for the 5.1 setup the microphones are often spaced considerably apart. This generates time differences between signals. The simulation of directional cues generated with this technique is problematic, since the auditory model used does not include the precedence effect. Thus the distances between the microphones are restricted to below 35 cm in this study.

For this simulation, a microphone array that has sufficiently short distances between the microphones was designed. This array has five cardioid microphones, two of them facing directions ±90° and one to 0°, separated by 5 cm from the center point, as shown in Fig. 10. The signals of these three microphones were applied to corresponding frontal loudspeakers. Microphones for ±30° loudspeakers were directed to ±90° to avoid overly strong cross talk between frontal loudspeakers. In practice, this is often done since cross talk may result in prominent coloration in the listening position. The two remaining microphones were in ±120° arrangement separated by 20 cm

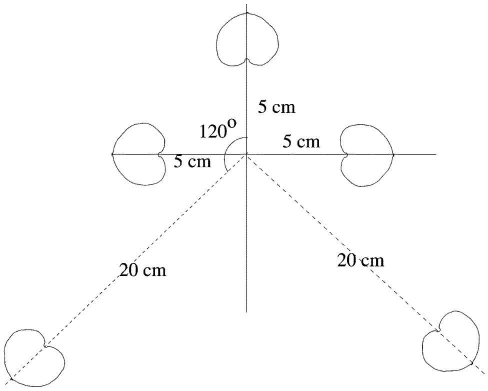

Fig. 10. Hypothetical microphone system for the 5.1 setup.

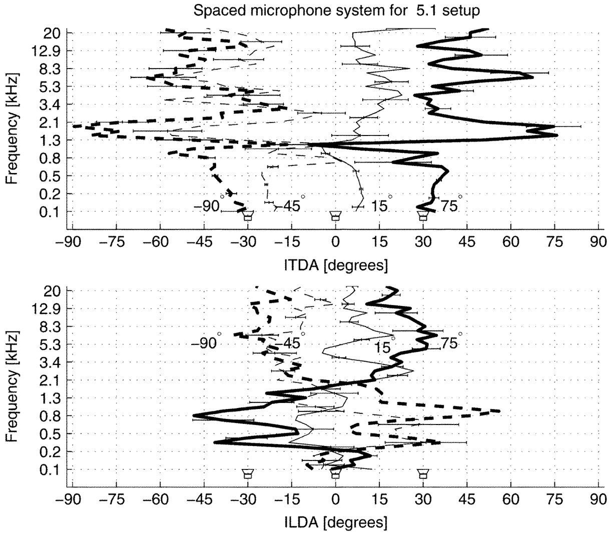

Fig. 11. ITDA and ILDA values simulated with a spaced microphone system (Fig. 10) in the 5.1 loudspeaker setup with target sound sources in directions $15°, -45°, 75°$, and $-90°$. Whiskers denote 25% of standard deviation.

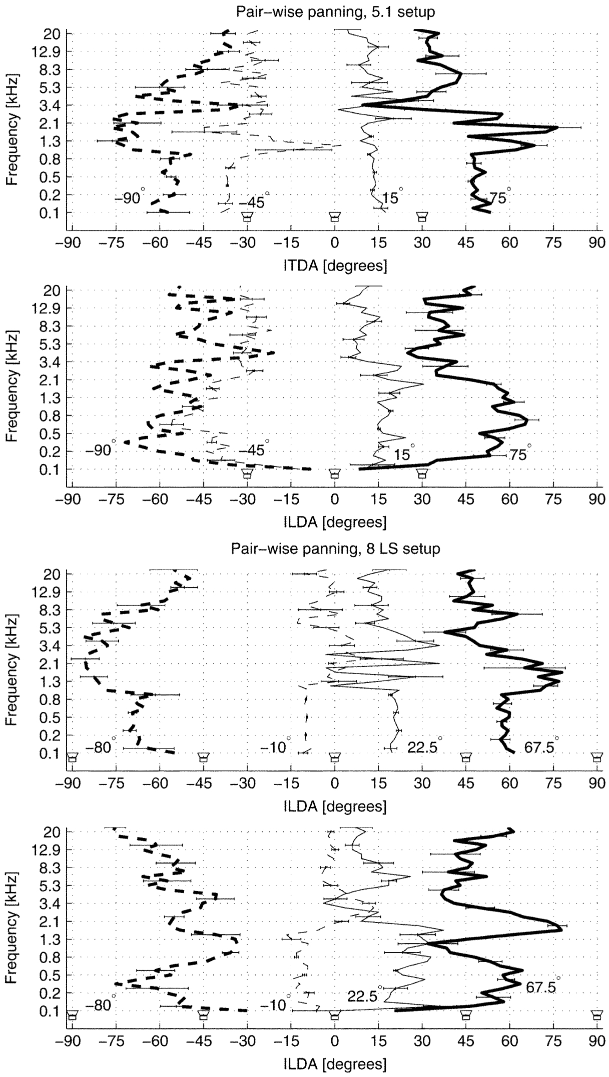

Fig. 12. ITDA and ILDA values simulated with pair-wise panning in the 5.1 loudspeaker and the eight-channel setup with four target sound sources. Whiskers denote 25% of standard deviation.

from the center. These signals were applied to speakers at ±110°.

The simulation results are shown in Fig. 11. The ITDA behaves fairly consistently at low frequencies. However, it fluctuates more than when using coincident techniques in which the values are compressed roughly between $-40°$ and $40°$. Even though the high-frequency ITDA is fairly inconsistent, the values are roughly coincident with sound source directions. The ILDA is generally inconsistent, especially at low frequencies as it has values on the other side of the median plane than the ITDA has. In contrast, at high frequencies the ILDA is roughly coincident with low-frequency ITDA.

D. Pair-Wise Panning

The results of simulations are presented in Fig. 12. The low-frequency ITDA functions are consistent up to 1 kHz. However, they are biased toward the median plane slightly with the loudspeaker pair (0°, 30°), and prominently with the loudspeaker pair (30°, 110°). High-frequency ITDA and ILDA act fairly consistently with frequency and coincide roughly with panning direction. The bias toward the median plane is known to occur when the loudspeaker pair is not symmetric with the median plane of the listener [32], [10]. With loudspeaker pair (30°, 110°) the bias is very large with a panning direction of 75°. This source for this bias is known. With amplitude panning the virtual source cone of confusion is always between the cones of confusions of the loudspeakers, as explained in Section III-D. In this case, the angles between the median plane and the cones of loudspeakers are 30° and 70°. The perceived direction should be about midway between these cones, corresponding to an azimuth of 52°, which matches with low-frequency ITDA.

When the loudspeaker system was changed to the eight-channel setup, there are some prominent changes in the simulated values, as seen in Fig. 12. When the loudspeaker pair (0°, 30°) changes to pair (0°, 45°) which has a larger spatial opening, the virtual source in between the loudspeaker produces ITDA and ILDA which are slightly more inconsistent with frequency. When changing the pair (30°, 110°) to pair (45°, 90°) where positioning should be possible to all azimuths between the loudspeakers, the bias indeed decreases dramatically. With this loudspeaker pair, virtual sources can be positioned to any direction between the loudspeakers, unlike

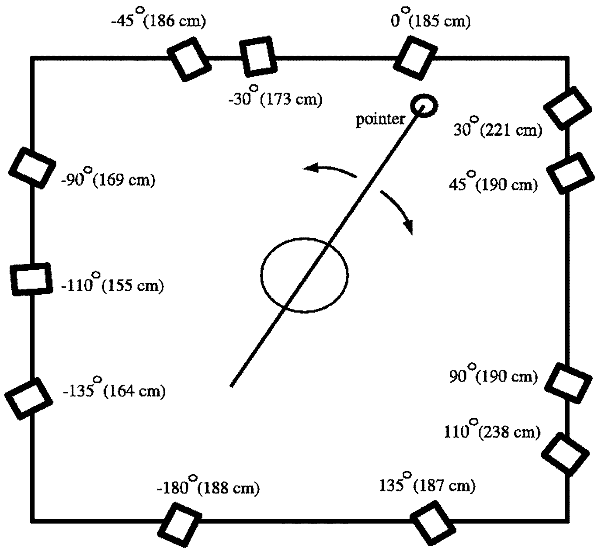

Fig. 13. Eight-channel and 5.1 Loudspeaker setups used in the listening tests. The different loudspeaker distances were compensated by appropriate delays.

with the pair (30°, 110°). There does not seem to be a significant change in consistency of ITDA and ILDA between pairs (30°, 110°) and (45°, 90°).

VI. Listening Tests

In the previous section, the results from a large set of simulations were presented. The validity of the results was assessed with listening tests. In the tests, a method of adjustment was used [33]. Listeners adjusted an auditory pointer to the same direction as a narrow-band virtual source. The physical direction of the auditory pointer was interpreted as the dominant, perceived direction of the virtual source.

A. Test Setup

The eight-channel and the 5.1 loudspeaker setups used in simulations were constructed inside an anechoic chamber. The subwoofer specified in the 5.1 system was not included in the test setup. The chamber used in the tests can be considered anechoic for frequencies higher than 100 Hz. The Genelec model 1029A loudspeaker was used for all loudspeaker positions in both setups. Fig. 13 illustrates the loudspeaker placement in the anechoic chamber as seen from above. The front loudspeaker at 0° was common for both setups. The eight-channel setup used the speakers at $0°, \pm 45°, \pm 90°, \pm 135°$, and 180°, whereas, the 5.1 setup employed the speakers at $0°, \pm 30°$ and $\pm 110°$. The optimal listening position, i.e., the sweet spot, was located below the rotary axis of the pointer. As the loudspeakers were at different distances from the listening position, the distance differences were compensated by adding appropriate delays to the signals of each channel. The loudspeaker amplifier gains were also level-aligned by measuring a reference broadband noise with an SPL meter at the listening position.

The acoustic pointer was a spherical loudspeaker with a radius of 5 cm attached to a rotary axis above the listener. The rotating level of the pointer was just above the level of the loudspeakers. The subjects were able to move the pointer by using a mechanism that did not disturb the incoming sound field; they rotated the pointer freely by using a circular band, as illustrated in Fig. 14.

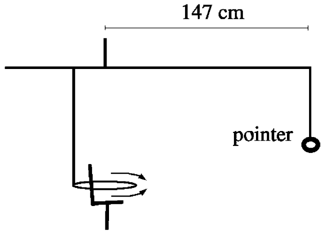

Fig. 14. Mechanism to rotate the auditory pointer around the listener. The band used to rotate the auditory pointer can be seen to float around the listening chair.

The position of the pointer was determined using three microphones placed on the walls of the anechoic chamber. The distances from the pointer to each microphone at a given position were calculated and 3-D positional coordinates of the pointer were computed. During the listening tests, the pointer loudspeaker emitted pink noise equalized with the inverse of the loudspeaker's magnitude response. Pink noise was assumed to be a kind of signal that would present the physical direction of the pointer well. Although the virtual source sounds had a narrow band width, the sound of the auditory pointer was always pink noise. Using narrow-band noise as an auditory pointer would have caused some signal-dependent effects in the directional perception of the pointer [3].

The signals used to produce virtual sources were band-limited pink noise. In this way, it was possible to investigate the localization of virtual sources frequency-dependently. Five octave-band noise signals with center frequencies of 200 Hz, 400 Hz, 800 Hz, 1600 Hz, and 3200 Hz with −40 dB/octave rolloff were used. Consequently, each virtual source was presented using five different frequency bands.

B. Test Procedure

During tests, the test subjects were seated on a chair which had been fixed so that the subject was facing toward the front speaker at 0°. The subjects were unpaid volunteers, mostly workers from the laboratory of the authors aged below 35. The subjects did not report any hearing deficiencies. A light-weight head rest ensured that the center of the subject's head remained in the sweet spot throughout the test. The loudspeakers were visible, but subjects were instructed to perform the localization task with eyes closed.

The auditory pointer was selected instead of some other pointing method, e.g., visual, motional etc., since it has been found that humans generate errors and bias when interpreting auditory perception with any method [3]. When they are comparing auditory perception to auditory perception, and adjusting the apparatus until the difference in direction cannot be perceived, there should be fewer artifacts.

The virtual source and the pointer signal were presented continuously one after another. Both signals were 500 ms long identical samples with a short fade-in and fade-out. The signals were repeated until the listener adjusted the pointer to the same direction as the virtual source and pressed a key on a keyboard on his lap to indicate that the adjustment was complete. After this, the location of auditory pointer was tracked and a signal was played to indicate that the next test item was on. If the virtual source was spread, the listeners were instructed to choose a random direction inside the virtual source. The test was organized so that one session for one loudspeaker setup consisted of 60 trials; five signals times three systems times four panning angles. Each trial took approximately one minute to perform. Sessions were divided into two 30-min parts with a break in between. The same session was completed two times by the same subject. The tasks were presented in randomized order for each session. The subjects were not aware of which reproduction system and which target direction were applied at a time.

In each system, four target directions corresponding to the worst cases with different layouts were employed, and positioned symmetrically around the median plane, e.g., $-15°, +15°, -75°$, and $+75°$. The data from targets on the left side of the median plane was inverted and combined with the corresponding right-side target directions. All target directions were in the frontal hemisphere. Since the simulation data produced values between $-90°$ and $+90°$, the front-back confusions were resolved to front before data analysis.

C. Statistical Analysis

To quantify the performance of different systems, error measures presented in [34] were used. They include the run RMS error $\langle\overline{D}\rangle$, which quantifies the absolute accuracy of a system. A RMS deviation D between perceived directions and one target direction is calculated at all frequency bands and for all repetitions for a single subject. The statistic $\langle\overline{D}\rangle$ is a mean over subjects, and is accompanied with standard deviation. The value is computed for a single target direction at a time. A run RMS error value is denoted, for example, as $\langle\overline{D75}\rangle = 12°(2.1)$, which would mean that the mean value over subjects' D-values for target direction 75° is 12° and the corresponding standard deviation is 2.1°.

A standard deviation value is computed for all subject's responses to one target direction for one system. The run standard deviation $\langle\overline{s}\rangle$ is the mean of these deviations, accompanied with the corresponding standard deviation over subjects' values. This value quantifies the response spread.

The mean error $\langle\overline{E}\rangle$ is the average displacement of the perceived direction from the target value for a system. A mean displacement is computed for each listener, and the average value and standard deviation is taken over subjects, thus producing the final values. Possible bias from targeted direction in virtual source perception is seen in this statistic. This value is also computed for each target direction, and presented analogously with run RMS error. The statistical significance of the bias was tested with one-sample t-test with 95% confidence level in each case.

One-way within-subjects ANOVA was used to find out if the frequency band of a stimulus had a significant effect to the perceived direction with each particular system and target direction. The dependent variable was perceived direction and the only factor was frequency band. The analysis was conducted to data from one target direction and one system at a time.

VII. Listening Test Results

The listening test results are presented with numerical statistics in Table I. Also, the results from each tests are shown by plotting the adjusted auditory pointer direction data together with simulated ITDA and ILDA values. The plots show the mean and standard deviation of the data. The ITDA and ILDA data has been taken from the simulation results presented in Section V. ITDA and ILDA values are averaged both over frequencies corresponding to each octave band and over ten individuals. The lower panels of the plots show the averaged frequency dependency of ITDA and ILDA inside each octave band. The results from the test that investigated the directional accuracy of the auditory pointer apparatus are first reported. After this, listening test results are shown for each tested loudspeaker setup separately.

TABLE I
Statistics for Listening Tests Results. The Symbol $\langle\overline{D}\rangle$ Denotes the Run RMS Error. A Separate Value Is Provided for Each Target Direction. Run Standard Deviation Is Presented as $\langle\overline{s}\rangle$ Values and Mean Error With $\langle\overline{E}\rangle$ Values

5.1		$\langle\overline{D15}\rangle$	$\langle\overline{D75}\rangle$	$\langle\overline{s15}\rangle$	$\langle\overline{s75}\rangle$	$\langle\overline{E15}\rangle$	$\langle\overline{E75}\rangle$
	Ambi I	11(1.6)	36(4.1)	6.9(2.9)	12(2.6)	-8.5(0.5)	-34(4.1)
	Spaced	9(1.4)	37(4.4)	5.3(1.6)	12(2)	-7.3(1.3)	-35(4.7)
	Pairws	4.1(0.9)	20(6.6)	3.9(1.1)	11(3.3)	0.2(1.5)	-15(9.5)
8-ch		$\langle\overline{D22}\rangle$	$\langle\overline{D67}\rangle$	$\langle\overline{s22}\rangle$	$\langle\overline{s67}\rangle$	$\langle\overline{E22}\rangle$	$\langle\overline{E67}\rangle$
	Ambi I	12(4.5)	20(3.6)	12(4.4)	16(1.2)	1.6(3.6)	-12(5.2)
	Ambi II	7.8(1.4)	14(2.9)	6.6(2.2)	13(2.9)	-3.6(2.3)	-2.5(2.9)
	Pairws	7.5(2.7)	10(1.8)	6.7(3.6)	9.7(1.6)	-2.8(2)	0.8(3.3)

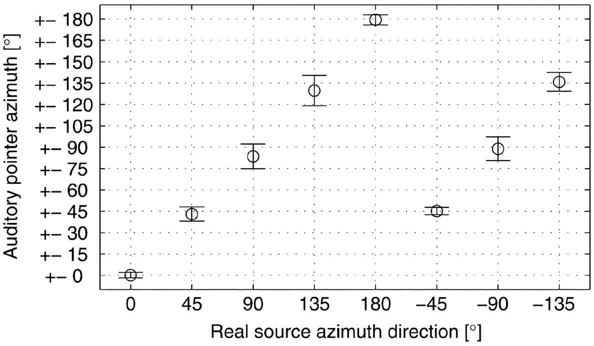

Fig. 15. Accuracy of the method of adjustment applied in these tests. Six listeners adjusted the auditory pointer to the direction of single loudspeakers three times. Circles denote the mean direction of adjustments, and whiskers the standard deviation.

A. Accuracy of Listening Test System

The apparatus for auditory pointer adjustment was tested to see how well the direction perceptions of the test attendees can be expressed with it. Six listeners matched the auditory pointer direction with single real sources in directions $[0°, 45°, \ldots, 315°]$. The real sources emitted pink noise and each trial was repeated three times. The results are shown in Fig. 15. It can be seen that the results correspond to the human directional resolution [3]. At 0°, the standard deviation of pointed directions is 1.9°. The deviation is slightly larger behind the listener, and considerably larger on the sides. Based on these results, it can be assumed that the auditory pointer apparatus provides sufficiently accurate data for these tests.

B. Tests With the 5.1 Loudspeaker Setup

The systems tested with the 5.1 setup were first-order Ambisonics, spaced microphone array, and pair-wise panning. The target directions were selected to be $\pm 15°$ and $\pm 75°$. The tests were conducted with six listeners who performed the adjustment to all four target directions twice. Three of the listeners also performed the test reported in the previous section. The results of the tests are shown in Fig. 16, together with the corresponding simulation results. Statistics for overall performance are shown in Table I.

Listening Test Results: The bias is characterized by a mean error $\langle \overline{E15} \rangle$. With a target direction of 15°, there was a prominent bias toward the median plane with Ambisonics and the spaced microphone array, $\langle \overline{E15} \rangle$ reached values $-8.5°(0.5)$ and $-7.3°(1.3)$, respectively. These values were found statistically significant with t-test ($p < 0.001$). With pair-wise panning, the mean of adjusted values did not depart from the target value significantly according to the t-test ($p = 0.6255$). With all systems, the listeners perceived the virtual source almost constantly to one direction independent of frequency, as seen in Fig. 16. However, there are some slight deviations with frequency, which were found to be statistically significant with ANOVA (Ambisonics: $p = 0.005$; Spaced array: $p = 0.005$ and pair-wise panning: $p = 0.037$).

In the *direction* 75° case, the adjusted values of all systems are biased toward the median plane. These effects were found statistically significant with t-test ($p < 0.001$ in all cases). With Ambisonics and the spaced array, the bias is on average $-34°(4.1)$ and $-35°(4.7)$, respectively, whereas, with pair-wise panning the bias is on average only $-15°(9.5)$. In this case, there is also a prominent frequency dependency with all systems. With Ambisonics, the perceived direction is biased more toward the median plane with increasing frequency (Fig. 16). With the spaced microphone array and pair-wise panning, the angle between the median plane and perceived direction grows slightly until 1600 Hz and then decreases (Fig. 16). The frequency-dependency was also found to be significant with ANOVA (Ambisonics: $p < 0.001$; spaced array: $p < 0.001$ and pair-wise panning: $p = 0.048$).

The bias toward the median plane with Ambisonics and spaced array systems also causes the run RMS error $\langle \overline{D} \rangle$ to also have large values for both target directions. The values with Ambisonics and spaced microphones are respectively $11°(1.6)$ and $9°(1.4)$. Both are more than two times larger than with pair-wise panning $4.1°(0.9)$. In the *direction* 75° case, there is bias also with pair-wise panning, which introduces a relatively large run RMS error $\langle \overline{D75} \rangle$.

The listeners have adjusted the auditory pointer quite consistently with different repetitions on the left and the right side of the median plane, which is seen in run standard deviation $\langle \overline{s} \rangle$ values in Table I. In the 15° case, the values are relatively low especially for pair-wise panning, although there has been more intra-subject variation in the 75° case.

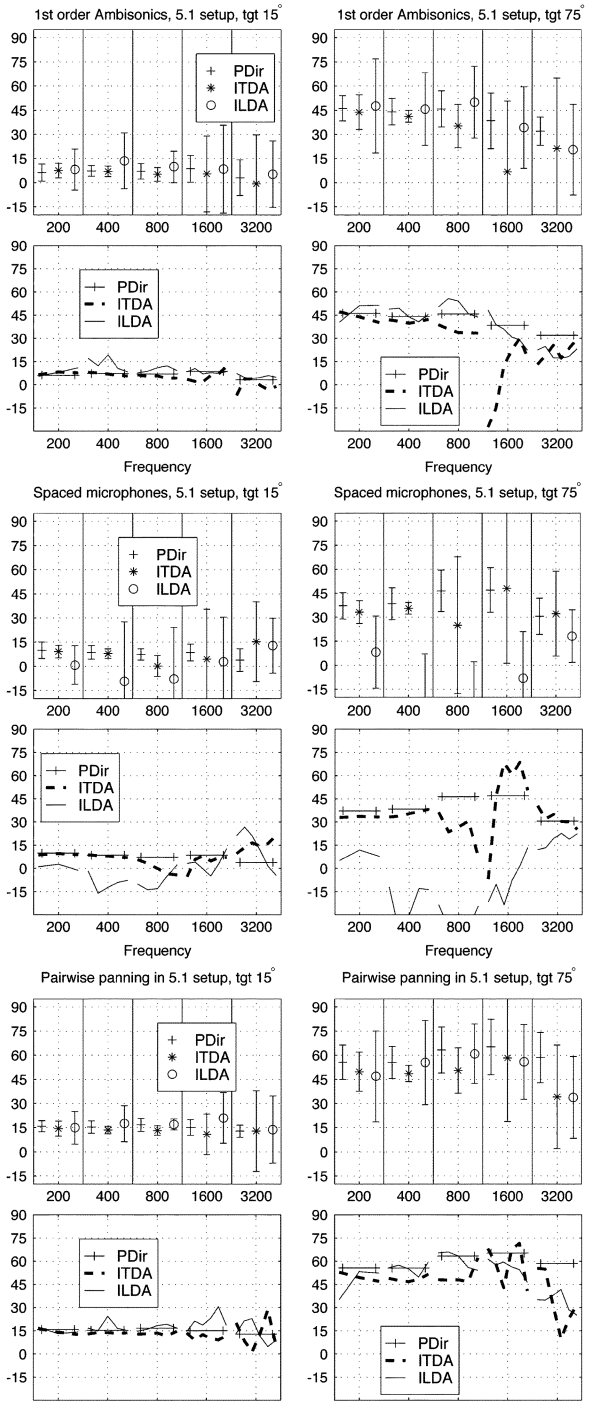

Fig. 16. Listening test results combined with corresponding modeling results. PDir denotes the perceived direction in the listening test. first-order Ambisonics, a spaced microphone technique shown in Fig. 10 and pair-wise panning were used to produce virtual sources to $\pm 15°$ and $\pm 75°$ directions with the 5.1 loudspeaker setup. Six listeners adjusted an auditory pointer to the same direction as a virtual source generated with octave-band pink noise. This procedure was repeated twice for all four virtual source directions at all frequency bands.

1) Comparison of Modeling Results With Listening Test Data: With a target direction of 15° and low frequencies, the ITDA corresponds well with listening test data, as seen in Fig. 16. At high frequencies it corresponds to either one of ITDA or ILDA or to an average value of them. With the spaced array, there are some deviations; with the 800 Hz band, the mean of neither ITDA or ILDA corresponds to perceived direction. However, the ITDA at the lowest part of the frequency band produces a match, which means that ITD can be the most prominent cue. With the spaced array there is some deviation between ITDA and ILDA, especially at low frequencies. It seems that at low frequencies ITD has dominated totally over ILD.

With a target direction of 75°, the values have large variations with frequency, between individuals and between ITDA and ILDA. At low frequencies, ITDA has been the most prominent, as seen with the spaced array case. At high frequencies, the relation between cues and perceived direction is often unclear. However, it seems that the virtual source has often been perceived slightly farther from the median plane than either of the cues suggest. Problematic cases are especially *spaced array* 75° 800 Hz and *pair-wise panning* 75° 3200 Hz where there seems to be only a weak correspondence between the ITDA or ILDA values and the perceived directions.

The auditory model simulation results in Section V-A suggest that the virtual sources created with first-order Ambisonics are perceived nearer the median plane, as frequency is increased. This is also seen in the listening test results, although the effect is not as strong as the model predicts. In the simulation results, directional estimates farthest from the median plane for the 5.1 system were about 50–60°, which were slightly exceeded in listening test data. However, on the simulation data and the listening test data it can be assumed that it is impossible to create direction perceptions farther than 70° from the median plane using the 5.1 loudspeaker system.

C. Tests With Eight-Channel Loudspeaker Setup

The listening tests with the eight-channel loudspeaker setup were ran with 1st- and second-order Ambisonics, and with pair-wise panning. The target directions were selected to be 22.5° and 67.5° since they lie between the loudspeakers and present the “worst case” at least for pair-wise amplitude panning. The tests were conducted with six listeners, three of whom also participated to the tests reported in Sections VII-A and VII-B. The adjustment was conducted to all four target directions twice, as in the 5.1 tests.

The results of the tests are shown in Fig. 17 together with the corresponding simulation results. Statistics for overall performance are shown in Table I.

1) Listening Test Results: It seems that the symmetric loudspeaker layout is more suitable for the first-order Ambisonics method. With a 22.5° target direction, $\langle\overline{E22}\rangle$ was not found to differ significantly from zero with t-test $(p = 0.1596)$, which indicates that there is no bias in this case. With pair-wise panning and second-order Ambisonics, there is a small negative bias, which was found to be significant with t-test $(p < 0.001)$.

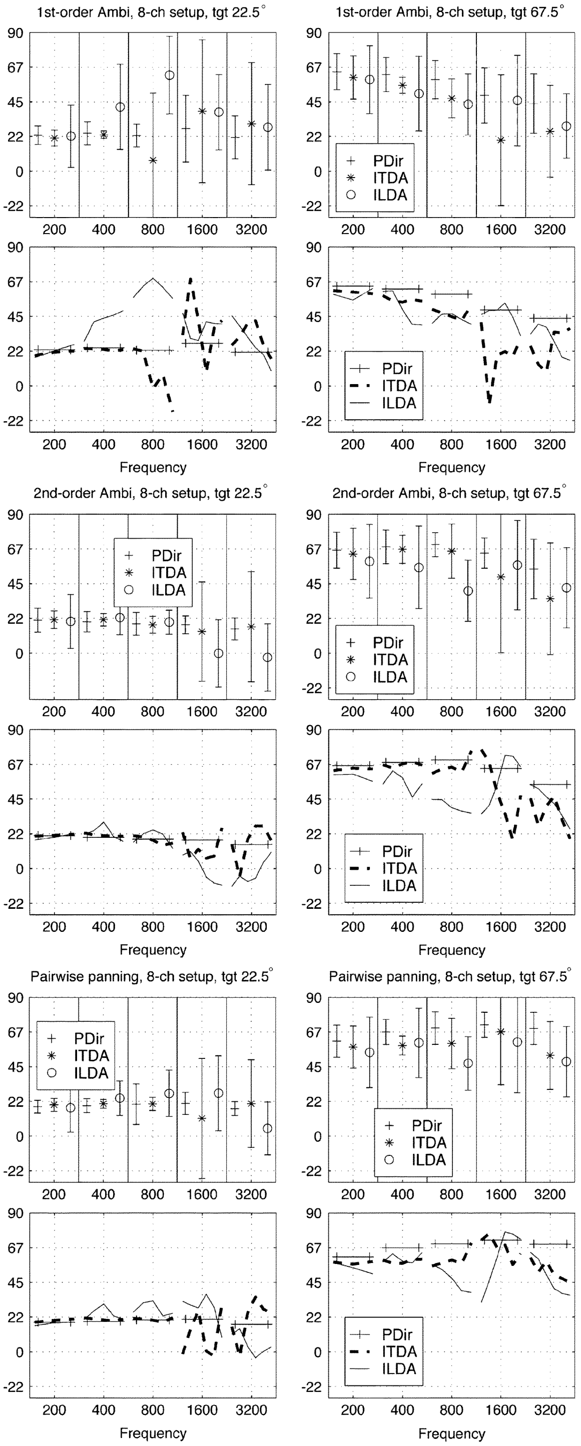

Fig. 17. Listening test results combined with corresponding modeling result. PDir denotes the perceived direction in the listening test. first-order and second-order Ambisonics and pair-wise panning were used to produce virtual sources at $\pm 22.5^\circ$ and $\pm 67.5^\circ$ directions in eight-channel listening. Six listeners adjusted an auditory pointer to the same direction as a virtual source generated with octave-band pink noise. This procedure was repeated twice for all four virtual directions at all frequency bands.

The perceived direction of virtual sources produced with first-order Ambisonics, and with pair-wise panning was not found to be dependent on frequency, whereas the perceived direction with second-order Ambisonics was found to depend on frequency in ANOVA tests (first-order Ambisonics: $p = 0.622$; second-order Ambisonics: $p = 0.025$ and pair-wise panning: $p = 0.463$). Although direction perception with second-order Ambisonics was found to depend on frequency, the variation is small, as seen in Fig. 17. With first-order Ambisonics, the standard deviation of perceived directions is large at high frequencies.

In the target direction, 67.5° case, there is a prominent bias with first-order Ambisonics ($-12°(5.2)$), and a small bias in second-order Ambisonics ($-2.5°(2.9)$), which were found significant in the t-test ($p < 0.001$ and $p = 0.0487$, respectively). The bias with with pair-wise panning was not found significant with the t-test ($p = 0.4166$).

The frequency-dependence is evident with the 8-channel setup in ANOVA tests (first-order Ambisonics: $p < 0.001$; second-order Ambisonics: $p = 0.005$ and pair-wise panning: $p = 0.040$). The dependencies are similar as with the 5.1 system. A decreasing curve occurs with first-order Ambisonics (Fig. 17). With pair-wise panning, a similar slightly increasing-decreasing curve, as with second-order Ambisonics can be seen.

When investigating the run RMS error with the eight-channel setup, it seems that the best average accuracy is obtained again with pair-wise panning. second-order Ambisonics competes equally in the 22.5° target case. Although the bias for first-order Ambisonics with 22.5° target has reduced significantly from the corresponding case with the 5.1 setup, the run RMS error $\langle\overline{D22}\rangle$ is relatively high, having the value 12°(4.5). This is explained by the large intra-listener variations shown in the $\langle\overline{s22}\rangle$ value 12°(4.4), and by large standard deviation of perceived direction which is present at some frequency bands (Fig. 17).

2) Comparison of Modeling Results With Listening Test Data: In the 22.5° case the simulation results match with listening test results in a similar way as with the 5.1 system, as seen in Fig. 17. Generally, at low frequencies, the ITDAs correspond with perceived directions and at higher frequencies either one of ITDA or ILDA or their average matches with perception. One interesting fact is that at high frequencies, the listening test data has a relatively low spreading with second-order Ambisonics and pair-wise panning, although the ITDA and ILDA values have a large variation with frequency and between individuals. Only with first-order Ambisonics, the large spreading has a relation to more spread listening test data.

In the 67.5° case there seems to be a systematic bias in ITDA and ILDA values with all systems (Fig. 17), similarly as was found in the 5.1 case. At all frequency bands of all systems, the mean of perceived directions is farther away from the median plane than the means of the ITDA or ILDA values suggest. The reason for this is not known. By investigating the frequency-dependent ITDA and ILDA, it seems that hearing mechanisms have selected the largest auditory cues available, and used them as most prominent direction. More studies need to be conducted on this subject.

VIII. Discussion on Validity of Simulation Results

This paper is the first attempt to analyze the directional perception of virtual sources created with multichannel reproduction techniques using a binaural auditory model and listening tests. The binaural auditory model computed the frequency-dependent ITDA and ILDA that predict the cone of confusion in which a sound source lies. These values were compared with the auditory pointer adjustment data from listening tests. This comparison is not straightforward, since there are two values. It is not accurately known which one is dominant, how they fuse into a single percept, or how they produce a spreaded auditory object.

The results show that the auditory model was able to explain some prominent features of the listening test data. The subjective directions of the virtual sources were mostly explained by examining the ITDA values at frequencies below 1 kHz and both the ITDA and the ILDA at high frequencies. When the virtual source was positioned farther from the median plane, there seemed to be a slight bias between cues and the listening test data. The listeners adjusted the auditory pointer farther from the median plane than ITDA had predicted. The reason for this effect is not known; it can be due to some inaccuracy in the auditory model, in the listening test setup, or due to some source of bias in the listening test method. A similar bias is found when real sources are analyzed with the model in Fig. 7, however with smaller magnitude. At higher frequencies, the interpretation of the simulation results is more problematic. Traditionally, it has been thought that the ILD cue should be salient at these frequencies. However, there is no clear relationship between ILDA and the auditory pointer adjustment data, although both ITDA and ILDA coincided relatively well with listening test data.

One reason for these deviations might be the fact that non-individual HRTFs were used in the simulations. Small changes in HRTFs and in listening test setups might have caused inaccuracy in simulation. Also, in the spaced array case, the fact that the precedence effect is not included in modeling may have caused deviations. It is possible that at some frequency bands the precedence effect has been effective, although the inter-channel delays were shorter than 1 ms with the spaced microphone array utilized.

When performing the test, the listeners gave their answer as a single auditory pointer direction. The amount of spreading of the virtual source, or the number of perceived auditory objects were not reported at all. Although some of the listeners reported that some virtual sources were diffuse, they apparently adjusted the directions very similarly as the rest of the subjects in these cases. It seems that although the source is spread, some of the cues are "leading," and the virtual source is judged according to these "leading" cues. The research on this topic is left for future studies.

Also, it has to be noted that these results are valid only in the best listening position. This analysis does not imply how the quality is degraded outside the best listening position, where the loudspeaker signals do not arrive at the listener simultaneously.

IX. Conclusion

In this study the directional qualities of different reproduction techniques were estimated using a binaural auditory model in the best listening position. The auditory model was used to analyze the virtual sources generated with different reproduction methods to a standard 5.1 setup without a subwoofer, and to an eight-channel setup. The simulation results were verified with psychoacoustical listening tests, in which the attendees adjusted an auditory pointer emanating broad-band noise to the same direction as their perception of the virtual source containing octave-band noise at five different frequencies. The listening test results matched the simulation results generally well, although there were some systematic deviations. The model gave the most reliable predictions with virtual sources near the median plane, and at low frequencies. Farther from the median plane, the output of the model was in general hard to interpret, and it suggested directions nearer the median plane than those that were actually perceived.

Both the simulation results and the listening tests suggest that with the 5.1 setup it is impossible to create virtual sources in directions farther than 70° from the median plane with the tested reproduction systems. These systems were first-order Ambisonics, a spaced microphone system and pair-wise amplitude panning. With the eight-channel setup, the bias toward the median plane was prominently smaller with the tested systems, which were 1st- and second-order Ambisonics and pair-wise panning.

The virtual sources produced with first-order Ambisonics generate ITDA and ILDA that are consistent with frequency when the sound source is near the median plane. However, when a sound source is farther from the median plane, ITDA and ILDA depend more on frequency. This results in frequency-dependent perception of the virtual source. A prominent bias toward the median plane was detected with all sound source directions. The corresponding results with the eight-channel setup have significantly less bias toward the median plane, although there is still a strong frequency-dependency in the lateral direction. The virtual sources generated by the second-order Ambisonics with the eight-channel layout have almost no bias and are only slightly frequency-dependent.

The results with the tested spaced microphone system were not as divergent as might have been expected based on the simulation results. It seems that although ITDA and ILDA behave differently at low frequencies, the listeners relied on the ITD cue only. Also, some of the listening test results could not be understood by examining auditory model output. The results from the pair-wise panning tests could be explained well with the auditory model, although when the target direction was above 50° the auditory model gave results that are biased toward the median plane by about 10°.

References

[1] A. D. Blumlein, "Audio Eng. Soc.," U.S. Patent 394 325 1931, Dec. 14, 1986.

[2] "Multichannel Stereophonic Sound System with and Without Accompanying Picture," International Telecommunication Union, Geneva, Switzerland, Tech. Rep., 1992-1994. I. R. BS.775-1.

[3] J. Blauert, *Spatial Hearing*, Revised ed. Cambridge, MA: MIT Press, 1997.

[4] R. H. Gilkey and T. R. Anderson, Eds., *Binaural and Spatial Hearing in Real and Virtual Environments*. Hillsdale, NJ: Lawrence Erlbaum, 1997.

[5] P. M. Zurek, "The precedence effect," in *Directional Hearing*, W. A. Yost and G. Gourewitch, Eds. New York: Springer-Verlag, 1987, pp. 3–25.

[6] G. Harris, "Binaural interactions of impulsive stimuli and pure tones," *J. Acoust. Soc. Amer.*, vol. 32, pp. 685–692, 1960.

[7] E. Hafter and C. Carrier, "Binaural interaction if low-frequency stimuli: The inability to trade time and intensity completely," *J. Acoust. Soc. Amer.*, vol. 51, pp. 1852–1862, 1972.

[8] F. Wightman and D. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Amer.*, vol. 91, pp. 1648–1661, 1992.

[9] ——, "Factors affecting the relative salience of sound localization cues," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1997.

[10] V. Pulkki, "Spatial Sound Generation and Perception by Amplitude Panning Techniques," Ph.D. dissertation, Dept. Elect. Comput. Eng., Helsinki Univ. Tech., [Online]. Available: http://lib.hut.fi/Diss/2001/isbn9 512 255 324/, 2001.

[11] F. Rumsey, *Spatial Audio*, Oxford, U.K.: Focal Press, 2001.

[12] M. A. Gerzon, "Panpot laws for multispeaker stereo," in *The 92nd Convention*, Vienna, Austria: Audio Engineering Society, Mar. 24–27, 1992. Preprint no. 3309.

[13] K. Farrar, "Soundfield microphone," *Wireless World*, vol. 85, pp. 99–102, 1979.

[14] M. J. Evans, A. I. Tew, and J. A. S. Angus, "Perceived performance of loudspeaker-spatialized speech for teleconferencing," *J. Audio Eng. Soc.*, vol. 48, no. 9, pp. 771–785, Sep. 2000.

[15] D. G. Malham, "Higher order ambisonic systems for the spatialization of sound," in *Proc. Int. Computer Music Conf.*, Beijing, China, 1999, pp. 484–487.

[16] G. Monro, "In-phase corrections for ambisonics," in *Proc. Int. Computer Music Conf.*, 2001, pp. 292–295.

[17] S. P. Lipshitz, "Stereophonic microphone techniques... are the purists wrong?," *J. Audio Eng. Soc.*, vol. 34, no. 9, pp. 716–744, 1986.

[18] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am*, vol. 93, no. 5, May 1993.

[19] J. Chowning, "The simulation of moving sound sources," *J. Audio Eng. Soc.*, vol. 19, no. 1, pp. 2–6, 1971.

[20] D. M. Leakey, "Some measurements on the effect of interchannel intensity and time difference in two channel sound systems," *J. Acoust. Soc. Amer.*, vol. 31, no. 7, pp. 977–986, Jul. 1959.

[21] J. C. Bennett, K. Barker, and F. O. Edeko, "A new approach to the assessment of stereophonic sound system performance," *J. Audio Eng. Soc.*, vol. 33, no. 5, pp. 314–321, May 1985.

[22] V. Pulkki, "Virtual source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, Jun. 1997.

[23] V. Pulkki, M. Karjalainen, and J. Huopaniemi, "Analyzing virtual sound source attributes using a binaural auditory model," *J. Audio Eng. Soc.*, vol. 47, no. 4, pp. 203–217, Apr. 1999.

[24] A. Härmä and K. Palomäki. HUTear—A free Matlab toolbox for modeling of auditory system. presented at Proc. Matlab DSP Conf. [Online]. Available: http://www.acoustics.hut.fi/software/HUTear/

[25] R. Patterson, K. Robinson, J. Holdsworth, D. Mckeown, C. Zhang, and M. H. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, L. D. Y. Cazals and K. Horner, Eds, Oxford, U.K.: Pergamon, 1992, pp. 429–446.

[26] B. C. J. Moore, R. W. Peters, and B. R. Glasberg, "Auditory filter shapes at low center frequencies," *J. Acoust. Soc. Amer.*, vol. 88, no. 1, pp. 132–140, Jul. 1990.

[27] T. C. T. Yin, P. X. Joris, P. H. Smith, and J. C. K. Chan, "Neuronal processing for coding interaural time disparities," *Binaural and Spatial Hearing in Real and Virtual Environments*, pp. 399–425, 1997.

[28] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psych.*, vol. 61, pp. 468–486, 1948.

[29] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Heidelberg, Germany: Springer-Verlag, 1990.

[30] B. C. J. Moore, "A model for the prediction of thresholds, loudness, and partial loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–240, 1997.

[31] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Amer.*, vol. 104, no. 5, pp. 3048–3058, Nov. 1998.

[32] G. Theile and G. Plenge, "Localization of lateral phantom sources," *J. Audio Eng. Soc.*, vol. 25, no. 4, pp. 196–200, Apr. 1977.

[33] B. L. Cardozo, "Adjusting the method of adjustment: SD vs. DL," *J. Acoust. Soc. Amer.*, vol. 37, no. 5, pp. 768–792, May 1965.

[34] W. Hartmann, "Localization of sound in rooms," *J. Acoust. Soc. Amer.*, vol. 74, no. 5, pp. 1380–1391, 1983.

Ville Pulkki received the M.Sc. and D.Sc. (Tech.) degrees in acoustics, audio signal processing, and information sciences from Helsinki University of Technology, Helsinki, Finland, in 1994 and 2001, respectively.

From 1994 to 1997, he was a full-time student at department of Musical Education, Sibelius Academy. In his doctoral dissertation he developed vector-base amplitude panning (VBAP), which is a method to position virtual sources to any loudspeaker configuration, and studied its performance with psychoacoustic listening tests and with modeling of auditory localization mechanisms. The VBAP method is widely used in multichannel virtual auditory environments and in computer music installations. His research activities cover methods to reproduce spatial audio and methods to evaluate quality of spatial audio reproduction. He has also worked on diffraction modeling in interactive models of room acoustics.

Toni Hirvonen was born in Vaasa, Finland, in 1976. He received M.Sc. (E.E.) degree from Helsinki University of Technology (HUT), Helsinki, Finland, in 2002.

Since 2003, he has been working in the HUT Laboratory of Acoustics and Audio Signal Processing conducting postgraduate research and studies. His main research topics are spatial hearing, auditory modeling, as well as audio reproduction.