Analyzing Virtual Sound Source Attributes Using a Binaural Auditory Model*

VILLE PULKKI, *AES Member*, **MATTI KARJALAINEN**, *AES Member*, **AND JYRI HUOPANIEMI****, *AES Member*

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

Virtual sound sources created with different techniques are evaluated using a computational model of auditory perception. A binaural model estimates the perceived localization cues and coloration. The cues of real and virtual sound sources are compared and the spatial attributes of the sources are discussed. The model predicts many known phenomena in loudspeaker listening and shows good agreement with empirical listening results.

0 INTRODUCTION

The quality of reproduced sound is most reliably evaluated by formal listening tests. The listening tests are, however, expensive and take a long time to conduct. To avoid them, a number of objective measures have been developed in order to assist in product development or to characterize known degradations from perfect sound.

There is an increasing need, however, to formulate objective measures based on the modeling of auditory perception that could combine the best sides of the two approaches—the relevance of listening tests and the efficiency as well as repeatability of objective measures. Such perceptual, auditory, or psychoacoustic modeling approaches are becoming an important tool for audio.

Computational modeling of the auditory system has already been applied successfully to some demanding problems, for example in audio and speech processing. An important path of research, related to the building of objective, perceptually based measures of sound quality, can be traced through papers such as [1]–[4], finally resulting in an ITU-R standard of objective sound quality (Task Group 10/4). The idea is to compare the original signal and the signal from the output of a device under test in a perceptual spectral or time–frequency domain. Based on such distance measures an index can be computed that correlates with listening test results, such as the mean opinion score (MOS) values. This work has not, however, been related to binaural aspects of sound quality.

The modeling of binaural hearing by computational means is a similar problem as in the advanced modeling of human perception in general. It has been hard to find explicit formulas and rules to describe the complex nonlinear behavior of these systems. Many traditional models of binaural hearing deal only with the lateralization of sound based on interaural cross correlation [5]–[7]. Recently there has been some progress in including directional features in more integrated ways [8]–[16], at least in preliminary formulations.

Relatively little has been done, however, in the computational modeling of spatial attribute perception for quality estimation of reproduced sound. The main reason for this is that binaural modeling is a task that includes difficult subproblems such as the precedence effect [17], [18]. Also, traditional multichannel sound systems include complex signal channels in the recording process, not only the reproduction subsystem. Synthetic audio is a newer paradigm, related to so-called virtual acoustics, where carefully controlled virtual sound sources can be synthesized using real sound sources such as loudspeakers or headphones. Methods for three-dimensional audio, such as crosstalk cancelation systems, binaural techniques, and auralization, have gained increasing popularity. For such synthetic virtual sound sources we can more easily define an ideal condition of reproduction.

An important paper on the evaluation of localization attributes in sound reproduction using auditory modeling was published by Macpherson [9]. The study contained the development of a binaural auditory model, including

* Presented at the 104th Convention of the Audio Engineering Society, Amsterdam, The Netherlands, 1998 May 16–19; revised 1998 July 9 and 1999 March 10.

** Also with Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA 94305-8180, USA, and Nokia Research Center, Speech and Audio Systems Laboratory, FIN-00045 Nokia Group, Finland.

©Audio Engineering Society, Inc. 1999

simplified processing of the precedence effect, and a table-matching procedure of directional cues to yield an estimate of the spatial distribution of the auditory event as a function of the azimuth angle. The model was applied to different sound recording and reproduction conditions, both anechoic and reverberant, and it showed good agreement with subjective perception, at least qualitatively.

Mac Cabe and Furlong [19] conducted also interesting work on the spatial attribute evaluation problem. Their method was based on measuring the ear input signals with a dummy head and calculating the localization cues with a simplified auditory model.

Since the publication of these papers the knowledge of spatial hearing and the art of auditory modeling as well as applications, especially in binaural technology, have made progress. Also, the computational power of modern computers allows much more demanding models to be run efficiently. Thus there is a need for revisiting the model-based evaluation of spatial sound reproduction.

The present study originated from the problem of evaluating the quality of spatial attributes of virtual sound sources created using the vector base amplitude panning (VBAP) technique published by Pulkki [20]. When amplitude panning of a signal to two or more loudspeakers is used in order to create a virtual sound source somewhere in a sector spanned by the loudspeakers, the degradations of the virtual source, compared to a real souce (a single loudspeaker), are the directional error or blur of the spatial image and the spectral coloration of the tone quality. A similar problem formulation emerges in binaural reproduction when using head-related transfer functions (HRTFs). This applies to any other multichannel technique as well, although the creation of virtual sources may not be as directly defined a problem in such cases.

In this study we have taken a new step toward the evaluation of spatial quality attributes of reproduced sound using the perceptual modeling approach. Our viewpoint here is primarily synthetic spatial audio, not recorded sound reproduction. For simplicity, only the direct sounds from sources to the listener are considered. The binaural model used is as minimalistic as possible without discarding the most important directional and timbral cues of virtual source perception. A gammatone filter bank and a simplified hair cell model are used together with loudness and cross-correlation computation. From the application point of view we first explored the most obvious case, the two-loudspeaker amplitude panning corresponding to normal stereo. Some special cases of interest are included, such as antiphase signals in stereophonic listening. Also, HRTF-based binaural reproduction is studied in order to see how well the binaural modeling approach describes the perceived quality attributes to HRTF filter designs. Further work will be needed to render the approach more instrumental and mature for practical purposes.

This paper is organized as follows. First we describe the binaural perceptual model that has been used in our tests. In Section 2 we review the main methods for virtual source creation: amplitude panning, time-delay panning, and HRTF processing. In Section 3 we describe and analyze the simulations. Stereophonic listening has been simulated with different panning methods and different head directions. The virtual sources were created using amplitude panning, time-delay panning, or HRTF processing.

1 BINAURAL PERCEPTUAL MODEL

Spatial and directional hearing has been studied intensively; for overviews, see, for example, [17], [21]. 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 due to wave propagation time difference (primarily below 1.5 kHz) and the shadowing effect of the head (primarily above 1.5 kHz), respectively. In the median plane, where the distances to both ears are equal, the ITD and ILD values are close to zero. Other effects, such as spectral cues and head movements, are considered to carry elevation and front–back information in the median plane. Spatial discrimination is difficult also in so-called cones of confusion, where ILD and ITD vary only slightly due to the nonsymmetry of the head. A cone of confusion can be approximated with a cone that has its symmetry axis in the line passing through the listener's ears, as illustrated in Fig. 1.

Many mathematical and computational models of directional hearing have evolved from coincidence and cross-correlation principles proposed by Jeffress [5]. It has been shown that the interaural cross correlation (IACC) conveys elevation and front–back discrimination ability in the median plane [22] although it is not clear in detail how the auditory system exploits this information.

Since Jeffress, several models have been published that extend the modeling of directional and spatial hearing using combined IACC and ILD [8], [11], the cocktail-party effect [12], and inclusion of simple models of the precedence effect [15]. In fact the precedence effect, that is, the suppression of early delayed versions of the direct sound in source direction perception, is found to be a complex phenomenon that is difficult to model in detail [17].

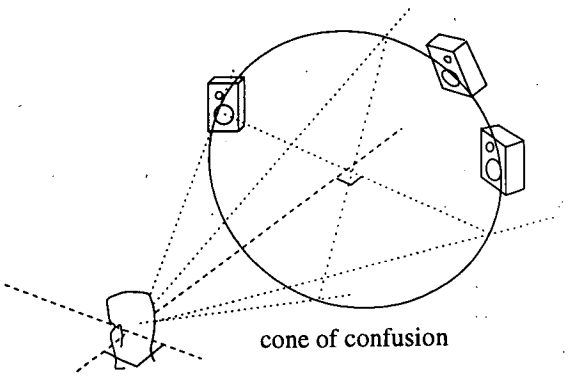


Fig. 1. Cone of confusion. ITDs and ILDs of sound sources vary only little in cone.

©Audio Engineering Society, Inc. 1999

A general drawback of advanced binaural models has been their computational complexity, which does not encourage full-scale experimentation. Thus in order to carry out extensive studies in binaural modeling of reproduced sound perception, some simplifications 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 does not have to be modeled. Since this effect is most prominent in attacks and onsets, we have chosen our test signal to be pink noise, which yields a spectrally balanced excitation to the auditory system, lacking major transients in its temporal envelope which would produce the precedence effect.

1.1 Implemented Binaural Auditory Model

The filtering effect of the external ear was modeled by measured KEMAR[1] (Knowles Electronic Manequin for Acoustic Research) HRTFs [23]. This is done by filtering the test signal with a digital filter that models the measured HRTFs between the ears and the sound source. The effect of the middle ear was not taken into account, since it influences the signal prominently at low frequencies, and the lowest frequency used here was as high as 200 Hz. However, its effect would have been symmetrical to both ears, so it would not change the ITD and the ILD.

The inner-ear frequency selectivity was modeled using a gammatone filterbank (GTFB) [24] of 42 bandpass ERB (equivalent rectangular bandwidth) channels (Fig. 2). The frequency scale in the illustrations in this paper is always the ERB scale, which represents the human auditory pitch scale. The parameters of the filter bank were as defined in Patterson [25].

The hair cell and auditory nerve behavior was simulated by half-wave rectification and 1-kHz low-pass filtering of the output of the gammatone filters. No adaptation model was included since our excitation signals were stationary (pink noise).

The interaural cross correlations are computed by

$$\Phi(\tau) = \int_{t=0}^{t_0} x_l(t)x_r(t-\tau)dt \quad (1)$$

where x_l and x_r are the signals of the left and right ears, respectively, t is the time parameter, τ is the time difference between ear signals, and t_0 is the length of the time window. This implements a rectangular time window starting at zero and ending at t_0. In the auditory system the time window is not rectangular. However, because we use stationary signals, the shape of the window has no influence on the result.

The range of the time difference was set to $\tau = [-1.1, +1.1]$ ms. The cross correlation was computed for each bandpass channel pair to yield the corresponding IACC (Fig. 3). An example of a set of IACCs is shown in Fig. 4 as a function of τ and the ERB channel. Typically the position of maximum in a single IACC curve indicates the ITD at the corresponding frequency. This is used to derive a simple estimate of the ITD as a function of the ERB channel frequency, as shown in Fig. 5.

Loudness values L, in sones, are computed for each ERB channel and each ear as

$$L = \sqrt[4]{\langle x^2 \rangle} \quad (2)$$

where $\langle x^2 \rangle$ is the time average of the signal power. The fourth root approximates the exponent of 0.23 used in [26]. The loudness values of the respective ERB channels are summed to estimate the composite loudness values. The loudness levels L_L (in phons) [26] are computed from the loudness values using the equation

$$L_L = 40 + 10 \log_2 L . \quad (3)$$

[1] KEMAR is a trademark of Knowles Electronics, Inc.

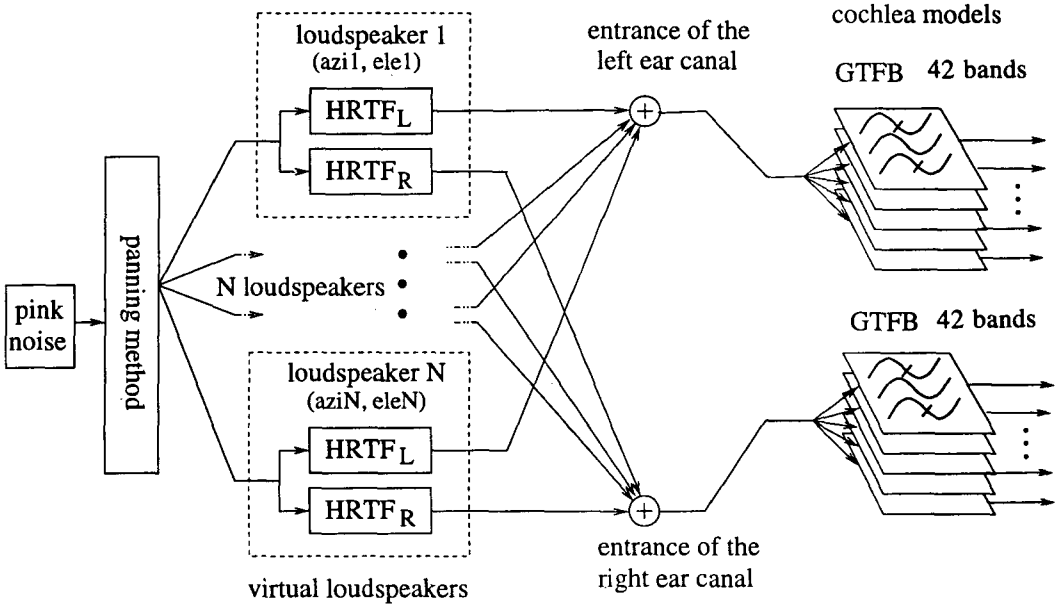


Fig. 2. Peripherical part of binaural model.

©Audio Engineering Society, Inc. 1999

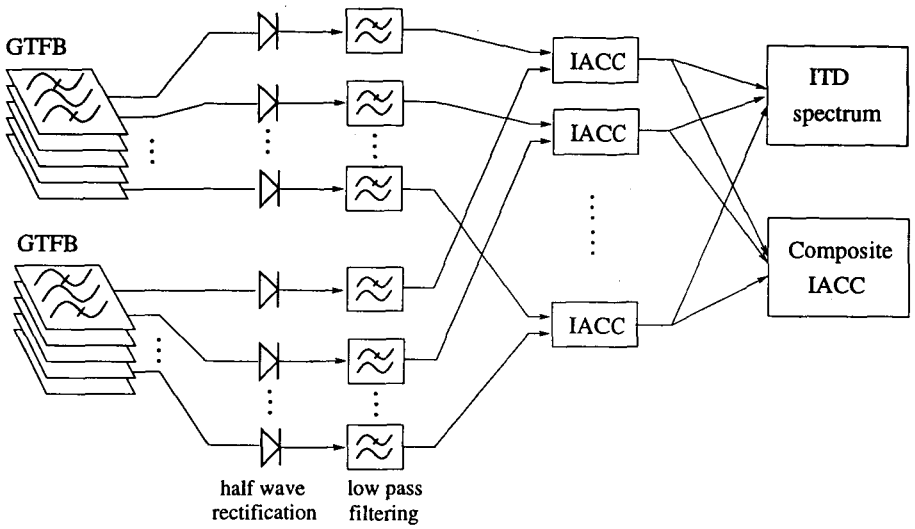


Fig. 3. Part of binaural model yielding ITD and IACC.

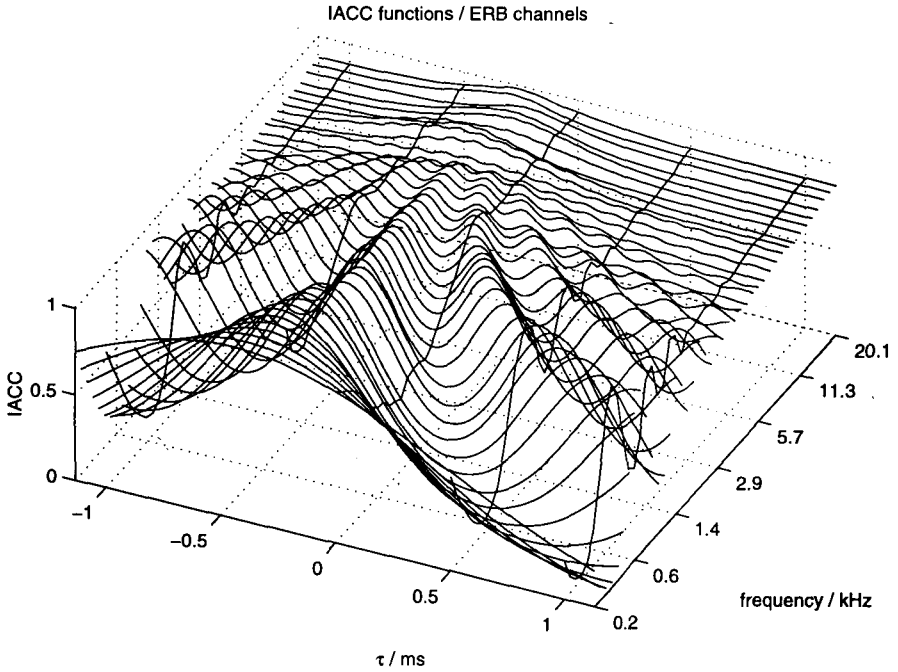
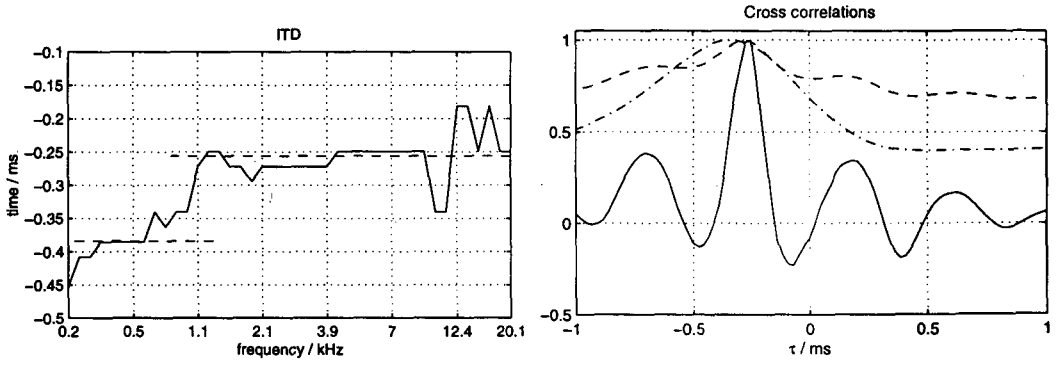


Fig. 4. Simulated IACC functions for real source at position (30°, 0°). For graphical reasons functions have been normalized.



(a) (b)

Fig. 5. Simulated ITD functions and cross correlations of real source at position (30°, 0°). (a) ——— simulated ITD; - - - ITD calculated with spherical model of head. (b) ——— cross correlation between left and right inputs; - - - sum of band cross correlations; — · — sum of band cross correlations for ERB channels below 1.5 kHz.

©Audio Engineering Society, Inc. 1999

The loudness levels of both ears calculated at each ERB channel are added together, which yields the composite (binaurally perceived) loudness level (CLL) spectrum (Fig. 6). An example of a CLL spectrum is shown in Fig. 7(a).

The difference of the loudness level spectra of the left and right ears is used as an ILD spectrum (Figs. 6 and 7(b)), also computed as a function of the ERB channel frequency.

In the following analyzes the IACC, ITD, and ILD curves are depicted to characterize directional cues and the CLL curve is used to represent perceived spectrum.

It is hypothesized that about 1 phon (dB) is the just noticeable difference (JND) in the CLL, that is, coloration between a real and a virtual sound source. There are no well-defined criteria, however, for the JND of the perceived source direction. The curves of the directional attributes of the virtual and the reference real sources are compared to estimate the distortion of the perceived direction. It remains a challenge for future research to derive a single attribute as a reliable estimate of the perceived direction and to prove its consistence with results of listening experiments.

The MATLAB source codes of the binaural model are available on a web site [27].

2 VIRTUAL SOUND SOURCES

A virtual sound source is an auditory perception which appears in a location in the auditory space which does not correspond to a real sound source. Virtual sound sources can be created with two or more loudspeakers, or with headphones. When a virtual source is created with loudspeakers, the mechanism relies on summing localization [17]. Sound signals are applied to loudspeakers and the signals may have amplitude or time differences, or they may be HRTF processed. The signals arriving from each loudspeaker to each ear are summed at the entrances of the ear canals. The summed signals that enter the auditory system contain information that creates the perception of a virtual source.

2.1 Amplitude Panning Techniques

Amplitude panning (intensity panning) is the most often used panning method due to its robustness. Two or more loudspeakers are placed in different directions and at equal distances from the listener. The same sound signal but with different amplitudes is applied to loudspeakers. The listener perceives a virtual source the direction of which is dependent on the ratio of the amplitudes.

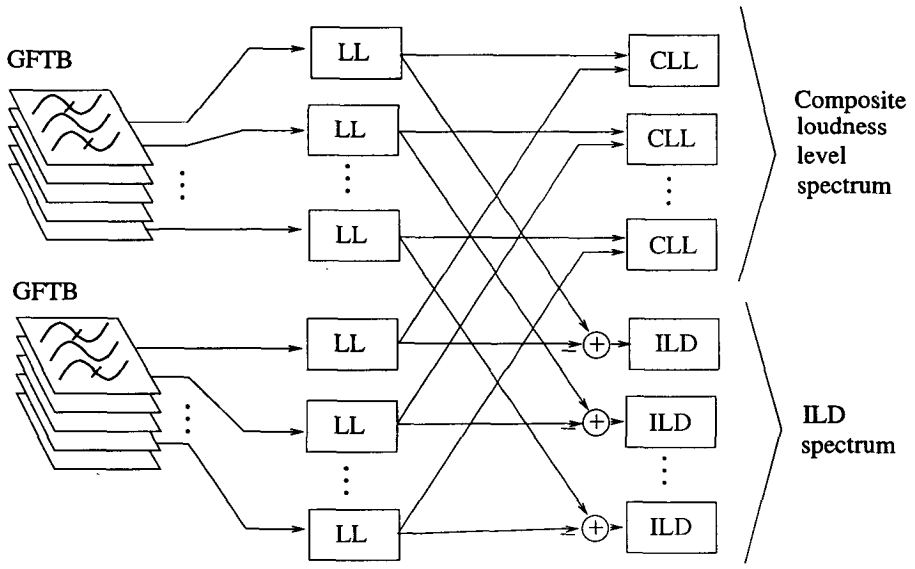
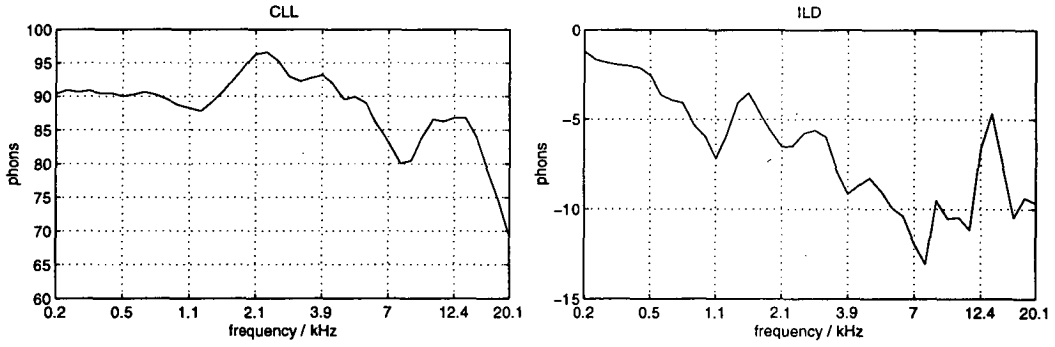


Fig. 6. Modeling of ILD and CLL. LL—loudness level calculation.



(a) (b)

Fig. 7. Simulated CLL and ILD spectra as functions of frequency for real source at position (30°, 0°).

©Audio Engineering Society, Inc. 1999

In most studies amplitude panning has been investigated in a standard stereophonic listening configuration. The amplitude panning techniques can also be applied in other configurations which may have any number of loudspeakers in any positioning [17, p. 216], [20].

At low frequencies, virtual source formation is based on summing localization. The head does not shadow the propagating signals prominently, thus a signal arrives from each louspeaker to each ear canal, forming a new signal for the left and right ear. Generally the amplitude difference of the loudspeakers is changed to a phase difference between ears, which yields the relatively good performance of amplitude panning.

At high frequencies the summing localization may not be valid due to the shadowing of the head. The strong direction- and frequency-dependent effects at high frequencies between the pinna and the arriving sound signals may also distort the localization cues. These phenomena affect mostly the ILD cue, which is also most salient at high frequencies. We may thus estimate that the ILD spectrum of the virtual source may behave unnaturally and yield a spatially spread virtual source.

However, under optimal conditions, for example, the standard stereophonic configuration, shadowing of the head causes that at high frequencies the signal arriving at each ear emanates mostly from the ipsilateral loudspeaker. Thus the condition is quite close to headphone listening: the amplitude level difference of loudspeaker signals is more or less turned to ILD.

2.1.1 *Amplitude Panning in Stereophonic Listening*

Amplitude panning is most often applied to two loudspeakers in a standard stereophonic listening configuration, as depicted in Fig. 8. Loudspeakers 1 and 2 are placed in front of the listener with an aperture of 60°. A signal of different amplitude is applied to each loudspeaker, and can be formulated as

$$x_i(t) = g_i x(t), \qquad i = 1, 2 \tag{4}$$

where $x_i(t)$ is the signal to be applied to loudspeaker i, g_i is the gain factor of the corresponding channel, and t is the time parameter.

It would be useful to know a relation between the gain factors and the perceived virtual source direction. Many approaches exist. Blumlein proposed the famous sine law in the 1930s [28], and Bauer reformulated it in phasor form [29]. In it the wave propagation time difference is taken into account, but the shadowing effect of the head is neglected. The sine law is presented as

$$\frac{\sin\varphi}{\sin\varphi_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{5}$$

where φ is the perceived angle of a virtual source and φ_0 is the loudspeaker base angle, as in Fig. 8. The equation is valid only when the frequency is below 600 Hz and when the listener's head is pointing directly forward. In the equation it is also assumed that the elevation is 0°. The equation does not set limitations on φ, but in most cases its value is set to satisfy $|\varphi| \leq \varphi_0$. If $|\varphi| > \varphi_0$, amplitude panning will produce antiphase loudspeaker signals, which may distort the virtual source [17], as shown in Section 3.2.3.

If the listener is facing toward the virtual source, the tangent law is more correct [30],

$$\frac{\tan\varphi}{\tan\varphi_0} = \frac{g_1 - g_2}{g_1 + g_2}. \tag{6}$$

This equation has the same limitations as the sine law.

2.2 Time-Delay Panning

When a constant delay is applied to one loudspeaker in stereophonic listening, the virtual source is perceived to migrate toward the loudspeaker that radiates the earlier sound signal [17]. When the delay is approximately 1.0 ms, the maximum effect is achieved.

In tests with different signals it has been found, however, that time panning is very dependent on frequency [31], [32]. In stereophonic microphone techniques that realize time panning it has been found that the resulting virtual source is unstable and may move as a function of frequency. Time panning is not widely used to position sources to desired directions; rather it can be used when some special effects are created.

2.3 HRTF Techniques

Amplitude and time-delay panning techniques are used in audio and music technology due to their simplicity, computational efficiency, and relatively good performance. Another viewpoint to panning is to model the physical characteristics of the human ear. It is possible to use measured or modeled free-field-to-ear transfer functions, HRTFs, for creating virtual sound sources in headphone or loudspeaker playback [33], [34]. From the point of view of creating the virtual source, HRTF processing (also called binaural processing or three-dimensional sound) is a complex amplitude and phase panning technique, manipulating the source signal in such a way that a virtual image in the three-dimensional space is created.

This method, when applied to loudspeaker listening,

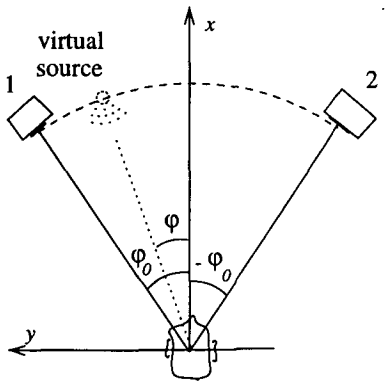


Fig. 8. Standard stereophonic listening configuration.

©Audio Engineering Society, Inc. 1999

also enables the creation of virtual sources outside the loudspeaker aperture. Thus the possibilities of virtual source imaging with HRTF techniques are superior to those of traditional panning, but there are also many drawbacks. Methods can be computationally demanding, HRTFs are generally highly individual and performance is not guaranteed for all listeners, there are different processing methods for headphone and loudspeaker reproduction, and the listening area in loudspeaker reproduction is very limited. In the following, methods for virtual source generation using HRTF processing are discussed.

2.3.1 HRTF Processing for Stereophonic Listening

As discussed in the previous section, HRTF-based panning methods fall into two main categories according to the method of reproduction—headphone and loudspeaker configurations. Fig. 9 illustrates the differences between binaural processing for the two listening conditions. In the case of headphone presentation, the HRTFs may be directly applied to create a virtual source (if proper equalization of measurement system and headphones has been carried out). In the example of Fig. 9(a) a monophonic time-domain signal x_m [short for $x_m(n)$] is filtered with two HRTF filter approximations $\boldsymbol{H}_l(z)$ and $\boldsymbol{H}_r(z)$ to create a single virtual source. Advantages of binaural processing are that the listening facilities and positions are not critical. On the other hand, individual HRTFs are to be used to yield natural three-dimensional panning, and care must be taken in the equalization and placing of headphones. Methods for HRTF filter design are discussed in [35].

In loudspeaker HRTF synthesis [Fig. 9(b)] signals $\hat{x}_l$ and $\hat{x}_r$ (processed binaural signals) are applied to the loudspeakers. This introduces the direction-dependent loudspeaker-to-ear transfer functions $\boldsymbol{H}_i(z)$ and $\boldsymbol{H}_c(z)$ (we consider here only the symmetrical listening position, i = ipsilateral, c = contralateral), and these must be taken into account in order to obtain a similar effect as in headphone listening. This calls for crosstalk canceling, because the sound from the left loudspeaker is heard both at the left and the right ears, and vice versa. One possibility is to see loudspeaker HRTF panning as a cascaded process, where HRTF filters are designed and implemented separately from the crosstalk canceling filters. Another alternative is to combine these processes and design crosstalk cancelation systems by using, for example, shuffler structures [36], [37]. The theory of crosstalk canceling for loudspeaker HRTF reproduction was first presented by Schroeder and Atal [38], and has been later investigated by many authors. Loudspeaker-based HRTF systems have two main limitations: 1) critical listening position, and 2) critical listening room conditions. The full spatial information can be retained only in controlled listening conditions, and the "sweet spot" for listening is fairly limited.

3 MODEL-BASED SIMULATION RESULTS

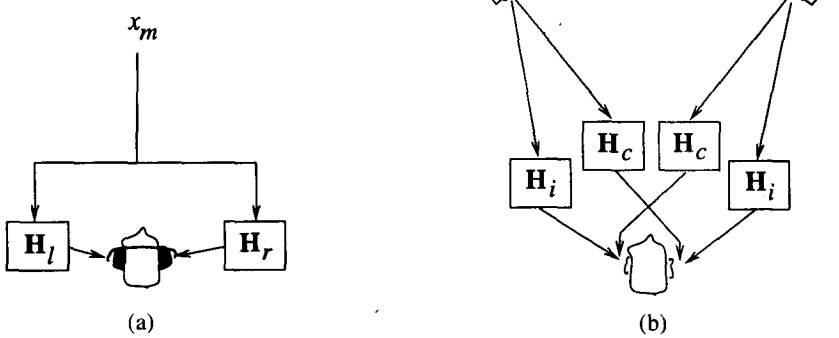
This study is an effort to understand the principles of hearing, which yield perceived virtual sources, when different virtual source production methods are applied. With this approach we can hypothesize the qualities of virtual sources by comparing the simulated virtual source cues with the simulated cues of the corresponding real sources.

The method to simulate virtual and real sources is shown in Fig. 2. The input signal is distributed to the simulated loudspeakers (loudspeakers for short) using the method to be tested. The frequency- and group-delay responses of the loudspeakers are assumed to be flat. The effects of the listener's head, torso, and pinna were introduced to the signal using the measured KEMAR HRTFs [23]. The signal from each loudspeaker was filtered with each ear's HRTFs. This produces the signals that would arrive at the listener's ear canals. The signals could also be recorded using a dummy or a real head.

When multiple loudspeakers are applied, the signals arriving at each ear are summed, resulting in left and right ear signals that are fed to the binaural auditory model. The signal that is applied to the loudspeakers is scaled to keep the amplitude reaching the ears constant. The sampling frequency was 44 100 Hz throughout the simulations.

3.1 Real Sound Sources

The localization cues of real sound sources are simulated. The result of a simulation of a real source at (30°,



(a) (b)

Fig. 9. Binaural processing. (a) Headphone listening. (b) Loudspeaker listening.

©Audio Engineering Society, Inc. 1999

0°) (azimuth, elevation) is shown in Figs. 4, 5, and 7.

We may first consider the simulated ITD. This cue can also be estimated by approximating the human head with a sphere and solving the wave equation. A useful approximation is presented by Kuhn [39],

$$\mathrm{ITD} = \begin{cases} (3a/c)\sin\phi & \text{at low frequencies} \\ (2a/c)\sin\phi & \text{at high frequencies} \end{cases} \quad (7)$$

where a is the radius of the sphere used to model the head, c is the speed of sound, and ϕ is the angle between the median plane and the direction of the sound wave. The transition between low and high frequencies is near 1.0 kHz.

We may now compare this result with our simulations. The average diameter of the KEMAR dummy head is 17.1 cm [40], and we may simulate a loudspeaker at $-30°$ azimuth; thus $\phi = -30°$. The equations would suggest to have an ITD of -0.38 ms at low frequencies and -0.25 ms at high frequencies. The simulated ITD function is presented in Fig. 5(a). It can be seen directly that the approximation of a sphere predicts roughly the simulated ITDs. The disparities may be caused by inaccuracies in the approximation of the human head with a sphere.

The simulated ILD and CLL spectra of this real source (Fig. 7) seem to include consistent data. The ILD spectrum is related to the magnitude spectrum of the interaural transfer function measured with human or dummy heads. Similar curves can be found in [17]. In the CLL spectrum the resonances and antiresonances of the outer ear are shown as peaks and dips.

3.2 Virtual Sources

In this section we create virtual sources with different techniques and compare their cues with cues of a corresponding real source. In simulations where amplitude panning was applied the sine law [Eq. (5)] was used in calculating the gain factors.

In the following we describe the simulations that have been conducted. In the simulations we have used all three main panning methods—amplitude panning, time-delay panning, and HRTF processing. The listening configuration in all cases is the standard stereophonic listening configuration.

3.2.1 Amplitude Panning

The test setup is shown in Fig. 10. The virtual source was panned to the azimuth direction 15° according to the sine law. The simulated ITDs for the virtual source coincide well with the ITDs of the corresponding real sources for frequencies below 1.1 kHz. At higher frequencies the ITD function of the virtual source deviates from the ITD function of a real source. However, in spatial hearing at frequencies above 1.6 kHz the ILD is more important than the ITD [17]. The disparities may, however, cause blur in the perceived direction.

The ILDs of the virtual source deviate prominently from the ILDs of the real sources. Only approximately half the simulated ERB channels have nearly the same ILD values as the real source. Especially between 1 and 3 kHz there exists a large gap between the cues. At frequencies above 3 kHz the ILD cue of the virtual source follows fairly well the ILD cue of the real source.

The disparities between virtual source and real source cues would predict that the virtual source would be spread spatially at frequencies between roughly 1 and 3 kHz.

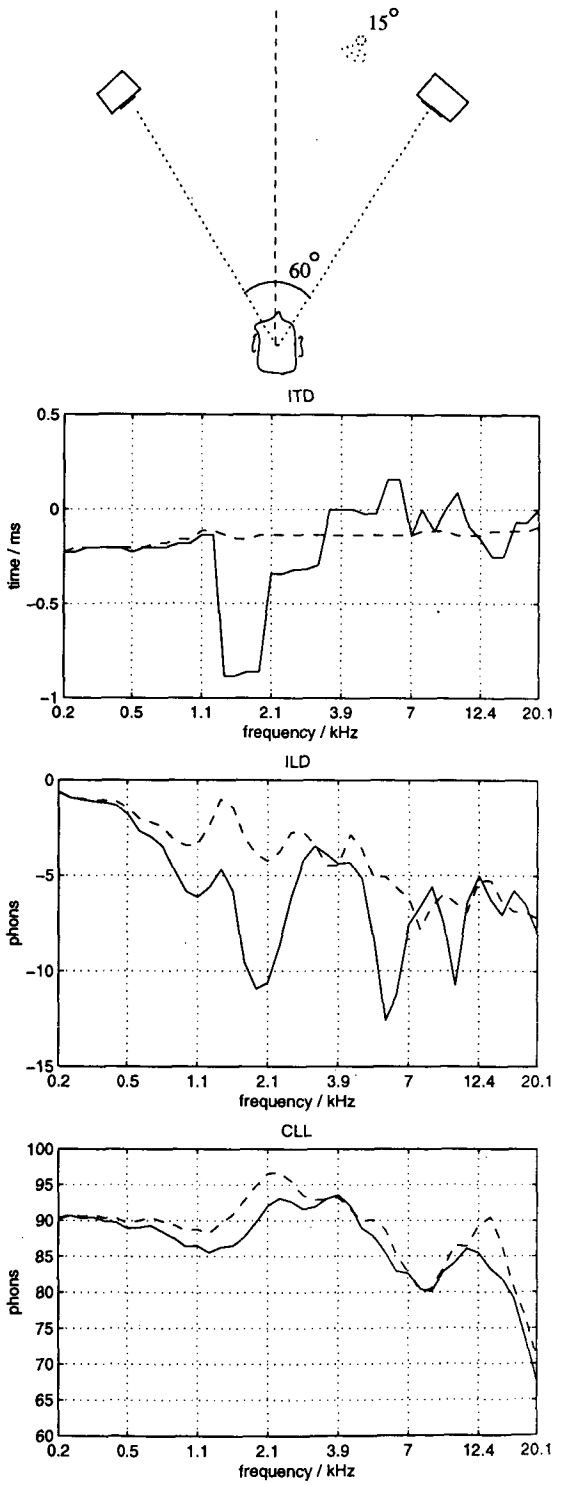


Fig. 10. Simulated cues of amplitude-panned virtual sound source in stereophonic listening. Azimuth angle of virtual source is 15°. - - - real source; —— virtual source.

©Audio Engineering Society, Inc. 1999

The perceived coloration can be estimated from the differences of CLL spectra of the virtual and real sources, which are shown at the bottom of Fig. 10. It can be seen that the loudness level of the virtual source is lower between 400 Hz and 3 kHz when compared to the reference real source. This is natural since the loudspeaker signals arriving at an ear canal cancel each other more at higher frequencies due to smaller wavelengths. At frequencies higher than about 3 kHz the crosstalk is not prominent due to head shadowing; thus the loudness level stays near the real source value. However, the virtual source CLL value deviates from the reference values to some extent. This may happen because of complex wave propagation near the listener's pinna.

3.2.2 Amplitude Panning: Effect of Head Direction

When the listener rotates his or her head in stereophonic listening, a virtual source should remain in its position as a corresponding real source would do. However, the amplitude-panned virtual sources move slightly in the same direction as the listener turns his or her head [29]. This may be perceived as a movement of the virtual source in azimuth or as a constant elevation of the virtual source.

This phenomenon was investigated by simulating stereophonic listening and turning the simulated listener's head 30° and 90°. The results and the system setups are shown in Figs. 11 and 12. In these tests the virtual sources were panned to the middle of the loudspeakers.

From the simulated cues of head direction 30° it can be seen that the disparities between real and virtual source ITDs and ILDs are larger when the listener's head points to the side. The ITD function below 1.1 kHz tends to have smaller absolute values than expected. It proposes that the virtual source would be located nearer the loudspeaker that is closer to the listener's median plane. This agrees with the phenomenon in the real world.

When the listener's head points toward 90°, the loudspeakers are in the same cone of confusion. In this case stable virtual sources cannot be positioned between the loudspeakers [41] (Fig. 12).

The simulated cues are presented in Fig. 12. It can be seen that the ITD function of the virtual source does not follow the cue of the real source at 90°; it rather follows the cue of the real source at 60°. This would suggest that the virtual source would be localized at 60°, or 120°, as happens in the real world.

In the CLL plots it can be seen that the coloration between the frequencies of 400 Hz and 3 kHz decreases when the listener turns his or her head toward 90°. This may be due to decreasing arrival-time differences at one ear from the loudspeakers, which causes the signals to be canceled in smaller amount in this frequency region.

3.2.3 Antiphase Signals

It is a well-known fact that when antiphase signals are applied to stereophonic listening the auditory event splits to multiple events. Especially the low-frequency components are localized indefinitely or inside the listener's head [17].

In our simulations the test signal was applied in antiphase to loudspeakers in a standard stereophonic listening configuration. The resulting cues are shown in Figs. 13 and 14. The ITD behavior is very abnormal. It exceeds 1 ms at low frequencies and its absolute value decreases very fast. The IACC curves of each ERB band

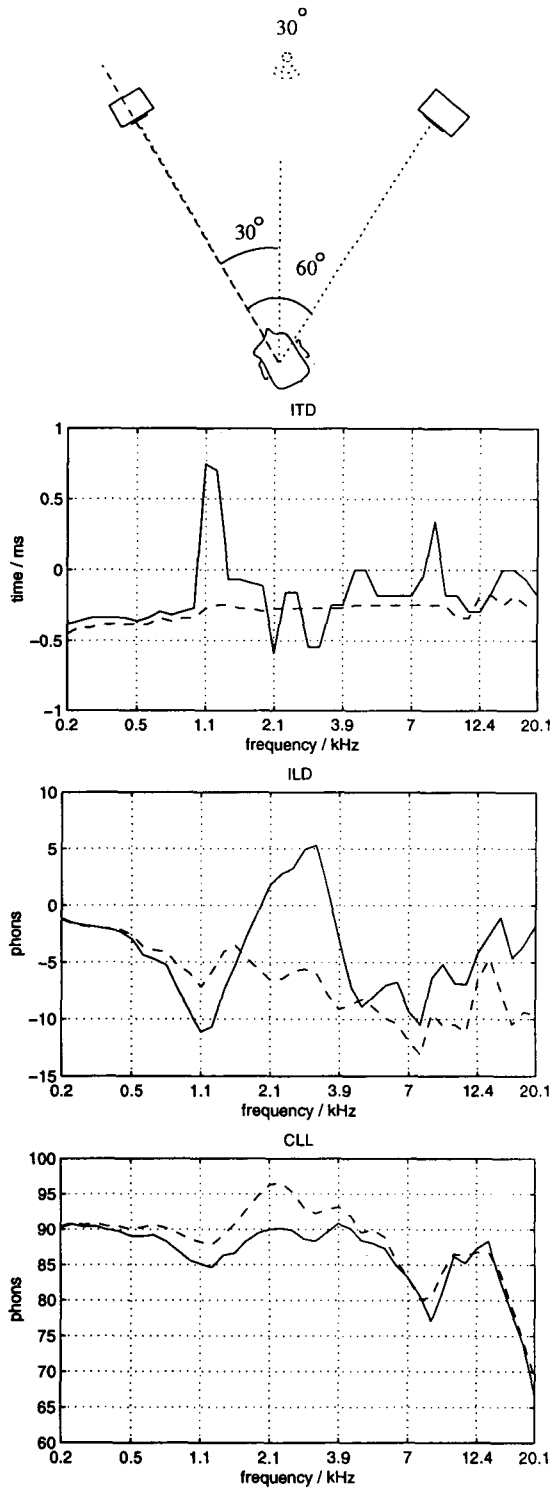


Fig. 11. Simulated cues of amplitude-panned virtual sound source in stereophonic listening. Listener's head is pointing to 30° left; angle of virtual source is 30°. - - - real source; ——— virtual source.

©Audio Engineering Society, Inc. 1999

are symmetrical. They have two equally high peaks which are located at the same absolute value of τ. Either of the peaks could be selected to be the time value of the ITD; the differences in height of the peaks are caused by small computational inaccuracies.

The ILD spectrum follows well the ILD of a real source in the median plane, since it remains zero at all frequencies. This could lead to a stable localization at high frequencies and very indefinite localization at low frequencies.

The CLL spectrum predicts that a prominent cancellation of low frequencies will occur, which actually happens. These results thus explain the phenomenon reported in the literature.

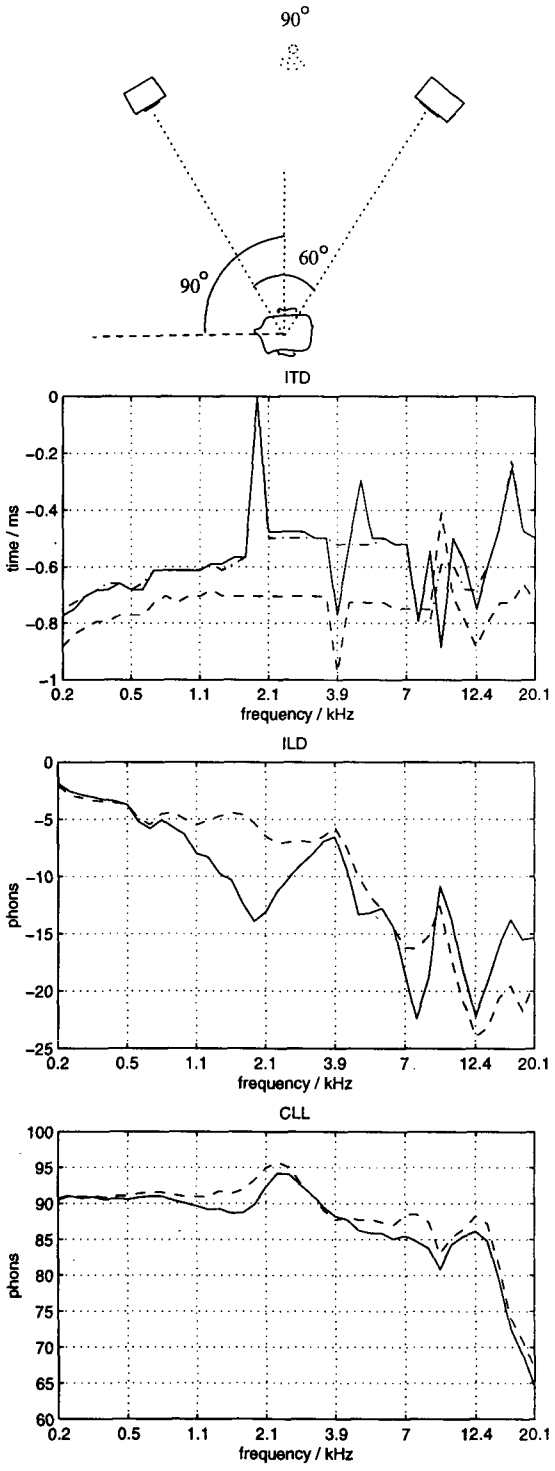


Fig. 12. Simulated cues of amplitude-panned virtual sound source in stereophonic listening. Listener's head is pointing to 90° left; angle of virtual source is 90°. - - - real source at 90°; — · — real source at 60°; ——— virtual source.

3.2.4 Time-Delay Panning

Several time-delay panned virtual sources in stereophonic listening were simulated. A delay of 0.4 ms was applied to one of the loudspeakers in the following case. The results of the simulations can be seen in Fig. 15. The simulations show that the ITD value is near zero at low frequencies and reaches the value of the time delay between the loudspeakers near 1.0 kHz. The ILD oscillates around zero phons and the oscillation amplitude is greater at low frequencies than at high frequencies.

This behavior can be explained as follows. At low frequencies head shadowing does not occur. Thus the loudspeaker signals are summed and the phase and the amplitude of the signal reaching the ear canal are modified. At high frequencies head shadowing neglects the summing effect and the signal that arrives at the ear canal is produced practically only by the ipsilateral loudspeaker. Thus the ILD spectrum and the ITD function are stabilized. Also, the ERB band resolution smoothes the "comb filter effect" of ILD oscillation prominently at high frequencies.

It must be noted that the ITD gets larger values at high frequencies than is possible when the loudspeakers are used separately. This would suggest that the virtual source would be outside the loudspeaker region, which happens in some situations [17].

When the ILD and the ITD are compared, this simulation suggests that the time-panning process produces very ambivalent localization cues. There exists evidence that when conflicting cues are applied at low frequencies, as in this case, the ITD cue is dominant in source localization [42]. This would suggest that the low-frequency image would be split due to curvedness of the ITD function. At high frequencies ITD and ILD cues conflict, which may cause a split image. Based on this information it can be predicted that the localization of time-panned virtual sources is not stable and pointlike, which is actually the case in the real world [32].

3.2.5 HRTF Processing: Effect of Shortening the Filter

An interesting application for a binaural auditory model is to study the fidelity of the auditory image created with HRTF processing. Computationally efficient and perceptually relevant three-dimensional sound processing is very desirable from the point of view of implementation and applicability. Therefore efficient filter design methods for HRTF approximations have been studied and various solutions proposed [35], [43], [37]. It has been shown that digital filter approximations of HRTFs taking into account the nonuniform frequency resolution of the human hearing yield better results at lower filter orders when compared to traditional design methods [35]. The binaural auditory model proposed in

©Audio Engineering Society, Inc. 1999

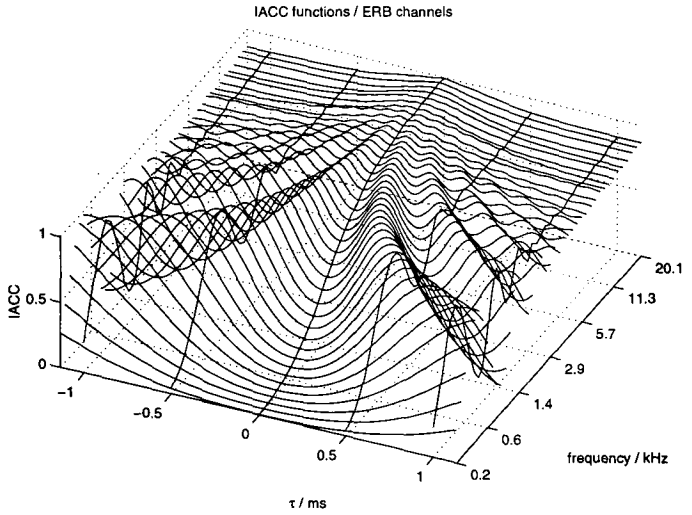


Fig. 13. Interaural cross-correlation functions of ERB bands when antiphase signals are applied to loudspeakers in stereophonic listening.

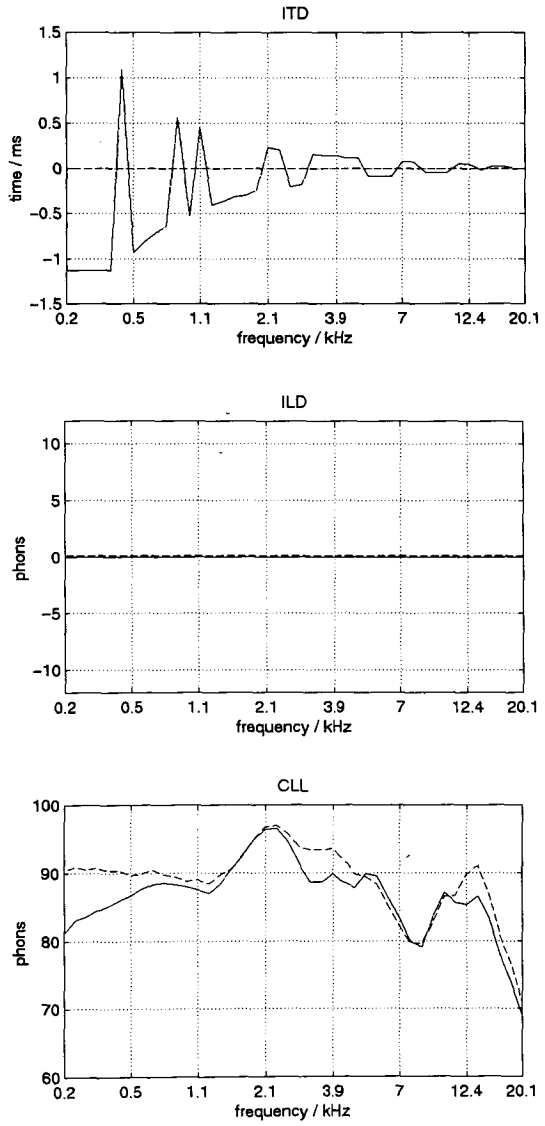


Fig. 14. Simulated localization cues when antiphase signals are applied to loudspeakers in stereophonic listening. - - - real source at 0° azimuth; —— virtual source.

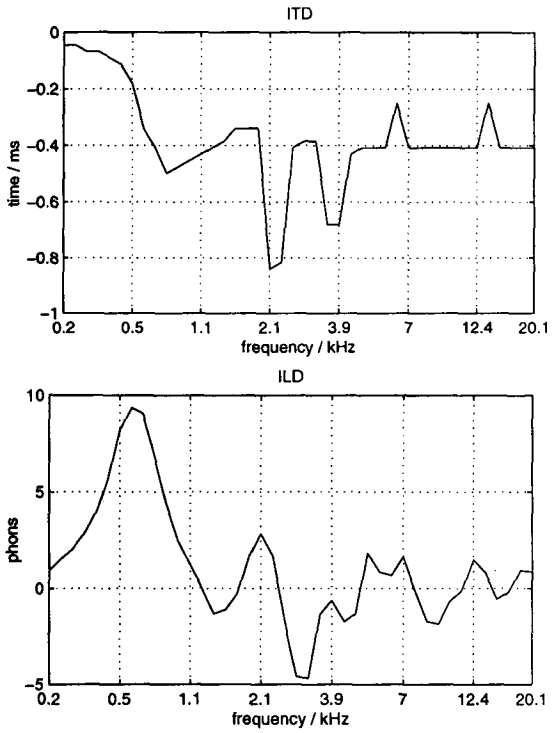


Fig. 15. Time delay of 0.4 ms is applied between loudspeakers of standard stereophonic listening configuration.

this paper could be used for creating a psychoacoustic error measure for HRTF filter design and for analyzing filter approximations as is done in the following. A more detailed study on estimating the HRTF filter design quality is presented in [44].

To illustrate the degradation of an HRTF-processed auditory image, we analyzed HRTFs of varying filter lengths using the binaural auditory model. Fig. 16 shows results for HRTFs at the angle (60°, 0°) approximated with FIR filters (designed using rectangular windowing, that is, truncating the impulse response) of length 88,

©Audio Engineering Society, Inc. 1999

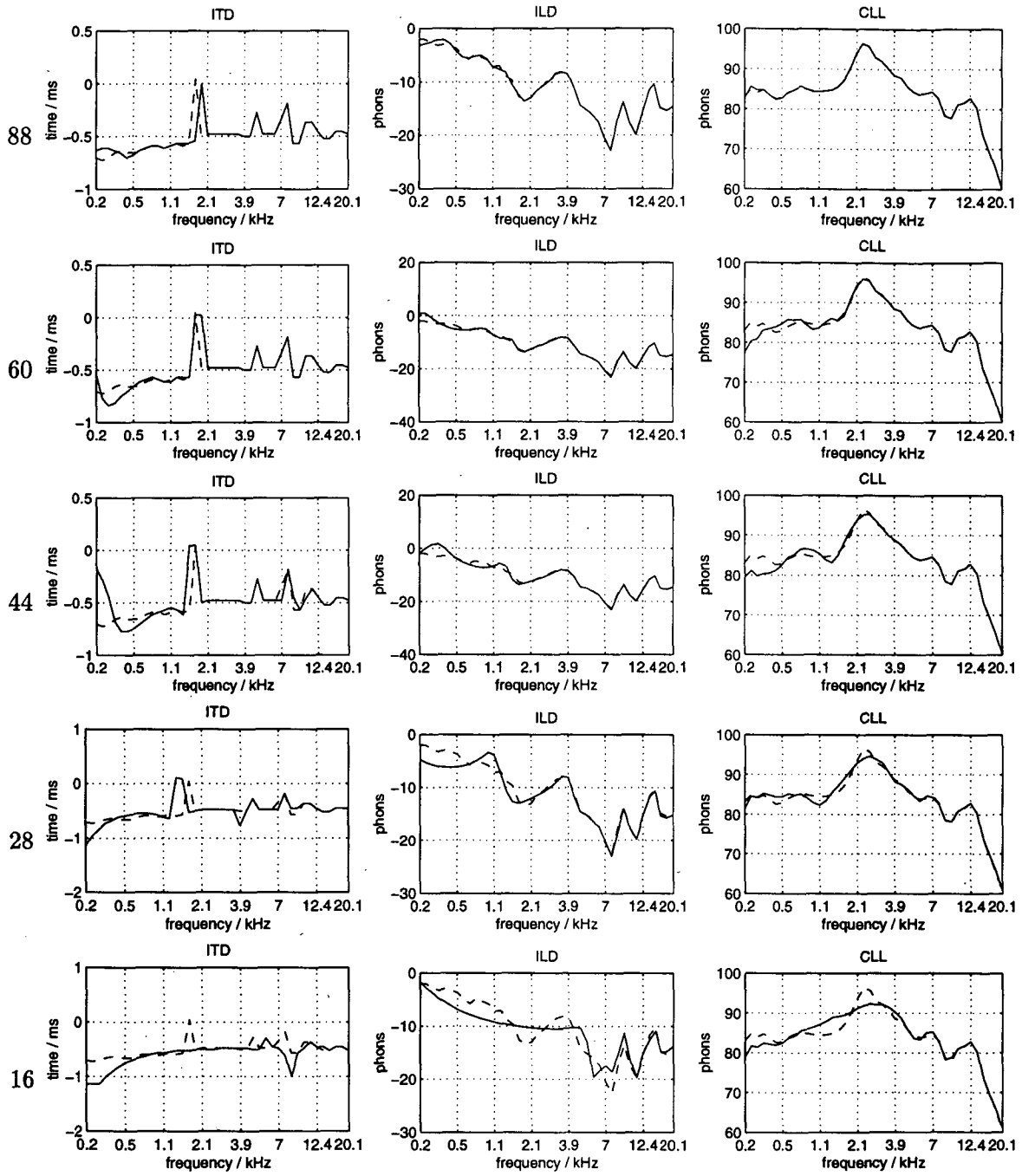


Fig. 16. Effect of shortening the HRTF FIR. Number of taps at left. Virtual source is in azimuth angle 60°. - - - real sources; ——— virtual sources.

60, 44, 28, and 16. The FIR filter was designed by using minimum-phase reconstruction and applying the ITD as a delay line, as in [35]. The first 128 filter coefficients (the impulse response samples) without the ITD are shown in Fig. 17. The results of this simulation are valid for both loudspeaker and headphone listening. This can be stated if we assume crosstalk canceling to be ideal in loudspeaker listening and the equalization and placement of the headphones to be ideal in headphone listening.

It can be seen that the ITD and ILD cues are well preserved down to a length of about 60, that is, less than 12% of 512, which is the length of the filter used when the KEMAR HRTFs were modeled. It can also be

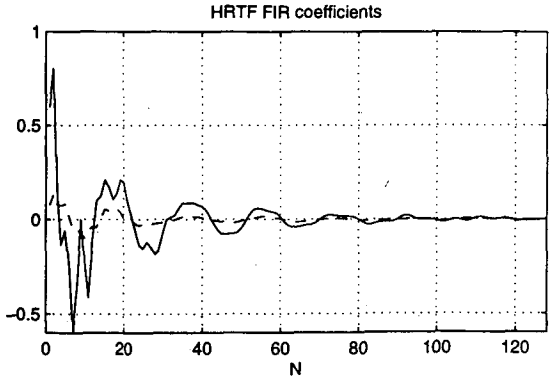


Fig. 17. Minimum-phase FIRs for HRTF processing. ——— left ear; - - - right ear.

©Audio Engineering Society, Inc. 1999

noted that the low-frequency behavior of the shortest filters is nonsatisfactory. Below filter lengths of 28, the auditory image is severely degraded and mislocalization is very likely to occur. In comparison to subjective listening tests carried out using nonindividualized (Cortex MK2 dummy head) HRTFs, these results are in good agreement. The listening test results suggested that 75% of the population (8 subjects) heard no difference between an FIR filter of length 40 and the reference (length 256). The 25% quartile was at length 25 [35].

4 DISCUSSION

The present paper has focused on binaural auditory modeling in order to evaluate the virtual sound source attributes created with loudspeaker and headphone sound reproduction techniques. Partly it parallels what was done in [9], such as evaluating directional cues in normal stereo, antiphase stereo, and so on, but using a different modeling approach. Estimation of source coloration has been added. It also adds the important case of HRTF-based binaural reproduction. The results show good agreement with auditory perception of the corresponding situations, at least qualitatively.

The auditory model used in this study was simplified to be valid only in cases where the precedence effect has minimal influence, that is, no delayed or reverberant versions of direct sound are received. So far no general model of the precedence effect is available. Also no "internal noise" was implemented, which would limit the accuracy of the model to the human performance level.

In the present study the simulated localization cues, that is, ITD and ILD, have been examined separately. A logical step forward would be the derivation of a single representation which could yield the perceived localization in each ERB channel and furthermore the perceived localization blur. In the derivation new listening experiments and computer simulations may be needed.

As a further step, the modeling of elevation perception could be added. Obviously ITD and ILD cues are not enough, especially in the median plane. One possibility is to use more of the information from IACCs since this enables the estimation of elevation [22].

For many practical purposes a computational model should be applicable in realistic environments where reflections from surfaces and reverberations occur. This requires the inclusion of more advanced temporal processing, in particular the precedence effect. Macpherson [9] has shown a simplified precedence processing scheme that yields useful results. Yet the phenomenon is far from well understood and modeled.

As a final goal from a practical point of view a binaural computational model should predict subjective quality measures of spatial sound quantitatively with good enough accuracy to be a useful tool in quality assessment, such as in product development. Since the human performance in spatial perception is personalized by learning, a technique to match this is to use learning principles, such as artificial neural networks, to calibrate a spatial sound measurement sytem according to human performance.

5 CONCLUSIONS

A simplified binaural auditory model was constructed based on known principles of the directional hearing process. This auditory model produces as its output the estimated main localization cues and coloration of audio signals. The signals can be recorded from a real or a dummy head. In this paper we have calculated them using dummy-head HRTFs.

The auditory model was used to analyze various phenomena in loudspeaker listening. In stereophonic listening the model was able to predict that the localization error of virtual sources is greater at high frequencies. The model was able to predict the phenomenon in which the head direction in stereophonic listening yields a shifted or elevated virtual source. Also the model predicted the well-known fact that virtual sources cannot be positioned between two loudspeakers at the listener's side.

The model gave ambiguous cues in time-delay panning and in antiphase stereo listening. In both cases the virtual sources are known to be somewhat spatially spread and diffuse. The model estimated the quality of virtual sources created using HRTF processing, consistently with listening test results.

6 ACKNOWLEDGMENT

The work of Mr. Pulkki was supported by the Graduate School in Electronics, Telecommunications and Automation (GETA). The authors wish to thank Bill Gardner at MIT for making the KEMAR HRTF measurements available to the scientific community and Richard O. Duda at San Jose State University for comments on the auditory model.

7 REFERENCES

[1] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties to the Human Ear," *J. Acoust. Soc. Am.*, vol. 66, pp. 1647–1652 (1979).

[2] M. Karjalainen, "A New Auditory Model for the Evaluation of Sound Quality of Audio Systems," in *Proc. ICASSP-85* (Tampa, FL, 1985), pp. 608–611.

[3] K. Brandenburg, "Evaluation of Quality for Audio Encoding at Low Bit Rates," presented at the 82nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 35, p. 382 (1987 May), preprint 2433.

[4] J. G. Beerends and J. A. Stemerdink, "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio Eng. Soc.*, vol. 40, pp. 963–978 (1992 Dec.).

[5] L. A. Jeffress, "A Place Theory of Sound Localization," *J. Comp. Physiol. Psych.*, vol. 61, pp. 468–486 (1948).

©Audio Engineering Society, Inc. 1999

[6] J. C. R. Licklider, "Audio Frequency Analysis," in C. Cherry, Ed., *Information Theory*, 3rd London Symp. (Butterworths, London, 1961), pp. 253–268.

[7] R. M. Stern and H. S. Colburn, "Theory of Binaural Interaction Based on Auditory-Nerve Data, IV. A Model for Subjective Lateral Position," *J. Acoust. Soc. Am.*, vol. 64, pp. 127–140 (1978).

[8] W. Lindemann, "Extensions of a Binaural Cross-Correlation Model by Contralateral Inhibition, I and II," *J. Acoust. Soc. Am.*, vol. 80, pp. 1608–1630 (1986).

[9] E. A. Macpherson, "A Computer Model of Binaural Localization for Stereo Imaging Measurement," *J. Audio Eng. Soc.*, vol. 39, pp. 604–622 (1991 Sept.).

[10] J. Backman and M. Karjalainen, "Modelling of Human Directional and Spatial Hearing Using Neural Networks," in *Proc. ICASSP-93* (Minneapolis, MN, 1993), pp. I-125–I-128.

[11] W. Gaik, "Combined Evaluation of Interaural Time and Intensity Differences: Psychoacoustic Results and Computer Modeling," *J. Acoust. Soc. Am.*, vol. 94, pp. 98–110 (1993).

[12] M. Bodden, "Modeling Human Sound-Source Localization and the Cocktail-Party Effect," *Acta Acustica*, vol. 1, pp. 43–55 (1993).

[13] C. Lim and R. O. Duda, "Estimating the Azimuth and Elevation of a Sound Source from the Output of a Cochlear Model," in *Proc. 28th Asilomar Conf. on Signals, Systems, and Computers* (Pacific Grove, CA, 1994).

[14] K. D. Martin, "A Computational Model of Spatial Hearing," master's thesis, Massachusetts Institute of Technology, Cambridge, MA (1995).

[15] K. D. Martin, "Echo Suppression in a Computational Model of the Precedence Effect," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA'97* (New Paltz, NY, 1997 Oct.).

[16] D. Nandy and J. Ben-Arie, "An Auditory Localization Model Based on High Frequency Spectral Cues," *Ann. Biomed. Eng.*, vol. 24, pp. 621–638 (1996 Nov.–Dec. 1996).

[17] J. Blauert, *Spatial Hearing*, rev. ed. (MIT Press, Cambridge, MA, 1997).

[18] P. M. Zurek, "The Precedence Effect," in W. A. Yost and G. Gourewitch, Eds., *Directional Hearing* (Springer, New York, NY, 1987), pp. 85–105.

[19] C. J. Mac Cabe and D. J. Furlong, "Virtual Imaging Capabilities of Surround Sound Systems," *J. Audio Eng. Soc.*, vol. 42, pp. 38–49 (1994 Jan./Feb.).

[20] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, pp. 456–466 (1997 June).

[21] R. H. Gilkey and T. R. Anderson, Eds., *Binaural and Spatial Hearing in Real and Virtual Environments* (Lawrence Erlbaum Assoc., Mahwah, NJ, 1997).

[22] J. Backman and M. Karjalainen, "Simulation of Human Spatial Hearing Using Artificial Neural Networks," in *Proc. Int. Congr. of Acoustics (ICA-14)* (Beijing, China, 1992), p. H4-3.

[23] B. Gardner and K. D. Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone," Tech. Rep. 280, MIT Media Lab Perceptual Computing, Cambridge, MA (1994).

[24] M. Slaney, "An Efficient Implementation of the Patterson–Holdsworth Filter Bank," Tech. Rep. 35, Apple Computer, Inc. (1993).

[25] R. D. Patterson, "The Sound of a Sinusoid: Spectral Models," *J. Acoust. Soc. Am.*, vol. 96, pp. 1409–1418 (1994).

[26] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models* (Springer, Heidelberg, Germany, 1990).

[27] V. Pulkki, "MATLAB Source Codes for a Simple Binaural Auditory Model," URL:www.acoustics.hut.fi/~ville/software/auditorymodel/ (1999).

[28] A. D. Blumlein, U.K. patent 394,325 (1931). Reprinted in *Stereophonic Techniques* (Audio Engineering Society, New York, 1986).

[29] B. B. Bauer, "Phasor Analysis of Some Stereophonic Phenomena," *J. Acoust. Soc. Am.*, vol. 33, pp. 1536–1539 (1961 Nov.).

[30] B. Bernfeld, "Attempts for Better Understanding of the Directional Stereophonic Listening Mechanism," presented at the 44th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 21, p. 303 (1973 May).

[31] D. H. Cooper, "Problems with Shadowless Stereo Theory: Asymptotic Spectral Status," *J. Audio Eng. Soc.*, vol. 35, pp. 629–642 (1987 Sept.).

[32] S. P. Lipshitz, "Stereophonic Microphone Techniques . . . Are the Purists Wrong?," *J. Audio Eng. Soc. (Features)*, vol. 34, pp. 716–744 (1986 Sept.).

[33] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-Related Transfer Functions of Human Subjects," *J. Audio Eng. Soc.*, vol. 43, pp. 300–321 (1995 May).

[34] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia* (AP Professional, Cambridge, MA, 1994).

[35] J. Huopaniemi and M. Karjalainen, "Review of Digital Filter Design and Implementation Methods for 3-D Sound," presented at the 102nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 45, p. 413 (1997 May), preprint 4461.

[36] D. H. Cooper and J. L. Bauck, "Prospects for Transaural Recording," *J. Audio Eng. Soc.*, vol. 37, pp. 3–19 (1989 Jan./Feb.).

[37] J. M. Jot, O. Warusfel, and V. Larcher, "Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony," presented at the 98th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 43, p. 396 (1995 May), preprint 3980.

[38] M. R. Schroeder and B. S. Atal, "Computer Simulation of Sound Transmission in Rooms," in *IEEE Conv. Rec.*, pt. 7 (1963), pp. 150–155.

[39] G. F. Kuhn, "Physical Acoustics and Measurements Pertaining to Direction Hearing," in W. A. Yost and G. Gourewitch, Eds., *Directional Hearing* (Springer, New York, NY, 1987), ch. 1, pp. 3–25.

[40] M. D. Burkhard and R. M. Sachs, "Anthropometric Manikin for Acoustic Research," *J. Acoust. Soc.*

©Audio Engineering Society, Inc. 1999

Am., vol. 58, pp. 214–222 (1975).

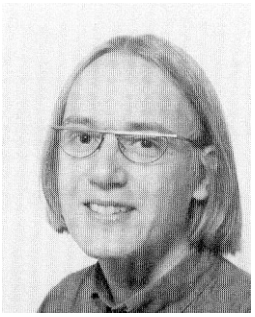
[41] G. Theile and G. Plenge, "Localization of Lateral Phantom Sources," *J. Audio Eng. Soc., (Project Notes/Engineering Briefs)*, vol. 25, pp. 196–200 (1977 Apr.).

[42] F. L. Wightman and D. J. Kistler, "The Dominant Role of Low-Frequency Interaural Time Differences in Sound Localization," *J. Acoust. Soc. Am.*, vol. 91, pp. 1648–1661 (1992).

[43] B. Gardner, "3-D Audio Using Loudspeakers," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA (1997).

[44] J. Huopaniemi, N. Zacharov, and M. Karjalainen, "Objective and Subjective Evaluation of Head-Related Transfer Function Filter Design," *J. Audio Eng. Soc.*, vol. 47, this issue.

THE AUTHORS



Ville Pulkki was born in Jyväskylä, Finland, in 1969. He received the M.Sc. degree in engineering from the Helsinki University of Technology (HUT), Espoo, Finland, in 1994. Between 1994 and 1997 he majored in studies at the Musical Education Department at the Sibelius Academy, Helsinki, Finland. Since 1997 he has been a Ph.D. candidate at HUT.

Mr. Pulkki was with the Laboratory of Computer and Information Sciences and the Neural Networks Research Centre at HUT from 1991 to 1995. He worked as a research assistant and later as a research scientist. Since 1995 he has been working as a research scientist in the Laboratory of Acoustics and Audio Signal Processing at HUT.

His research interests include three-dimensional sound, virtual acoustics, and music. Mr. Pulkki is a member of the AES and the Polar Bear Club of Finland. Email: ville.pulkki@hut.fi.

Biographies for Matti Karjalainen and Jyri Huopaniemi were published in the January/February issue of the *Journal*.

©Audio Engineering Society, Inc. 1999