Phantom materialization: A novel method to enhance stereo audio reproduction on headphones

Jeroen Breebaart and Erik Schuijers

EDICs : AUD-SMCA

***Abstract*— Loudspeaker reproduction systems are subject to a compromise between spatial realism and cost. By simulating loudspeaker reproduction on headphones, the resulting spatial realism is limited accordingly, despite the virtually unlimited spatial imaging capabilities of binaural audio rendering technology. More particularly, phantom imaging as often used for stereo audio material intended for loudspeaker reproduction is subject to various restrictions in terms of loudspeaker positioning in simulated space. As a consequence, phantom imaging should preferably be avoided when simulating virtual loudspeakers over headphones, especially if head tracking is incorporated or if a wide sound stage is desired. A novel method is described to extract phantom sound sources from stereo audio content and convert these to sound sources in a virtual listening environment.**

I. Introduction

During the last two decades, headphones as audio reproduction system have gained significant interest. Since the introduction of the portable cassette players in the early 80's, mobile music players have become extremely popular, especially amongst youngsters. Since then, mobile players have developed quite rapidly. During the mid 80's, the analog, magnetic tape storage medium was replaced by optical storage methods (CDs) with content stored in digital format that significantly improved the audio quality. Around the year 2000, flash memory was introduced to store music in a digital, compressed format, resulting in a significant increase in storage capacity expressed in hours of content. Very recently, mobile players have been extended with video playback capabilities as well. Given the large availability of video content accompanied by surround-sound audio, it can be expected that 3D sound positioning will find its way in the mobile domain in the near future.

Mobility and social constraints in most cases dictate headphones as a reproduction device on mobile players. In contrast to loudspeaker playback, stereo audio content reproduced over headphones is perceived *inside* the head [1]. The absence of the effect of the acoustical pathway from sources at certain physical positions to the eardrums causes the spatial image to sound unnatural, since the cues that determine the perceived azimuth, elevation and distance of a sound source are essentially missing or very inaccurate.

To resolve the unnatural sound stage caused by inaccurate or absent sound source localization cues on headphones, various systems have been proposed to simulate a *virtual loudspeaker setup*. The idea is to superimpose sound source localization cues onto each loudspeaker signal. The technology for such binaural audio rendering on headphones has attracted quite some interest in various research areas and has found its way to consumer electronic devices.

It is well known that accurate synthesis process of virtual sources over headphones is not straightforward. In particular, it has shown to be very difficult to match the perceived and intended sound source position and distance using a generic, non-individualized system [2]. Another challenge is related to the effect of head movements. In normal listening conditions, head rotations will result in a change of stationary sound source positions relative to the orientation of the head. When using headphones, on the other hand, the position of virtual sound sources moves along with head rotations. The absence of the effect of head rotations in virtual auditory displays has a detrimental effect on naturalness [3], [4] and sound source localization abilities [5], [6], [7], [8]. One method to resolve this inconsistency that has been proposed is to measure the orientation of the head by means of a so-called head tracker, and modify the parameters of the synthesis process accordingly to simulate sound source

positions that have stationary physical locations.

For music or movie content, there is an *additional* challenge that is often not mentioned explicitly. By simulation of a *virtual loudspeaker setup*, the constraints and limitations that apply to the suboptimal reproduction system will also limit the spatial image quality in the virtual environment *on top* of the difficulties that exist for binaural synthesis in general.

In the next section, limitations and constraints of loudspeaker reproduction will be discussed in more detail. Subsequently, the general challenges for binaural synthesis will be outlined, followed by a discussion on the consequences of combining the two: the synthesis of virtual loudspeaker setups. Finally, an alternative solution for binaural synthesis is described that resolves some of the problems associated with virtual loudspeaker setups.

II. The stereophonic compromise

The aim of an audio or recording engineer is to provide an artistically meaningful and preferably realistic reproduction of a certain event or recording, given the limitations of a certain target reproduction system. Basically two approaches can be pursued (cf. [9]). The first approach, referred to as *there and then*, aims at an accurate reproduction of the spatial characteristics that were present during the recording. The idea is to create the illusion that the listener is situated in the same room as the recording was made, for example in a concert hall. The second method, referred to as *here and now*, aims at placing various auditory objects in the *reproduction room*, i.e., in the same room the listener is situated during playback.

One of the major challenges that audio engineers are facing is that the target reproduction system is suboptimal and restricted in terms of spatial imaging capabilities. These restrictions follow from various cost and esthetic considerations. The most popular loudspeaker reproduction system is based on two-channel stereophony, using two loudspeakers ideally positioned at +30 and -30 degrees azimuth and at equal distance from the listener. Under these circumstances, the listener is positioned in the so-called *sweet spot*. Given this setup, a technique referred to as amplitude panning can position a *phantom* sound source *between* the two loudspeakers. Thus, in the context of this paper, the term phantom source reflects a perceived sound source at a position in between the loudspeakers that is created using panning techniques. For a phantom source generated using amplitude panning, the employed inter-channel level differences (ICLDs) result in inter-aural time differences (ITDs) and inter-aural level differences (ILDs) at the level of the listener's eardrums that roughly correspond to those of the desired phantom sound source position [10]. However there are indications that the match between sound source localization cues resulting from a phantom source and those from a real source can differ significantly, especially in the mid and high frequency range [10], [11], [12].

Secondly, the area of feasible phantom source positions is quite limited. Basically, phantom sources can only be positioned at an arc between the two loudspeakers. The angle between the two loudspeakers has an upper limit of about 60 degrees [13] and hence the resulting frontal image is limited in terms of width. But even for such limited aperture angle of the speakers, the perceived phantom source position is not fully deterministic and depends on the temporal and spectral characteristics of the source signals [14], [11].

Thirdly, in order for amplitude panning to work correctly, the position of the listener is very restricted. The sweet spot is usually quite small. As soon as the listener moves outside the sweet spot, panning techniques fail and audio sources are perceived at the position of the closest loudspeaker [15].

A fourth restriction applies to the orientation of the listener. If due to head or body rotations both speakers are not positioned symmetrically on both sides of the median plane, the conversion from inter-channel relations to correct inter-aural cues fails and the perceived position of phantom sources is wrong or becomes ambiguous [16], [17], [18], [19], [20].

A fifth potential issue is the spectral coloration that is induced by amplitude panning. Due to dissimilar path-length differences to both ears and the resulting comb-filter effects, phantom sources may suffer from pronounced spectral modifications compared to a real sound source at the desired position [17], [21]. It has also been shown that speech intelligibility is significantly improved for a real (center) loudspeaker source compared to a phantom source, which may also be attributed to the presence of comb-filter effects in the phantom-source case [22].

A final, more generic potential problem of loudspeaker reproduction is that the room-acoustic properties in which the reproduction system is placed will be superimposed on the room-acoustic properties of the recording. In the case of a *here and now* type recording, this is a desired property, but for recordings that were mixed according to the *there and then* principle, the reproduction room acoustics may interfere with those captured by the recording.

III. Virtual challenges

If the panning techniques to create phantom sources on loudspeaker systems are used on headphones, the auditory objects are perceived inside the head [23], [24], [10], [1]. This is because the simple, frequency-independent level and/or time differences between the audio channels only approximate the sound-source localization cues that appear in the real world. Inter-aural level differences and inter-aural time differences depend on azimuth and elevation of a sound source in a complex way, due to path-length differences, diffraction, reflections and head-shadow effects. In order to create a realistic *virtual sound source*, the acoustical pathway from a certain sound source position to both eardrums must be modelled in great detail. The most common method to describe and process virtual sound sources is by means of Head-Related Transfer Functions (HRTFs)[25], [26], [27], [28]. Typically, HRTFs come in pairs (one for each ear) and due to their strong dependence on position, they are typically measured at a very fine spatial resolution (typically 5 to 10 degrees spacing, cf. [29], [30]). Besides dependence on azimuth and elevation, HRTFs vary also as a function of distance [31], [32]. The large amount of data, associated with an HRTF database, and the required processing power are the main challenges in binaural audio processing [2], especially on mobile devices with limited processing power and battery life.

Another difficulty is the dependence of HRTFs on the specific anthropometric properties of each individual [33], [34]. If signals are processed with *individualized* HRTFs in anechoic conditions and a fixed head orientation/position, listeners can not discriminate between a real sound source and a virtual sound source [26], [28]. Despite such accurate reproduction of virtual sources, subjects show localization errors and front/back confusions with generic [35] as well as with individualized HRTFs, especially for positions on the so-called *cones of confusion* that have virtually identical ITD and ILD cues [36], [32], [35]. If non-individual HRTFs are used, sound source localization performance degrades [37], [38], [33]. Interestingly, if head rotations are allowed during sound stimulation using a physical sound source, or if the effect of head rotations is taken into account in the synthesis process of virtual sources, the localization accuracy of human sound source localization increases considerably (especially in terms of front/back reversals), both for individual *as well as* non-individual HRTFs [38], [5], [35]. Moreover, there is some evidence that if head rotations are included in the binaural synthesis, non-individualized and individualized HRTFs give *approximately equal* localization performance [5], [6], [7], [8]. This observation seems to suggest that HRTF personalization is only required in the case that the effect of head rotations are not taken into account in the synthesis process.

If anechoic HRTFs are employed, the perceived distance of the virtual source is often very limited. The addition of a room-acoustic model, involving early reflections (i.e., the first part of the time-domain reverb impulse response that mainly contains distinct peaks) is known to increase the *out-of-head* percept and realism of a virtual sound source [39], [35], [40]. The addition of late reverberation (the remainder of the impulse response that is in most cases described in terms of statistical properties such as the decay time and modal density) does not seem to have any effect on externalization [35]. The incorporation of room acoustics, however, may interfere with those captured by the recording itself and hence may have negative effects as well, especially for *there and then* types of content.

IV. The user perspective

When binaural synthesis techniques are used to simulate a fixed virtual loudspeaker setup using headphones, all restrictions that are valid for loudspeaker systems on the one hand, and the difficulties associated with binaural synthesis on the other hand will all apply simultaneously. In other words, a listener may find him or herself with his or her head tightened to a chair without being able to move or change orientation. The pinnae of the listener's ears are replaced by those from someone else. The acoustical properties of the virtual listening room may interfere with those captured by the recording. The sound stage is quite narrow due to the compromised loudspeaker setup, phantom sources have a somewhat diffuse image and may suffer from comb-filter artefacts, and it is quite difficult to determine whether the virtual loudspeakers are placed in front or behind the listener. This is not a very desirable user perspective and chances are that this approach does not provide the desired immersive and realistic listening experience.

Given all observations described above, we can formulate a set of design recommendations for headphone enhancement algorithms to playback audio produced for loudspeakers.

1) Incorporate head tracking. Head tracking is currently the *only* known factor to resolve front/back reversals. A second important advantage of the incorporation of the proprioceptive sense is that it relaxes the need for individualized HRTFs and all associated measurement challenges that are difficult to realize for consumer-oriented devices.

2) Convert phantom sources to virtual sources. If stereo audio content contains phantom sources, reproduction of this audio content will result in *v*irtual phantom sources if reproduced over a stereophonic virtual loudspeaker setup. The necessity of phantom imaging to overcome the quality vs cost compromise of loudspeaker systems is in many cases not required with headphone-based systems. By selecting the appropriate HRTFs, sound sources can be positioned anywhere around the listener without the restrictions associated with phantom imaging. This is especially important when head tracking is incorporated; a change in the orientation of the listener's head violates the restrictions imposed by amplitude panning and hence the combination of head tracking and phantom imaging should be avoided. Thus, phantom sources will have to be converted to (virtual) sound sources at the intended spatial position.
3) Minimize virtual reproduction-room acoustic interference or at least provide user control. Early reflections are required for externalized sound sources, but late reverberation is not. To minimize interference between the acoustical properties of the recording and the reproduction room, it is recommended to use early reflections only, and possibly provide means to add late reverberation if desired by the listener.

V. Materialize the Phantom

For many applications, phantom sound sources are difficult to circumvent. A significant portion of recorded stereo music, especially in the popular category, is produced using amplitude panning as main spatial imaging method. Hence playback of such material over two virtual loudspeakers will result in "virtual phantom" sources as depicted in the left panel of Fig. 1. Using amplitude panning, the left and right virtual loudspeakers produce a virtual phantom sound source which is subject to the various drawbacks described in Sect. II. The preferred virtual playback scenario is shown in the right panel of Fig. 1. The virtual phantom source is replaced by a virtual source using HRTFs that correspond to an azimuth angle a_b conforming to the perceived azimuth angle of the virtual phantom source shown in the left panel of Fig. 1. This process is referred to as "phantom materialization".

The main challenge of the development of such a scheme is to decompose a set of signals into one or more phantom source signals, including their corresponding perceived positions, and a residual signal, that represents signal components that do not fit in the amplitude panning model, such as room reflections and reverberation or effects that may have been added to certain elements in a stereo mix that modify their spatial attributes. Since it is not a-priori known how many phantom sources are present in a certain audio segment, and (blind) separation of such sources is very difficult, the proposed decomposition method is based on recent trends in spatial audio processing and compression. It has been shown that the spatial image of an auditory scene can be captured by interpreting individual time/frequency tiles of a signal as pseudo auditory objects [41], [42], [43], [44], [45], [46], [47] that have a certain perceived position and a perceived width. The perceived position depends on sound-source localization cues (e.g., interaural time and level differences) while the perceived width predominantly depends on the coherence of the underlying stereo signal pair.

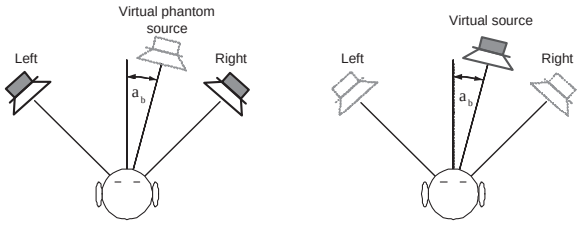

Fig. 1. Two virtual loudspeakers result in a virtual phantom source at an azimuth angle a_b resulting from amplitude panning techniques (left panel). The virtual phantom source is replaced by a virtual source at the same position (right panel).

The proposed method applied to stereo signals is outlined in Fig. 2. A spatial analysis stage decomposes the stereo input signal into various time/frequency tiles. For each time/frequency tile, an estimated perceived position angle a_b of the phantom source is derived based on analysis of sound-source localization cues. Subsequently, for each time/frequency tile the stereo input signal is decomposed into three intermediate signals:

- A phantom source signal S that is constructed from the stereo input signal,
- a residual signal D_l for the left input channel representing components that are not associated with the phantom source signal S, and
- a residual signal D_r for the right input channel representing components that are not associated with the phantom source signal S.

The idea to extract phantom sources from stereo or multi-channel content has been part of extensive research in the past (cf. [48], [49], [46], [50], [51]). This current decomposition method differs however from earlier proposals that focus on extraction of primary and ambience signals. In earlier approaches, the primary

signal is assumed to be directional, while the ambience signal is assumed to be unidirectional to create a sense of spaciousness. The risk with such approach is that if the input signals do not fit the underlying model for the primary components, the resulting signals may erroneously be interpreted as ambience. In the current residual approach, on the other hand, the idea is to enhance the spatial image for signal components that fit in an amplitude-panning model, while leaving other components (the residual signals) untouched.

The three intermediate signals and the position angle a_b of the phantom source are conveyed to a spatial synthesis stage that employs HRTFs to the three intermediate signals to generate the desired virtual sound sources, and subsequently converts the resulting stereo binaural signal to the time domain. Each individual time/frequency tile of the signal S is convolved with HRTFs of the corresponding azimuth angle a_b, while the residual components D_l and D_r are processed with HRTFs of predetermined positions. These predetermined positions are preferably equal to the positions of the loudspeakers in two-channel stereophonic listening.

A. Spatial analysis

The spatial analysis stage comprises a time/frequency analysis transform to obtain the required processing bands. These processing bands preferably follow a nonlinear frequency resolution according to the Equivalent Rectangular Bandwidth (ERB) scale [52] to mimic the assumed perceptual spatial decomposition. Various methods based on Fourier transforms [53], [44] or filter banks [54], [46] have been proposed in the past. In the current implementation, the two input signals $x_l[n]$ and $x_r[n]$ (sampled at 44.1 kHz sampling frequency) were segmented in 1024-sample, 50% overlapping frames. Each segment was subsequently windowed using a square-root Hanning window, and transformed to the frequency domain using a K=1024-point discrete Fourier transform (DFT) to result in frequency-domain representations $X_l[k], X_r[k]$. Finally, the various frequency bins k ($k = 0, ..., K-1$) were grouped into processing bands b ($b = 0, ..., B-1$) to approximate the ERB-scale resolution.

For each processing band b, the following transformation of the signals $X_l[k]$ and $X_r[k]$ was used:

$$S[k] = \frac{X_l[k] + X_r[k]}{\sin(\gamma_b) + \cos(\gamma_b)}, \tag{1}$$

$$D[k] = X_l[k] - \sin(\gamma_b) S[k]. \tag{2}$$

Here, $S[k]$ represents the phantom-source signal and $D[k]$ a single, out-of-phase residual signal according to:

$$D_l[k] = D[k], \tag{3}$$

$$D_r[k] = -D[k]. \tag{4}$$

A single, out-of-phase residual signal $D[k]$ was found to result in a more robust estimate of a residual component than our attempts to estimate independent residual signals $D_l[k]$ and $D_r[k]$. The angle γ_b represents a "sine-cosine pan law" parameter, also referred to as "tangent pan law" parameter [55], [56], [14] and has a value between 0 and 90 deg. The inverse relationship between $S[k], D[k], X_l[k]$ and $X_r[k]$ gives a better insight in the employed signal model:

$$\begin{bmatrix} X_l[k] \\ X_r[k] \end{bmatrix} = \begin{bmatrix} \sin(\gamma_b) & +1 \\ \cos(\gamma_b) & -1 \end{bmatrix} \begin{bmatrix} S[k] \\ D[k] \end{bmatrix}. \tag{5}$$

The solution for γ_b follows from an energetic analysis under the constraints that $S[k]$ and $D[k]$ are independent and $D[k]$ is minimized for each band b independently (see [44], [46] for more details):

$$\tan(\gamma_b) = \frac{\sigma_{X_l,b} \cos(v_b + \beta_b)}{\sigma_{X_r,b} \cos(-v_b + \beta_b)}, \tag{6}$$

with

$$v_b = \frac{1}{2} \arccos(\rho_b). \tag{7}$$

The variable ρ_b denotes the normalized cross-correlation coefficient of the signals $X_l[k]$ and $X_r[k]$ in processing band b:

$$\rho_b = \Re \left(\frac{\sum_{k \in b} X_l[k] X_r^*[k]}{\sigma_{X_l,b} \sigma_{X_r,b}} \right), \tag{8}$$

with $\Re(x)$ the real part of x, X^* the complex conjugate of X, $k \in b$ denoting all bins k that are grouped in parameter band b and $\sigma^2_{X_l,b}, \sigma^2_{X_r,b}$ the energy in the left and right input signals for processing band b, respectively:

$$\sigma^2_{X,b} = \sum_{k \in b} X[k] X^*[k]. \tag{9}$$

The parameter β_b is given by:

$$\beta_b = \tan\left(\frac{\sigma_{X_r,b} - \sigma_{X_l,b}}{\sigma_{X_r,b} + \sigma_{X_l,b}} \arctan(v_b)\right). \tag{10}$$

The amplitude-panning parameter γ_b is mapped to a perceived position parameter a_b for a given (predetermined) virtual loudspeaker setup with a left-channel loudspeaker positioned at an angle a_l and a right-channel

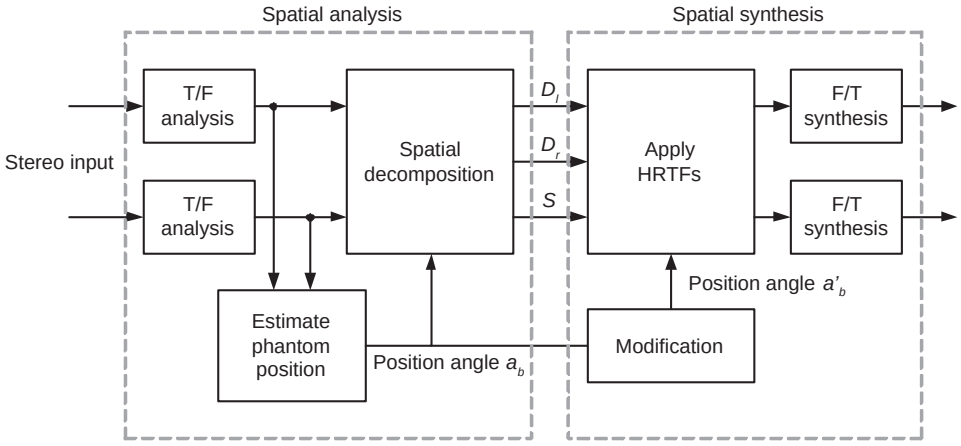

Fig. 2. Outline of the proposed analysis followed by synthesis steps to materialize virtual phantom sources.

loudspeaker at an angle a_r (typically -30 and $+30$ deg, respectively):

$$a_b = a_r + \frac{\gamma_b}{90}\left(a_l - a_r\right). \tag{11}$$

An example of the angle a_b as a function of the ICLD of a single sound source (i.e., $\rho_b = 1$) is given by the solid line in Fig. 3. For an ICLD of -30 dB (i.e., the left loudspeaker radiates 30 dB more power than the right loudspeaker), the angle a_b approaches -30 deg, which corresponds to the position of the left loudspeaker ($a_l = -30$) in this example. For an ICLD of 0 dB, both loudspeakers have the same power and hence the (perceived) position angle a_b equals 0. Further increases of the ICLD result in a larger angle a_b towards $a_r = 30$.

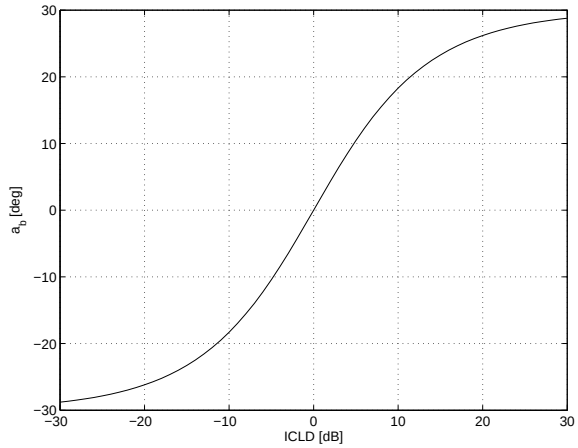

Fig. 3. The angle a_b as a function of the ICLD defined as $20\log_{10}\frac{\sigma_{X_r}}{\sigma_{X_l}}$ for a single panned sound source (i.e., $\rho_b = 1$ represented by the solid line). The positions of the loudspeakers are given by $(a_l, a_r) = (-30, 30$ deg).

The ratio of the power of D and S (expressed in dB) as a function of the ICLD and the normalized cross-correlation ρ_b is visualized in Fig. 4. As can be observed, the (relative) power of D is maximum in the case of an ICLD of zero dB and a correlation $\rho_b = 0$. If the ICLD or correlation increases, the relative power of D decreases accordingly. For extreme panning values (ICLD equal to +30 or -30 dB), the power of signal D becomes extremely small. Hence in that case, the spatial image is effectively modelled by a single source S at the loudspeaker position corresponding to the ICLD (left speaker for ICLD = -30 dB, and right speaker for ICLD = +30dB).

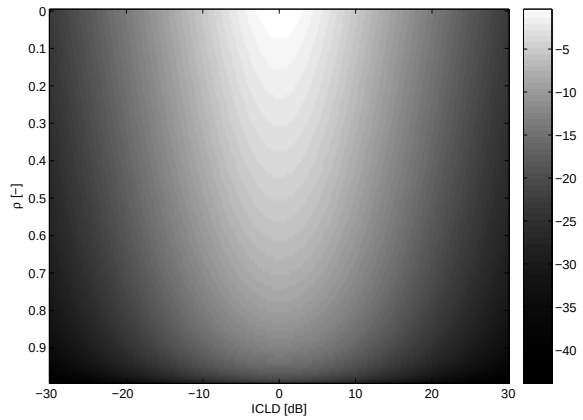

Fig. 4. The ratio of the power of D and S (expressed in dB) as a function of the ICLD and ρ_b.

B. Spatial synthesis

The spatial synthesis stage reconstructs the spatial sound stage based on the signals $S[k]$, $D[k]$ and position angle for the S signal a'_b, and the positions corresponding to the residuals D (a'_l and a'_r), which may be modified versions of a_b, a_l and a_r, respectively. These modified position angles may result from a desired modification in the sound-stage aperture (by increasing the angle

between the two loudspeakers placed at a_l and a_r), or a sound-stage angular offset resulting from a head-tracker according to:

$$a'_l = a_l c_1 + c_0, \tag{12}$$

$$a'_r = a_r c_1 + c_0, \tag{13}$$

$$a'_b = a_b c_1 + c_0, \tag{14}$$

with c_1 a sound-stage aperture scale factor and c_0 a sound-stage offset angle. The sound-stage aperture is visualized in Fig. 5.

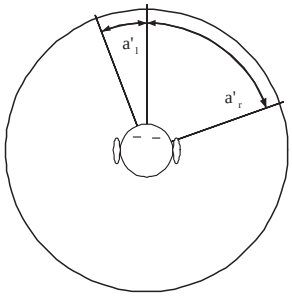

Fig. 5. Aperture of the reconstructed soundstage. The angles a'_l and a'_r represent the modified positions of the left and right loudspeaker, respectively.

The synthesis process comprises generation of three sources for each parameter band: a source $S[k]$ at an azimuth angle a'_b, a source $D[k]$ at an azimuth angle a'_l, and a source $-D[k]$ at an azimuth angle a'_r. A conventional method to create such virtual sound sources is by means of convolution using HRTFs. This process could in principle be employed on sub-band signals as well. Alternatively, in the FFT domain, convolution can be efficiently implemented by multiplication with HRTFs provided that zero-padding is employed to the signal frames to resolve the cyclic behavior of the Fourier transform. Recently, however, it has been shown that HRTF processing can be efficiently implemented in a lossy parametric form [57], [46] without negative perceptual consequences. The parametric representation is especially suitable for the application described here since (1) in contrast to many other approaches, no zero padding is required, and (2) the parametric HRTF description closely matches the processing-band approach that is pursued here. This means that the spatial analysis and spatial synthesis can be employed in the same (transform) domain without creating audible time-domain aliasing distortion.

Using the parametric approach, pairs of HRTFs are described by the following parameters that are defined for each processing band independently:

1) An (average) level parameter of the left-ear HRTF $p_{l,b,a,e}$,
2) An (average) level parameter of the right-ear HRTF $p_{r,b,a,e}$,
3) An average phase difference parameter $\phi_{b,a,e}$,

with a the azimuth angle and e the elevation angle. The (average) level parameters $p_{l,b,a,e}, p_{r,b,a,e}$ describe the spectral envelopes of the HRTFs (and hence the inter-aural level differences). The phase difference parameter $\phi_{b,a,e}$ provides a step-wise constant approximation of the inter-aural time difference. Experiments have shown that if the processing bands are sufficiently small, such step-wise approximation does not result in audible differences compared to the original HRTFs [46]. Using such parametric representation, the synthesis process of a binaural signal pair Y_l, Y_r for a virtual source with signal $X[k]$ within a subband or FFT domain comprises multiplication of the signal with a complex-valued scalar:

$$Y_l[k] = X[k] p_{l,b,a,e} e^{-j\phi_{b,a,e}/2}, \tag{15}$$

$$Y_r[k] = X[k] p_{r,b,a,e} e^{+j\phi_{b,a,e}/2}. \tag{16}$$

The phase difference $\phi_{b,a,e}$ is expressed in radians and is divided symmetrically across the two output signals, which requires $\phi_{b,a,e}$ to be an estimate of the *unwrapped* phase difference between the respective HRTFs.

Applying the parametric method in the current context, the binaural output signals $Y_l[k]$, $Y_r[k]$ given the decomposition of the input signals into $S[k]$, $D[k]$, γ_b, and the position data a'_b, a'_l, and a'_r are given by:

$$\begin{aligned} Y_l[k] = \; & S[k] p_{l,b,a'_b,0} e^{-j\phi_{b,a'_b,0}/2} + \ldots \\ & D[k] p_{l,b,a'_l,0} e^{-j\phi_{b,a'_l,0}/2} - \ldots \\ & D[k] p_{l,b,a'_r,0} e^{-j\phi_{b,a'_r,0}/2} \end{aligned}, \tag{17}$$

$$\begin{aligned} Y_r[k] = \; & S[k] p_{r,b,a'_b,0} e^{+j\phi_{b,a'_b,0}/2} + \ldots \\ & D[k] p_{r,b,a'_l,0} e^{+j\phi_{b,a'_l,0}/2} - \ldots \\ & D[k] p_{r,b,a'_r,0} e^{+j\phi_{b,a'_r,0}/2} \end{aligned}. \tag{18}$$

The values for $p_{l,b,a,e}$, $p_{r,b,a,e}$ and $\phi_{b,a,e}$ are typically stored in a (parametric) HRTF database for a discrete set of virtual sound source positions. The spatial resolution of such a database is typically one parameter set for 5 to 15 degrees in azimuth and/or elevation. If parameters are required for an azimuth and/or elevation angle that is not present in the database, (bi)linear interpolation is employed.

Finally, the time-domain output signals are obtained by an inverse DFT, windowing with a square-root Hanning window and overlap-add (using 50% overlapping frames).

C. Evaluation

1) Stimuli and method: A listening test was conducted to evaluate the subjective implications for the processing scheme as described above. Subjects had to rate three different processing configurations:

1) A standard stereo virtual loudspeaker setup with virtual speakers positioned at -30 and +30 degrees (labeled as "30DegFix");
2) A "widened" stereo virtual loudspeaker setup with virtual speakers positioned at -60 and +60 degrees (labeled as "60DegFix");
3) The phantom materialization method using $c_1 = 2$ resulting in $a_l' = -60$ and $a_r' = +60$ degrees ("60DegDyn").

Nine subjects were asked to provide scores for a set of items for each of the processing methods above on a 100-point scale in a double-blind listening test. The 100-point scale had equidistant anchors labelled "Excellent", "Good", "Fair", "Poor", and "Bad". The test procedure provided means to switch between processing methods in real time and to loop user-definable segments within the excerpts. The (unprocessed) stereo signal was provided as "reference". Subjects were instructed to rate the perceived quality of the processed items, and only use the "reference" to allow identification of artefacts, sound source coloration or unnatural spatial attributes introduced by the processing and to perform a cross-check whether these artefacts are absent or present in the unprocessed content. In other words, the resulting rates can only be interpreted as relative scores across the employed processing methods, but do not indicate any absolute quality rating for the processing itself (compared to the case without any processing). The subjects were seated in a sound-isolated listening room using Stax reference headphones.

Eight excerpts were used that covered a wide variety of content and stereo imaging, including classical music, popular music, speech, and speech with background music or background ambience. A short description of each excerpt is given in Table I. The audio excerpts had a duration between 9 and 30 seconds and were sampled at 44.1 kHz, 16 bits.

Anechoic dummy-head HRTF measurements were employed to generate virtual sound sources. No head tracking was employed in the test. The HRTFs were sampled at a 6-degree azimuth and elevation resolution.

Excerpt	Name	Description
1	Bovio	Classical orchestra
2	Classic	Classical orchestra
3	Elliot1	Speech dialog with ambience
4	Panvces	Two voices inversely panned
5	Popvce	Popular music with female vocalist
6	Scoppin	Jazz instrumental music
7	Spchmusic	Male speech with background music
8	Y tal vez	Latin instrumental music

TABLE I
LISTENING TEST EXCERPTS.

Subsequently, the HRTFs were equalized for the diffuse-field and transformed to the parametric domain (cf. [57], [46]). In total, 28 parameter bands were employed to cover the audible frequency range using a spectral spacing that is identical to that employed in MPEG Surround [46], [58]. Spatial positions in-between HRTF measurement positions were obtained by means of linear interpolation of the HRTF parameters. A simple early-reflections stage was incorporated to increase the percept of distance. This stage operated on a mono down mix in parallel to the HRTF processing stage and consisted of a band-pass filter, a delay, a set of first-order all-pass sections connected in series and a Lauridsen decorrelator [59] to create stereo output. The resulting early-reflections signal was added to the output of the HRTF processing stage with a gain of -6 dB. This level gave an audible but subtle increase in the perceived distance while minimizing the interference with the spatial attributes of the various excerpts.

2) Results: The results of the listening test are shown in Fig. 6. The excerpts are shown along the abscissa. The last entry ("Mean") represents the mean score averaged across excerpts. The scores averaged across subjects are given along the ordinate. The error bars denote the 95% confidence intervals of the means. The various symbols represent the different processing configurations.

The results indicate that the "60DegFix" (diamonds) and "30DegFix" (squares) configurations are more or less equal in terms of scores with a small preference for the "60DegFix" configuration. For some items, the wider loudspeaker base has higher scores (items "Classic" and "Panvces") while the opposite effect is observed for two other items ("Elliot1" and "Scoppin"). The "60DegDyn" configuration (downward triangles) shows considerably higher scores on average than the other two configurations for 7 out of the 8 items, and is on par with the "60DegFix" configuration for the "Spchmusic" item.

A 3-way analysis of variance was carried out using the configuration, excerpt and subject as independent variables (including second-order interactions), and the

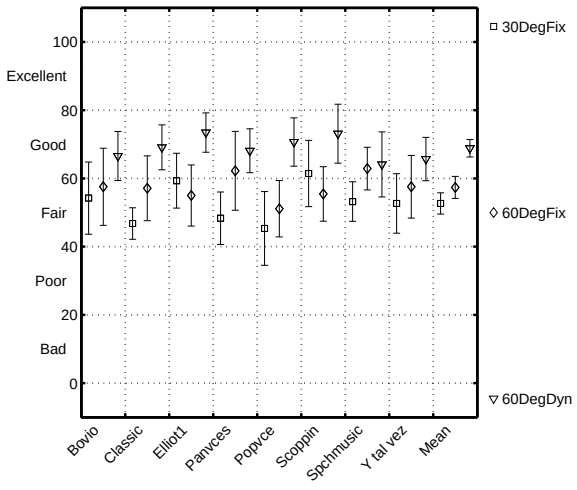

Fig. 6. Subjective preference results for each item averaged across subjects. Error bars denote 95% confidence intervals for the means.

preference score as dependent variable. The main effect of the configuration yielded an F ratio of $F(2, 112) = 50.82$ ($p < 10^{-10}$) indicating that the configuration is a significant factor in the results. The same is true for subject ($F(8, 112) = 11.87$, $p < 10^{-10}$) but not for excerpt ($F(7, 112) = 1.69$, $p > 0.11$). A post-hoc comparison of marginal means revealed that all configurations differ significantly at the 95% confidence level. The interaction between configuration and excerpt resulted in $F(14, 112) = 1.87$ ($p < 0.037$), indicating that some excerpts were more critical to find differences between the various processing methods than others. The same is true for the interaction between configuration and subject ($F(16, 112) = 3.11$, $p < 0.00023$) which indicates differences between subjects to rate the various configurations. Finally, the interaction between excerpt and subject was not found to be significant ($F(56, 112) = 1.18$, $p > 0.22$).

3) Discussion: The subjective ratings indicate a small preference for two virtual loudspeakers placed at +/- 60 degrees compared to placement at +/- 30 degrees. This seems in contradiction to earlier statements that the aperture angle of loudspeakers should be limited to 60 degrees (cf. [13]) to obtain correct phantom source imaging. The degradation of phantom source image quality is confirmed by informal retrospective listening to the various items and processing configurations. For the 120-degrees aperture angle ("60DegFix"), phantom sources tend to sound more "inside" the head and are elevated compared to the corresponding images resulting from the other two processing configurations. On the other hand, the "30DegFix" configuration has a quite narrow spatial extent, which may cause lower preference scores. Possibly, the preference for a wider sound stage for "60DegFix" counteracts the corresponding degradation in phantom-imaging accuracy and "out-of-head" localization.

The phantom materialization method resulted in equal or higher scores than the two other (conventional) processing methods. Especially for those items that consisted of a mixture of multiple sound sources at discrete spatial positions, the quality difference is most prominent (items "Elliot1", "Popvce" and "Scoppin"). These observations support the notion that spatial analysis and synthesis methods can overcome limitations of fixed virtual loudspeaker setups. More specifically, the results can be explained at least qualitatively by considering two main effects: (1) the overall spatial extent (or sound stage width), and (2) the spatial naturalness or imaging quality of the various sound sources. By widening or narrowing the sound stage without the use of spatial analysis and synthesis techniques, these two effects seem to be counteracting. On the other hand, the use of phantom materialization results in a positive-sum result: a more natural and simultaneously wider sound stage. In fact, informal tests revealed that if the orientation of the listener with respect to the loudspeaker setup is changed (which could be the result of the incorporation of a head tracker), the differences are even more pronounced in favor of the phantom materialization method. This observation suggests that the proposed method results in even more natural sound stage imaging when head-tracking is employed.

VI. Conclusions

Playback of conventional stereo content over headphones is usually perceived "inside" the head. A more natural reproduction can be obtained using so-called virtual loudspeakers employing HRTFs. However, when a conventional stereo speaker setup is simulated over headphones, all corresponding compromises that are associated with such a reproduction system will limit the spatial accuracy in the virtual listening environment as well. In this paper an approach was presented that exploits the extended spatial imaging capabilities for headphone reproduction. A stereo signal is decomposed into a number of (phantom) sound sources with corresponding perceived positions. Subsequently, a spatial synthesis step materializes the (virtual) phantom sources by synthesis of the estimated phantom-source signals using HRTFs corresponding to the perceived positions. A listening test was conducted to illustrate the potential of the phantom materialization method. The results indicate that (1) subjects prefer a larger spatial extent of the sound stage and (2) subjects prefer materialized sources rather than (virtual) phantom sources in a virtual listening test

setup. The angular approach of the proposed method is very suitable for applications that include head tracking.

VII. Acknowledgments

The authors would like to thank their colleagues Okke Ouweltjes and Arno van Leest, the associate editor and the anonymous reviewers for their very helpful comments and contributions to improve the manuscript.

References

[1] J. Blauert. *"Spatial hearing: the psychophysics of human sound localization"*. The MIT Press, Cambridge, Massachusetts, 1997.

[2] D. R. Begault. Challenges to the successful implementation of 3-D sound. *J. Audio Eng. Soc.*, 39:864–870, 1991.

[3] J. M. Loomis, C. Hebert, and J. G. Cincinelli. Active localization of virtual sounds. *J. Acoust. Soc. Am.*, 88:757–1764, 1990.

[4] E. M. Wenzel. What perception implies about implementation of interactive virtual acoustic environments. In *Proc. 101st AES Convention*, Los Angeles, USA, 1996.

[5] F. L. Wightman and D. J. Kistler. Resolution of front-back ambiguity in spatial hearing by listener and source movement. *J. Acoust. Soc. Am.*, 105:2841–2853, 1999.

[6] U. Horbach, A. Karamustafaoglu, R. Pellegrini, P. Mackensen, and G. Theile. Design and applications of a data-based auralization system for surround sound. In *Proc. 106th AES convention*, Munich, Germany, 1999.

[7] P. J. Minnaar, S. K. Olesen, F. Christensen, and H. Møller. The importance of head movements for binaural room synthesis. In *Proc. ICAD*, pages 21–25, Espoo, Finland, July 2001.

[8] P. Mackensen. *Head movements, an additional cue in localization*. PhD thesis, Technische Universitaet Berlin, Berlin, 2004.

[9] F. Rumsey. Spatial quality evaluation for reproduced sound: Terminology, meaning and a scene-based paradigm. *J. Acoust. Soc. Am.*, 50:651–666, 2002.

[10] S. P. Lipshitz. Stereo microphone techniques; are the purists wrong? *J. Audio Eng. Soc.*, 34:716–744, 1986.

[11] D. Griesinger. Stereo and surround panning in practice. In *Proc. 112th AES convention*, München, Germany, 2002.

[12] E. Benjamin and P. Brown. The effect of head diffraction on stereo localization in the mid-frequency range. In *Proc. 122nd AES convention*, Vienna, Austria, 2007.

[13] J. C. Bennett, K. Barker, and F. O. Edeko. A new approach to the assessment of sterephonic sound system performance. *J. Audio Eng. Soc.*, 33:314–321, 1985.

[14] V. Pulkki and M. Karjalainen. Localization of amplitude-panned virtual sources I: Stereophonic panning. *J. Audio Eng. Soc.*, 49:739–752, 2001.

[15] H. A. M. Clark, G. F. Dutton, and P. B. Vanderlyn. The 'Stereosonic' recording and reproduction system: A two-channel systems for domestic tape records. *J. Audio Eng. Soc.*, 6:102–117, 1958.

[16] G. Theile and G. Plenge. Localization of lateral phantom sources. *J. Audio Eng. Soc.*, 25:196–200, 1977.

[17] V. Pulkki, M. Karjalainen, and V. Valimaki. Localization, coloration and enhancement of amplitude-panned virtual sources. In *Proc. 16th AES conference*, pages 257–278, Rovaniemi, Finland, 1999.

[18] G. Martin, W. Woszczyk, J. Corey, and R. Quesnel. Sound source localization in a five-channel surround sound reproduction system. In *Proc. 107th AES convention*, New York, USA, 1999.

[19] V. Pulkki. Localization of amplitude-panned virtual sources II: Two-and three-dimensional panning. *J. Audio Eng. Soc.*, 49:753–767, 2001.

[20] J. Corey and W. Woszczyk. Localization of lateral phantom images in a 5-channel system with and without simulated early reflections. In *Proc. 113th AES convention*, Los Angeles, USA, 2002.

[21] V. Pulkki. Coloration of amplitude-panned virtual sources. In *Proc. 110th AES convention*, Amsterdam, The Netherlands, 2001.

[22] B. Shirley, P. Kendrick, and C. Churchill. The effect of stereo crosstalk on intelligibility: Comparison of a phantom stereo image and central loudspeaker source. *J. Audio Eng. Soc.*, 55:852–863, 2007.

[23] B. Sayers. Acoustic image lateralization judgments with binaural tones. *J. Acoust. Soc. Am.*, 36:923–926, 1964.

[24] W. A. Yost. Lateral position of sinusoids presented with interaural intensive and temporal differences. *J. Acoust. Soc. Am.*, 70:397–409, 1981.

[25] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. I. Stimulus synthesis. *J. Acoust. Soc. Am.*, 85:858–867, 1989.

[26] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. II: Psychophysical validation. *J. Acoust. Soc. Am.*, 85:868–878, 1989.

[27] W. M. Hartmann and A. Wittenberg. On the externalization of sound images. *J. Acoust. Soc. Am.*, 99:3678–3688, 1996.

[28] E. H. A. Langendijk and A. W. Bronkhorst. Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *J. Acoust. Soc. Am.*, 107:528–537, 2000.

[29] T. Ajdler, L. Sbaiz, and M. Vetterli. The plenacoustic function on the circle with application to HRTF interpolation. In *Proc. ICASSP*, pages 273–276. IEEE, 2005.

[30] T. Ajdler, L. Sbaiz, and M. Vetterli. Head related transfer functions interpolation considering acoustics. In *Proc. 118th AES convention*, Barcelona, Spain, May 2005.

[31] D. S. Brungart, N. I. Durlach, and W. M. Rabinowitz. Auditory localization of nearby sources. II. Localization of a broadband source. *J. Acoust. Soc. Am.*, 106:1956–1968, 1999.

[32] B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco. Tori of confusion: binaural localization cues for sources within reach of a listener. *J. Acoust. Soc. Am.*, 107:1627–1636, 2000.

[33] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi. Binaural technique: Do we need individual recordings? *J. Audio Eng. Soc.*, 44:451–469, 1996.

[34] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen. Evaluation of artifical heads in listening tests. *J. Audio Eng. Soc.*, 47:83–100, 1999.

[35] D. R. Begault, E. M. Wenzel, and M. R. Anderson. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *J. Audio Eng. Soc.*, 49:904–916, 2001.

[36] F. L. Wightman and D. J. Kistler. Individual differences in human sound localization behavior. *J. Acoust. Soc. Am.*, 99:2470–2500, 1996.

[37] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.*, 94:111–123, 1993.

[38] A. Bronkhorst. Localization of real and virtual sound sources. *J. Acoust. Soc. Am.*, 98:2542–2553, 1995.

[39] D. R. Begault. Perceptual effects of synthetic reverberation on 3-D audio systems. In *Proc. 91th AES convention*, New York, USA, 1991.

[40] B. G. Shinn-Cunningham. The perceptual consequences of creating a realistic, reverberant 3-D audio display. In *Proc.*

of the international congress on acoustics, Kyoto, Japan, April 2004.

[41] F. Baumgarte and C. Faller. Why binaural cue coding is better than intensity stereo coding. In *Proc. 112th AES convention*, Munich, Germany, 2002.

[42] C. Faller and F. Baumgarte. Binaural cue coding applied to stereo and multi-channel audio compression. In *Proc. 112th AES convention*, Munich, Germany, 2002.

[43] F. Baumgarte and C. Faller. Binaural cue coding - part I: Psychoacoustic fundamentals and design principles. *IEEE Trans. SAP*, 11:509–519, 2003.

[44] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers. Parametric coding of stereo audio. *EURASIP J. on Applied Signal Processing*, 9:1305–1322, 2004.

[45] J. Breebaart. Analysis and synthesis of binaural parameters for efficient 3D audio rendering in MPEG Surround. In *Proc. ICME 2007*, pages 1878–1881, Beijing, China, 2007.

[46] J. Breebaart and C. Faller. *"Spatial audio processing: MPEG Surround and other applications"*. John Wiley & Sons, Chichester, 2007.

[47] V. Pulkki. Spatial sound reproduction with directional audio coding. *J. Audio Eng. Soc.*, 55:503–516, 2007.

[48] C. Avendano and J.-M. Jot. Frequency-domain techniques for stereo to multichannel upmix. In *Proc. 22nd AES conference*, Espoo, Finland, 2002.

[49] R. Irwan and R. M. Aarts. Two-to-five channel sound processing. *J. Audio Eng. Soc.*, 50:914–926, 2002.

[50] M. M. Goodwin and J. M. Jot. Binaural 3-D audio rendering based on spatial audio scene coding. In *Proc. 123rd AES convention*, New York, USA, 2007.

[51] J. Merimaa, M. M. Goodwin, and J. M. Jot. Correlation-based ambience extraction from stereo recordings. In *Proc. 123rd AES convention*, New York, USA, 2007.

[52] B. R. Glasberg and B. C. J. Moore. Auditory filter shapes in forward masking as function of level. *J. Acoust. Soc. Am.*, 71:946–949, 1982.

[53] J. Breebaart, S. van de Par, A Kohlrausch, and E. Schuijers. High-quality parametric spatial audio coding at low bit rates. In *Proc. 116th AES convention*, Berlin, Germany, 2004.

[54] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård. Low complexity parametric stereo coding. In *Proc. 116th AES convention*, Berlin, Germany, 2004.

[55] A. D. Blumlein. UK Patent 394,325 (1931), reprinted in Stereophonic techniques, Audio Eng. Soc., NY, 1986.

[56] B. Bernfeld. Attempts for better understanding of the directional stereophonic listening mechanism. In *Proc. 44th AES convention*, Rotterdam, The Netherlands, 1973.

[57] J. Breebaart, L. Villemoes, and K. Kjörling. Binaural rendering in MPEG Surround. *EURASIP J. on Applied Signal Processing*, Volume 2008, Article ID 732895, 2008.

[58] J. Rödén, J. Breebaart, J. Hilpert, H. Purnhagen, E. Schuijers, J. Koppens, K. Linzmeier, and A. Hölzer. A study of the MPEG Surround quality versus bit-rate curve. In *Proc. 123rd AES convention*, New York, USA, 2007.

[59] H. Lauridsen. Experiments concerning different kinds of room-acoustics recording. *Ingenioren*, 47, 1954.