PHANTOM MATERIALIZATION FOR HEADPHONE REPRODUCTION

Jeroen Breebaart

Philips Research
NL-5656 AE Eindhoven The Netherlands

Erik Schuijers

Philips Applied Technologies
NL-5656 AE Eindhoven The Netherlands

ABSTRACT

Conventional stereo audio material is often produced using amplitude panning techniques to achieve flexible positioning of sound sources with a limited number of loudspeakers. Consequently, for faithful playback the orientation and position of the listener are very restricted. If stereo audio is reproduced over virtual loudspeakers on headphones incorporating head-tracking, these position and orientation restrictions limit the realism and spatial accuracy of the reproduction. In this paper, it will be outlined that amplitude panning and the corresponding phantom sound sources cause spatial quality limitations for headphone rendering. A novel method to circumvent phantom imaging on headphones is presented and evaluated in a listening test.

Index Terms— Headphones, Audio systems, phantom imaging, amplitude panning

1. INTRODUCTION

Mobile audio has become increasingly popular during the last two decades. Mobility and social constraints dictate headphones as a reproduction device on mobile players, often resulting in sound sources perceived *inside* the head [1]. The absence of the effect of the acoustical pathway from sound 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. It has been shown however that it is very difficult to match the perceived and intended sound source position and distance using a generic, non-individualized system [2]. Moreover, by simulation of a *virtual loudspeaker setup*, the constraints and limitations that apply to a loudspeaker system will also limit the spatial image quality in its virtual counterpart.

2. AMPLITUDE PANNING

One of the major challenges that audio engineers are facing is that a stereophonic loudspeaker system is restricted in terms of spatial imaging capabilities. These restrictions follow from various cost and esthetic considerations. For example, if the listener is located in the sweet spot, a technique referred to as amplitude panning can position a phantom sound source between the two loudspeakers. The employed inter-channel level differences result in inter-aural time differences (ITDs) and inter-aural level differences (ILDs) at the level of the listeners eardrums that only roughly correspond to those of the desired phantom sound source position [3, 4]. 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 [5]. Larger aperture angles result in significant errors between desired and actual sound source localization attributes [6]. Thirdly, the position and orientation of the listener are very restricted. 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 [7]. If the listeners orientation does not correspond to the intended orientation, further discrepancies between intended and desired sound source locations occur [8, 9]. Finally, amplitude panning can result in sound source coloration [10] and/or degraded speech intelligibility [11].

3. HEADPHONE REPRODUCTION

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) [12]. 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 typically limited processing power and battery life.

Another difficulty is the dependence of HRTFs on the specific anthropometric properties of each individual [13]. Even if individualized HRTFs are used, subjects show localization

errors and front/back confusions [14]. The only known solution that resolves front/back confusions that is known to work robustly is the incorporation of head rotations by means of a head tracker [15, 16]. Unfortunately, incorporation of head tracking is cumbersome for stereo audio material employing amplitude panning given the inability of amplitude panning to work in a front/back direction. Hence if head tracking is to be combined with amplitude-panned stereo material, phantom images should be processed with care to avoid ambiguous or unnatural sound source localization cues. A proposed solution is outlined in the next section.

4. PHANTOM MATERIALIZATION

Playback of stereo 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. 2. 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".
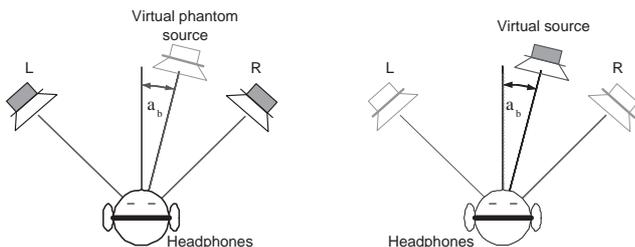

Fig. 1. Replacement of a virtual phantom source at azimuth angle a_b (left panel) by a virtual source (right panel).

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 residual signals, that represent 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. The current method extends existing approaches (for example [17]) by interpreting individual time/frequency tiles of a signal as pseudo auditory objects [18, 19] that have a certain perceived position and a perceived width. The perceived position depends on sound-source localization cues (e.g., inter-aural time and level differences) while the perceived width predominantly depends on the coherence of the underlying stereo signal pair in each specific time/frequency tile.

The proposed method is outlined in Fig. 2 and is described in more detail in [20]. A spatial analysis stage decomposes the stereo input signal into various time/frequency tiles (not shown) according a perceptual frequency scale. For each time/frequency tile, an estimated perceived position angle a_b of the phantom source is derived from an analysis of sound-source localization cues between the left and right input signals L and R. Based on the assumption that the majority of stereo content is produced using amplitude panning and hence inter-channel time differences do not play a significant role, the (1) inter-channel level differences (ICLDs) and (2) the inter-channel cross-correlation (ICC) are calculated for each time/frequency tile. Subsequently, the stereo input signal is decomposed into three intermediate signals: A phantom source signal S, and two residual signals D_l and D_r. These residual signals represent components that are not associated with the phantom source signal S. The estimates of the intermediate signals S, D_l and D_r follow from the following signal model:

$$\left[\begin{array}{c} L \\ R \end{array} \right] = \left[\begin{array}{ccc} \sin(\gamma_b) & 1 & 0 \\ \cos(\gamma_b) & 0 & 1 \end{array} \right] \left[\begin{array}{c} S \\ D_l \\ D_r \end{array} \right]. \qquad (1)$$

The angle γ_b represents the employed panning parameter (i.e., $\gamma_b = 0$ or $\gamma_b = 90$deg corresponds to a source panned to the left or right speaker, respectively). The solution for γ_b was found by maximizing the power of S given the ICLD and ICC parameters and the assumption that the expected values of the cross-products $< S, D_l >$ and $< S, D_r >$ are zero. A further signal model simplification was employed by assuming $D_l = -D_r$.

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 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 loudspeaker positions.

In the synthesis process, the azimuth angle a_b, and the predetermined positions for the residual components may be modified according to head-tracker data or a desirable modification in the sound-stage aperture (cf. [20]).

4.1. Evaluation

4.1.1. Stimuli and method

A listening test was conducted to evaluate the subjective implications for the processing scheme as described above. Nine
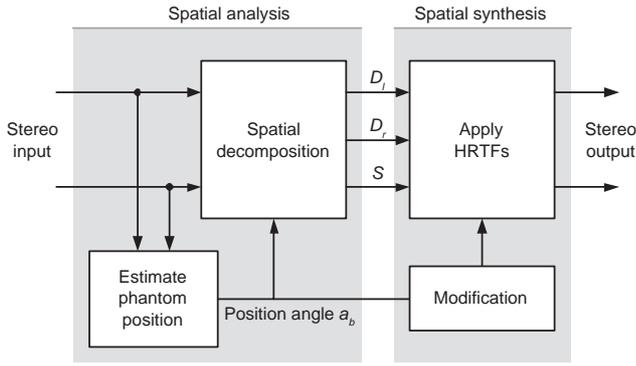
Fig. 2. Spatial analysis and synthesis approach.

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 "30Deg");

2. A "widened" stereo virtual loudspeaker setup with virtual speakers positioned at -60 and +60 degrees (labeled as "60Deg");

3. The phantom materialization (PM) method with a total aperture of 120 degrees ("60DegPM").

The subjects were asked to rate a set of items for each of the processing methods above on a 100-point scale in a double-blind listening test. 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 that allowed subjects to check whether sound source coloration or artefacts were introduced by the processing or were already present in the original excerpts. Subjects were seated in a sound-isolated listening room using Stax reference headphones. No head tracking was employed in the test.

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 ambiance. The audio excerpts had a duration between 9 and 30 seconds and a sampling rate of 44.1 kHz, 16 bits.

Anechoic dummy-head HRTF measurements were employed to generate virtual sound sources. The HRTFs were sampled at a 6-degree azimuth and elevation resolution and were equalized for their diffuse-field response. The HRTFs were converted to a lossy but perceptually transparent parametric form [21, 22, 19] to allow seamless integration of spatial analysis and synthesis in the same domain without the need for additional transforms or zero padding operations.

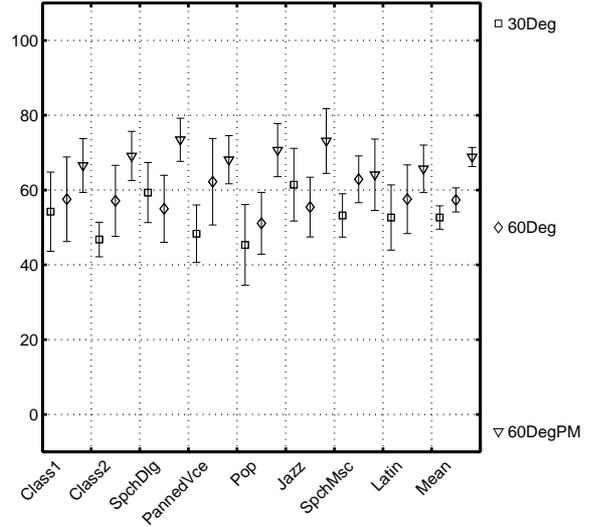A simple early-reflections stage was incorporated to increase the percept of distance. This stage operated on a mono

Fig. 3. Subjective preference results for each item averaged across subjects. Error bars denote 95% confidence intervals for the means.

down mix in parallel to the HRTF processing stage. The resulting stereo 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.

4.1.2. Results

The results of the listening test are shown in Fig. 3. The excerpts are shown along the horizontal axis. The last entry ("Mean") represents the mean score averaged across excerpts. The scores averaged across subjects are given along the vertical axis. The error bars denote the 95% confidence intervals of the means. The various symbols represent the different processing configurations.

The results indicate that the "60Deg" (diamonds) and "30Deg" (squares) configurations differ only slightly, especially if their means across subjects and excerpts are considered. The "60DegPM" configuration (downward triangles) shows considerably higher scores than the other two configurations for 7 out of the 8 items.

4.1.3. Discussion

The subjective ratings indicate an approximately equal preference for two virtual loudspeakers placed at either 30 or 60 degrees. This seems in contradiction to earlier statements that the aperture angle of loudspeakers should be limited to 60 degrees (cf. [5]) 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 ("60Deg"), 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 "30Deg" configuration has a quite narrow spatial extent, which may cause lower preference scores. Possibly, the preference for a wider sound stage for "60Deg" 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. These observations support the notion that spatial analysis and synthesis methods can overcome limitations of fixed virtual loudspeaker setups. 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.

5. CONCLUSIONS

In this paper an approach was presented that exploits the spatial imaging flexibility 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. Results from a listening test reveal that subjects prefer a larger spatial extent of the sound stage, provided that the degradation in phantom source imaging that would normally occur with conventional stereo content is compensated for.

6. 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.*, vol. 39, pp. 864–870, 1991.

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

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

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

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

[7] 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.*, vol. 6, pp. 102–117, 1958.

[8] G. Theile and G. Plenge, "Localization of lateral phantom sources," *J. Audio Eng. Soc.*, vol. 25, pp. 196–200, 1977.

[9] 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*, 1999.

[10] V. Pulkki, M. Karjalainen, and V. Valimaki, "Coloration, and Enhancement of Amplitude-Panned Virtual Sources," in *Proc. 16th AES conference*, 1999.

[11] 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.*, vol. 55, pp. 852–863, 2007.

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

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

[14] 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.*, vol. 49, pp. 904–916, 2001.

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

[16] U. Horbach, A. Karamustafaoglu, R. Pellegrini, P. Mackensen, and G. Theile, "Design and aplpications of a data-based auralization system for surround sound," in *Proc. 106th AES convention*, 1999.

[17] C. Avendano and J.-M. Jot, "Frequency-domain techniques for stereo to multichannel upmix," in *Proc. 22nd AES international conference on virtual, synthetic and entertainment audio*, Espoo, Finland, 2002, pp. 121–130.

[18] C. Faller and F. Baumgarte, "Binaural cue coding applied to stereo and multi-channel audio compression," in *Preprint 5574, 112th AES convention*, Munich, Germany, 2002.

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

[20] J. Breebaart and E. Schuijers, "Phantom materialization: A novel method to enhance stereo audio reproduction on headphones," *IEEE Trans. On Audio, Speech and Language processing*, p. Under review, 2008.

[21] J. Breebaart and A. Kohlrausch, "The Perceptual (ir)relevance of HRTF magnitude and phase spectra," in *Preprint 5406, 110th AES convention*, Amsterdam, The Netherlands, 2001.

[22] J. Breebaart, L. Villemoes, and K. Kjörling, "Binaural rendering in MPEG Surround," *EURASIP J. on Applied Signal Processing*, vol. Accepted, 2008.