ANALYSIS AND SYNTHESIS OF BINAURAL PARAMETERS FOR EFFICIENT 3D AUDIO RENDERING IN MPEG SURROUND

Jeroen Breebaart

Philips Research Laboratories
5656 AE Eindhoven, The Netherlands

ABSTRACT

This paper describes a novel method to simulate an multi-channel audio experience using stereo headphones. In contrast to conventional, convolution-based methods, the current approach is based on parametric representations of spatial audio. An audio scene with multiple virtual sound sources is represented by a mono down-mix signal of all sound source signals, accompanied by certain statistical (spatial) properties. These statistical properties of the sound sources are combined with statistical properties of head-related transfer functions to estimate "binaural parameters" that represent the perceptually-relevant aspects of the auditory scene. Subsequently, a binaural rendering stage re-instates the estimated binaural parameters on the down mix. The advantage of this approach is that the computational complexity of the rendering process is virtually independent of the number of simultaneous sound sources. If combined with parametric multi-channel audio coders such as MPEG Surround, the proposed method is advantageous over conventional methods in terms of perceived quality and computational complexity.

1. INTRODUCTION

The synthesis of virtual auditory scenes has been an ongoing research topic for many years. The aim of so-called binaural rendering systems is to evoke the illusion of one or more sound sources positioned around the listener using stereo headphones. Binaural rendering has benefits in the field of research, simulation and entertainment [1]. Especially in the field of entertainment, the virtual auditory scene should sound very compelling and "real". However, because of the complex nature of current state-of-the-art systems, several concessions are required for feasible implementations, especially if the number of sound sources that has to be rendered simultaneously is large.

Recent trends in consumer audio show a shift from stereo to multi-channel audio content, as well as a shift from solid state to mobile devices. These developments cause additional constraints on transmission and rendering systems. Firstly, the number of audio channels that has to be transmitted increases from two to five. The corresponding increase in transmission bandwidth that would result from conventional, discrete-channel audio coders is often undesirable and sometimes even unavailable. Secondly, consumers often use headphones for audio rendering on a mobile device. To experience the benefit of multi-channel audio, a compelling binaural rendering system is required. This is quite a challenge given the limited processing power and battery life of mobile devices.

In this paper, a novel binaural rendering process will be described that exploits recent advances in parametric multi-channel audio compression. The method is based on the analysis and synthesis of perceptually-relevant parameters of a virtual auditory scene. The analysis and synthesis of these so-called "binaural parameters" is outlined in Secs. 3-4; the integration of this method in the recently finalized MPEG Surround standard [2] for multi-channel audio compression is described in Sec. 5.

2. BINAURAL PARAMETERS

Sound source localization in the horizontal plane is facilitated by inter-aural time differences (ITDs) and inter-aural level differences (ILDs) [3], caused by relative path lengths and the acoustic shadow effect of the head. The properties of sound propagation also result in an intricate frequency-dependence of these cues. Sound source elevation is predominantly facilitated by elevation-dependent spectral peaks and notches that are superimposed on the original sound source spectrum [4].

The acoustical transfer from a certain sound source position to both eardrums in an anechoic environment can be accurately described by a pair of head-related transfer functions (HRTFs). However, several investigations have shown that HRTFs may comprise pronounced signal properties that seem perceptually *irrelevant*. For example, it has been shown that for low frequencies, ITDs dominate sound source localization, while at high frequencies, ILDs and spectral cues are more important [5]. Other researchers have successfully demonstrated that the frequency-dependent ITD can be replaced by a constant, position-dependent ITD without perceptual consequences [6, 7]. A related finding is that the inter-aural time difference can be replaced by a constant inter-aural phase difference (IPD) within various frequency bands. The resulting piece-wise constant phase characteristic does not re-

1-4244-1017-7/07/$25.00 ©2007 IEEE

sult in audible differences provided that the frequency bands are not broader than critical bands [7].

There is also considerable evidence that certain details of the HRTF *magnitude* spectra are irrelevant [8, 9]. Specifically, it seems that *constant* spectral cues within critical bands are a sufficient requirement for high-quality binaural rendering.

Besides such local *spectral* stationarity (in critical bands) as a sufficient prerequisite for high-quality binaural rendering, there are several indications that *temporal* limitations can be exploited as well. Several experiments have revealed a temporally sluggish response of the binaural hearing system [10]. Fast variations of binaural cues (in the order of 10 Hz or faster) are perceived as a change in the "compactness" or "wideness" [3] rather than a temporally varying position. A statistical property that is often associated with this perceived compactness is the inter-aural coherence (IC).

The observation that one set of spatial parameters for each time/frequency tile describes the most relevant perceptual aspects of a sound scene is the basis for recent developments in audio compression schemes. So-called "spatial audio coding" or "binaural cue coding" algorithms describe stereo or multi-channel audio by means of a down mix accompanied with "spatial parameters", that describe the perceptually relevant spatial properties between various audio channels of the original content (cf. [11, 12, 13]).

3. BINAURAL PARAMETER ANALYSIS

In conventional binaural rendering systems, sound sources i with associated time-domain signals $x_i(t)$ are rendered at certain positions by convolving each signal with a pair of head-related impulse responses $h_{L,i}(t)$, $h_{R,i}(t)$, for the left and right ears, respectively, to result in binaural signals $y_{L,i}(t)$, $y_{R,i}(t)$. This process is visualized in the left panel of Fig. 1.

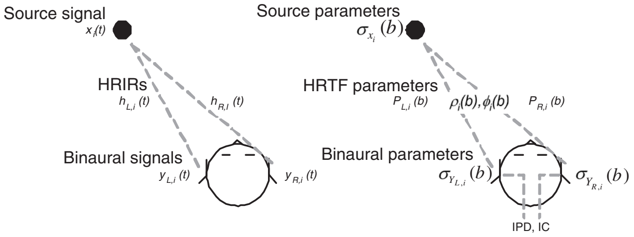

Fig. 1. Synthesis of a virtual sound source by means of HRIR convolution (left panel) and by means of parametric representations (right panel).

It is often convenient to express the convolution in the frequency domain using a frequency-domain representation $X_i(f)$ of a short segment of $x_i(t)$:

$$Y_m(f) = \sum_i H_{m,i}(f) X_i(f), \tag{1}$$

with $H_{L,i}(f)$,$H_{R,i}(f)$ the frequency-domain representations (head-related transfer functions) of $h_{L,i}(t)$,$h_{R,i}(t)$, respectively, and $m \in \{L, R\}$. The binaural parameter analysis approach aims at the estimation of the power $\sigma^2_{Y_{m,i}}$ in each time/frequency tile (a so-called "parameter band" which represents a specific frequency range) of signals Y_m, as well as the inter-aural phase difference (IPD) and inter-aural coherence (IC) that result from all virtual sound sources simultaneously. It turns out that under certain constraints, these binaural parameters can be estimated accurately from certain statistical properties of the signals $x_i(t)$ and the HRTFs $H_{L,i}(f)$,$H_{R,i}(f)$. In particular, (1) the powers of the source signals in each time frequency tile, $\sigma^2_{X_i}$, and (2) the mutual cross-correlation coefficients c_{i_1,i_2} between source signals x_{i_1} and x_{i_2} have to be known. The parameters of each HRTF pair that are required comprise (1) the powers in a specific parameter band (b) of each HRTF, represented by $P^2_{L,i}(b)$, $P^2_{R,i}(b)$ for the left and right ears, respectively, (2) the average phase difference $\phi_i(b)$ and (3) the coherence $\rho_i(b)$ between corresponding HRTFs.

4. BINAURAL PARAMETER SYNTHESIS

The synthesis process comprises re-instating the binaural parameters on a down mix signal $X(f)$ of the object signals. Using a frequency-domain representation, one frame of the down-mix signal is given by:

$$X(f) = \sum_i X_i(f). \tag{2}$$

The reconstructed binaural signals $\hat{Y}_L, \hat{Y}_R$ are obtained using a matrix operation $\mathbf{W}_b$ that is derived for each parameter band (b) independently and applied to all frequency components (f) that belong to the parameter band (b) following:

$$\begin{bmatrix} \hat{Y}_L(f) \\ \hat{Y}_R(f) \end{bmatrix} = \mathbf{W}_b \begin{bmatrix} X(f) \\ D(X(f)) \end{bmatrix}, \tag{3}$$

with $D(.)$ a so-called "decorrelator" which generates a signal that has virtually the same temporal and spectral envelopes as its input but is independent from its input. The independence is achieved by delays, all-pass filters and spectro-temporal envelope adjustment tools. This method of binaural synthesis is identical to the parameter synthesis method applied in "parametric stereo" decoders [12]. The matrix coefficients ensure that for each frame, the two binaural output signals $\hat{Y}_L, \hat{Y}_R$ have the correct levels, as well as IPD and IC relations.

5. APPLICATION TO MPEG SURROUND

MPEG Surround [2] is a novel parametric method for efficient transmission of multi-channel audio. In this audio coding format, a multi-channel audio signal is represented as a downmix signal and a set of "spatial parameters" that, among other aspects, describe the statistical relations of the original multi-channel signals in terms of (relative) signal powers and correlation coefficients. Thus, the parameters that are transmitted represent identical statistical properties between audio channels as those required for the binaural analysis and synthesis approach described in Secs. 3 and 4. It is therefore possible to integrate the binaural analysis and synthesis approach in a dedicated MPEG Surround *binaural decoding mode* for headphone playback. The architecture of this mode is visualized in Fig. 2. Instead of directly applying the transmitted spatial parameters to the output signals, the parameters are used in the binaural parameter analysis stage to compute the binaural parameters that would result from the combined spatial decoding and binaural rendering process. Thus, the binaural parameter analysis stage estimates the binaural parameters σ_{Y_L}, σ_{Y_R}, IPD and IC for each parameter band and each newly-transmitted parameter set. The binaural output signals are subsequently synthesized by the binaural parameter synthesis stage.

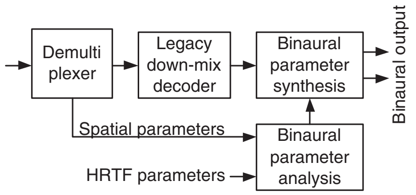

Fig. 2. Overview of the binaural decoding mode.

5.1. Evaluation

5.1.1. Procedure

A listening test was pursued to evaluate the subjective quality of the proposed binaural synthesis method. In this test, the quality of the MPEG Surround binaural decoding mode ("MPS binaural") is compared to a reference condition. This reference condition comprised convolution of an original multi-channel audio excerpt with HRIRs. As a control condition, the combination of MPEG Surround multi-channel decoding followed by conventional HRIR convolution was employed (denoted "MPS + HRIR"). For all configurations, anechoic KEMAR HRIRs [14] were used with a length of 128 samples at a sampling frequency of 44.1 kHz.

For both the binaural decoding mode as well as the control condition, the same MPEG Surround bit stream was employed. This bit stream was generated using a state-of-the-art MPEG Surround encoder using a mono down mix configuration. This mono down mix was subsequently encoded using a high-efficiency AAC encoder at 44 kbps. The spatial parameters generated by the MPEG Surround encoder occupied approximately 4 kbps. This rather low bit rate of 48 kbps total was selected because it is foreseen that the binaural decoding mode is especially suitable for mobile applications that are under severe transmission bandwidth and complexity constraints.

Twelve listeners participated in this experiment. In a double-blind MUSHRA test [15], the listeners had to rate the perceived quality of several processed excerpts against the original (i.e., unprocessed) excerpts on a 100-point scale with 5 anchors. A hidden reference and the low-pass filtered anchor (reference with a bandwidth limitation of 3.5 kHz) were also included in the test.

A total of 11 critical excerpts were used. The excerpts are the same as used in the MPEG Call for Proposals (CfP) on Spatial Audio Coding [16], and range from pathological signals (designed to be critical for the technology at hand) to movie sound and multi-channel music productions.

5.1.2. Results

The results of the listening test are shown in Fig. 3. The various excerpts are given along the abscissa, while the ordinate corresponds to the average MUSHRA score across listeners. Different symbols refer to different configurations. The error bars denote the 95% confidence intervals of the means.

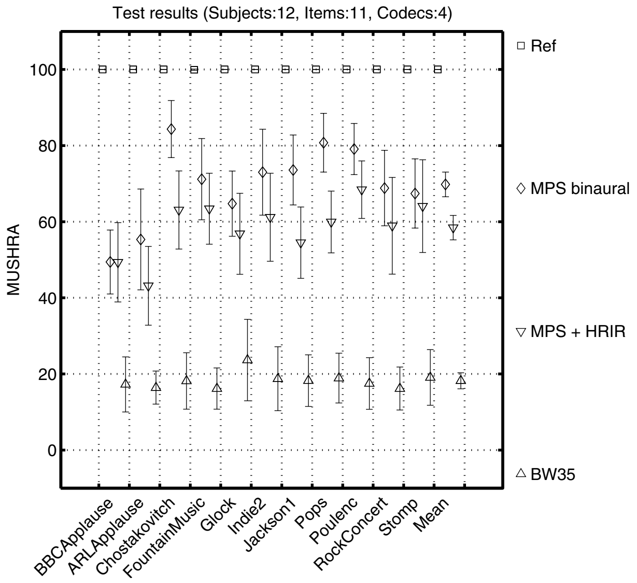

Fig. 3. Subjective test results.

The hidden reference (square symbols) has the highest scores. The results for the binaural decoding mode are denoted by the diamonds; the control condition using convolution is represented by the downward triangles. Although the scores for these methods vary between 45 and 85, the binaural decoding approach has scores that are higher than the conventional method for all excerpts. Finally, the low-pass anchor has the lowest scores of around 20.

If the computational complexity of the binaural decoder and the conventional systems are compared, also interesting differences are observed. The number of operations (expressed in equivalent multiply-accumulates per stereo output sample pair) amounts to 251 for the binaural decoder and 1066 for the MPEG Surround multi-channel decoder followed by convolution using Fast Fourier Transforms. Hence the binaural decoding mode has a computational complexity that is approximately four times lower than the conventional, convolution-based method.

5.1.3. Discussion

The results of the perceptual evaluation indicate that both binaural rendering methods (the binaural decoding mode and the conventional HRIR convolution method) are distinguishable from the reference. This is most probably due to the low bit rate (48 kbps total) that was employed to represent the multi-channel signal in MPEG Surround format. For loudspeaker playback, the perceived quality of MPEG Surround operating at 48 kbps has been shown to equal a MUSHRA score of 65 in other tests [13]. In that respect, the quality for the test and control conditions are in line with earlier reports.

The parametric representation of MPEG Surround aims at perceptual reconstruction of multi-channel audio. As such, at the bit rate that was under test, MPEG Surround does not deliver full waveform reconstruction of the multi-channel output signals. Given the low scores for MPEG Surround decoding followed by HRIR convolution, the multi-channel signals resulting from the parametric representation seem unsuitable for further post processing using HRIRs. The binaural decoding mode, however, which does not rely on processing of decoded signals, outperforms the convolution-based method for the tested configuration, both in terms of perceived quality and computational complexity. As a result, the proposed parametric method is currently adopted as an integral part of the MPEG Surround standard, including extensions to stereo down mixes, support for echoic impulse responses and high-quality filtering modes [2, 13].

6. CONCLUSIONS

The novel, parametric method for binaural rendering has been presented that can be integrated with spatial audio coders. The combination of low transmission bit rates, low complexity and high quality make this method especially suitable for mobile applications.

7. REFERENCES

[1] B. G. Shinn-Cunningham, “Applications of virtual auditory displays,” in *20th Ann. Conf. IEEE Eng. Med. Biol. Soc.*, Hong Kong, China, October 1998.

[2] ISO IEC, “MPEG audio technologies - Part 1: MPEG Surround,” ISO/IEC FDIS 23003-1:2006(E), 2004.

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

[4] F. L. Wightman and D. J. Kistler, “Individual differences in human sound localization behavior,” *J. Acoust. Soc. Am.*, vol. 99, pp. 2470–2500, 1996.

[5] 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.

[6] A. Kulkarni and H. S. Colburn, “Role of spectral detail in sound-source localization,” *Nature*, vol. 396, pp. 747–749, 1998.

[7] J. Breebaart, F. Nater, and A. Kohlrausch, “A parametric approach to represent Head-Related Transfer Functions,” *J. Acoust. Soc. Am.*, p. In preparation, 2007.

[8] J. Huopaniemi and N. Zacharov, “Objective and subjective evaluation of head-related transfer function filter design,” *J. Audio. Eng. Soc.*, vol. 47, pp. 218–239, 1999.

[9] 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.

[10] B. Kollmeier and R. H. Gilkey, “Binaural forward and backward masking: evidence for sluggishness in binaural detection,” *J. Acoust. Soc. Am.*, vol. 87, pp. 1709–1719, 1990.

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

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

[13] L. Villemoes, J. Herre, J. Breebaart, G. Hotho, S. Disch, H. Purnhagen, and K. Kjörling, “MPEG Surround: the forthcoming ISO standard for spatial audio coding,” in *Proc. 28th AES conference*, Pitea, Sweden, 2006.

[14] B. Gardner and K. Martin, “HRTF Measurements of a KEMAR dummy-head microphone,” MIT Media Lab Perceptual Computing Technical Report 280, May 1994, 1994.

[15] ITU-R, “Method for the subjective assessment of intermediate quality level of coding systems (MUSHRA),” ITU-R Recommend. BS.1534, 2001.

[16] Audio Subgroup, “Call for proposals on spatial audio coding,” ISO/IEC JTC1/SC29/WG11 N6455, 2004.