Parametric binaural synthesis: Background, applications and standards

Jeroen Breebaart[1], Fabian Nater[2], Armin Kohlrausch[1,3]
[1]*Philips Research, 5656 AE Eindhoven, The Netherlands, Email: jeroen.breebaart@philips.com*
[2]*Swiss Federal Institute of Technology, CH - 8092 Zürich, Switzerland*
[3]*Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands*

Abstract

The amount of information present in HRTFs and the required processing capabilities for real-time and interactive binaural rendering have long been a challenge for many applications for binaural rendering. More recently, parametric methods to capture the perceptually-relevant information from HRTFs have been developed. By means of extracting perceptually-relevant attributes from HRTF pairs, binaural rendering can be performed at lower complexity compared to the employment of HRTF convolution. Furthermore, parameter-based binaural rendering can be efficiently integrated with parametric audio coders. Last but not least, parametric spatial processing can be used to provide a more convincing spatial reproduction for conventional stereo signals. This paper provides an overview of the perceptual consequences and limitations of HRTF parameterization, its applications, and relevant standardization efforts.

Introduction

The process to generate virtual sound sources using headphones that evoke the same percept as real sound sources have been subject to research for several decennia. The most popular approach to this challenging problem is to measure head-related transfer functions (HRTFs) or binaural room impulse responses (BRIRs) and to convolve each source signal with a pair of HRTFs or BRIRs that correspond to the desired sound source position [1, 2, 3]. Although very successful results have been reported using this approach, several challenges to implement these methods in consumer devices have been identified as well [4, 5, 6]:

- The large amount of data in HRTF/BRIR databases and the associated amount of required computing power is often difficult to realize;
- The incorporation of the measurement and resulting effect of head rotations (i.e., head tracking) with low latency is a challenging problem;
- The necessity to use personalized HRTFs/BRIRs required for a convincing effect may suit laboratory tests but is virtually impossible for consumer-grade applications.

HRTF parameterization

The problem of large amounts of data and the associated processing power has been investigated by several groups. Initial methods predominantly exploited statistical redundancy for data reduction [7] while more recent approaches focused more on removal of perceptually irrelevant information in HRTF spectra. Several approaches for perceptual irrelevancy removal have been published. Most of these are based on the hypothesis that binaural cues do not need a higher frequency resolution than the critical bandwidth [8, 9, 10] and the fact that the frequency-dependent inter-aural delay can in many cases be simplified with a constant inter-aural delay [11, 8] that depends on the sound source position. These observations form the basis for so-called parametric methods to describe HRTFs. These methods describe an HRTF pair by three vectors of length b that contain HRTF parameters for a limited set of frequency bands [12, 13]:

- a level parameter vector for the left ear $\vec{p}_l$;
- a level parameter vector for the right ear $\vec{p}_r$; and
- an inter-aural phase parameter vector $\vec{\phi}$.

The parameter analysis and binaural synthesis processes can be implemented using a variety of time-frequency transforms. In the following simple example, HRTFs are assumed to be represented by discrete-sampled complex-valued frequency spectra $\vec{H}_l$ and $\vec{H}_r$ of length k, for the left and right ears, respectively, and a parameter-band definition matrix $\mathbf{Q}$ of size (k, b) that determines the contribution of each spectrum component k to each parameter value b. This matrix typically describes a set of partially-overlapping, adjacent parameter bands. The level parameters $\vec{p}_x$ that represent spectral energy densities in this example, are then given by:

$$\vec{p}_x = \mathbf{Q}^+ \vec{H}_{xx}, \tag{1}$$

with $\vec{H}_{xy}$ the element-wise cross-product of $\vec{H}_x$ and the complex-conjugate of $\vec{H}_y$:

$$H_{xy}(k) = H_x(k) H_y^*(k), \tag{2}$$

and $\mathbf{Q}^+$ the pseudo-inverse of $\mathbf{Q}$:

$$\mathbf{Q}^+ = \left(\mathbf{Q}^T \mathbf{Q}\right)^{-1} \mathbf{Q}^T. \tag{3}$$

The inter-aural phase parameters $\vec{\phi}$ are extracted according to:

$$\vec{\phi} = \mathbf{Q}^+ \angle_u \left(\vec{H}_{lr}\right), \tag{4}$$

with $\angle_u(\vec{H})$ the *unwrapped* phase angles of the elements of $(\vec{H})$. Reconstruction of the power spectrum $\vec{H'}_{xx}$ or inter-aural phase spectrum now follows from multiplication of the representative parameter vector with the matrix $\mathbf{Q}$:

$$\vec{H'}_{xx} = \mathbf{Q}\vec{p}_x. \quad (5)$$

An example HRTF and parametric reconstruction are visualized in Fig. 1. The left-ear magnitude spectrum (in dB) of a sound source positioned at 45 degrees azimuth and 0 degrees elevation is shown as a function of frequency by the dotted line (subject "3" from the CIPIC HRTF database [14]). The reconstructed spectrum, based on 20 magnitude parameters (one parameter for two critical bands) is shown by the solid line. The analysis and reconstruction matrix $\mathbf{Q}$ was based on triangular, overlapping parameter analysis bands (see [12] for details). As can be observed in the figure, the match between original and reconstructed spectrum is very accurate at low frequencies, and less accurate at higher frequencies (from 6 kHz onwards).

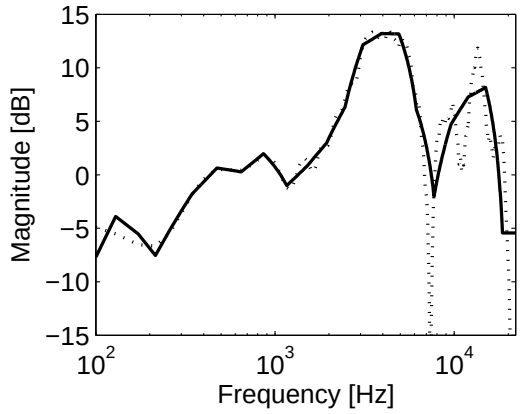

Figure 1: Original (dotted) and reconstructed (solid) HRTF spectrum based on 20 spectral parameters.

The parametric approach for HRTF analysis and synthesis has two important advantages:

- The amount of HRTF-related information that has to be stored in a database is significantly reduced. Depending on the spectral details of the HRTFs (as a result of subject and position dependencies), it seems that 20 to 40 non-linearly spaced parameter bands are sufficient to capture all perceptually-relevant cues, both in terms of perceived timbre [12] as well as sound source localization [13]. It has also been suggested that for many applications, inter-aural phase parameters only have to be reconstructed for frequencies up to about 2 kHz [12] based on the observation that the human auditory system is insensitive to inter-aural fine-structure phase differences above that frequency [15, 16], and that low-frequency time differences perceptually dominate high-frequency envelope time differences [17]. If the inter-aural delay is described by a single delay value for each sound-source direction, 41 to 81 real-valued parameters thus suffice for each sound source position. Moreover, since these parameters closely resemble perceptual attributes, and the auditory sensitivity to changes in these parameters is well known, the number of bits required to store HRTF parameters is relatively small. As simple example of such a "perceptual HRTF compression" method, one could assume the range of spectral levels to be 64 dB with a resolution of 1 dB, thus resulting in 6 bits to describe one parameter value. This number can be significantly reduced by employing additional lossless entropy coding techniques.
- It allows relatively simple integration of HRTF processing with block-based or multi-rate audio compression or processing algorithms for example those based on Discrete Fourier Transforms (DFTs) or Quadrature Mirror Filter (QMF) banks. In this type of transforms, the HRTFs are only reconstructed indirectly from the extracted parameters by applying the parameters directly on the (transformed) signal to be processed. More specifically, the processing boils down to (complex-valued) scaling of DFT bins or sub-band samples, without the need for zero-padding as in conventional DFT-based convolution. Furthermore, static or dynamic interpolation of HRTFs to render virtual sound sources at positions for which no HRTF parameters are available can be performed on the parameter level directly, facilitating simple incorporation of dynamic positioning and head tracking.

Applications and standards

1. MPEG Surround

The first audio codec that supports binaural rendering integrated in the stardard specification is MPEG Surround [18, 19, 20]. MPEG Surround is a relatively new audio compression standard that provides unprecedented compression efficiency for multi-channel audio while providing backward compatibility with legacy mono and stereo infrastructures. It achieves the high compression efficiency by parameterization of the perceptually-relevant properties of multi-channel content, combined with a down mix that is encoded with a (conventional) perceptual audio codec. This method is also referred to as "spatial audio coding". One of the application scenarios where MPEG Surround really shines is audio transmission to mobile devices. In that scenario, headphone reproduction of multi-channel content can be greatly enhanced by binaural processing. In practice, this is however difficult to achieve since battery constraints often do not allow to decode more than two audio channels and to employ high-quality binaural rendering. MPEG Surround solves this hurdle by blending content-related (spatial) parameters and HRTF parameters in an integrated decoding and binaural rendering process, which is outlined in Fig. 2. A legacy mono or stereo bit stream is first decoded by a legacy decoder to result in a mono or stereo down-mix signal. The spatial parameters that were conveyed in the MPEG Surround bit stream are

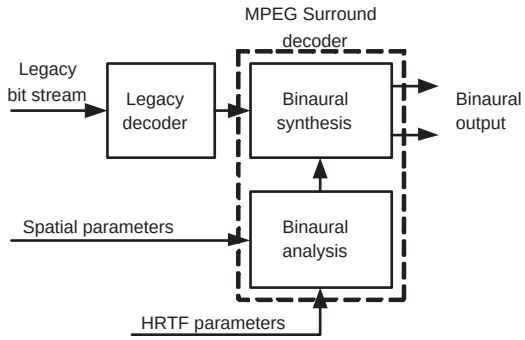

Figure 2: Binaural rendering in MPEG Surround.

sent to a binaural analysis stage. This stage combines the spatial parameters that describe the spatial attributes of the content with the HRTF parameters that describe the spatial attributes of the (virtual) reproduction medium. The combination of spatial and HRTF parameters results in so-called "binaural parameters" that represent the combined spatial attributes of the content and HRTFs and are reinforced on the down-mix signal in a binaural synthesis stage.

The combined process of spatial decoding and HRTF processing provides a significant reduction in processing requirements compared to conventional decoding followed by binaural synthesis while the perceived quality has been reported to *increase* [13].

2. Spatial Audio Object Coding (SAOC)

The next step in spatial audio coding is to allow interactive positioning and modification of individual sound sources or objects contained in a down mix. The aim is to let users freely position sound sources on stereo or multichannel loudspeaker systems, or in a virtual environment reproduced on headphones. Sound sources can also be modified in level and timbre, and user-variable levels of effects such as reverberation can be applied (cf. [21, 22]). This interactivity is provided by means of "object parameters" that are transmitted alongside a legacy down mix that describe certain statistical properties of the object signals. The basic structure for SAOC binaural decoding is given in Fig. 3. The transmitted object parameters are combined with a render matrix and HRTF parameters. The render matrix and HRTF parameters are provided by the user to map the objects to a large set of (virtual) loudspeakers with associated HRTF parameters. The synthesis process is similar to the MPEG Surround decoding process, although several additions exist (see [22]) for extended functionality.

3. Rendering of legacy stereo signals

In the consumer domain, binaural applications often map legacy-stereo signals to virtual speakers, assuming that spatial imaging techniques designed for loudspeaker reproduction will operate correctly in virtual environments. In [23], however, the authors state that this approach is suboptimal since virtual loudspeaker systems are not necessarily compromised by esthetic and cost limitations. In their proposal, stereo audio signals are decomposed

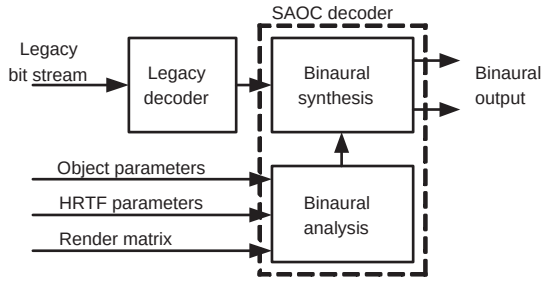

Figure 3: Binaural rendering in a spatial audio object decoder.

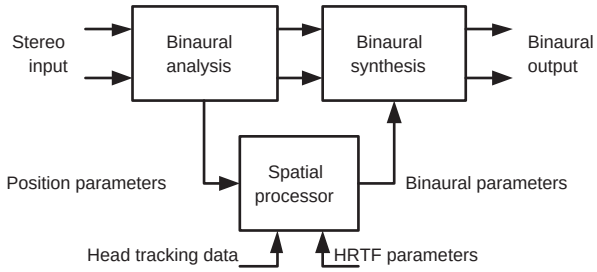

Figure 4: Phantom materialization for headphones playback.

into various frequency bands that coincide with HRTF parameter bands. Subsequently, the intended positions of phantom sound sources are extracted for each band-limited signal (spatial analysis) and may be subject to dynamic modification, for example as a result of head tracking (spatial processing). The position data are combined with HRTF parameters corresponding to the position data. The result of this process is a set of binaural parameters that represent the reproduction of the decomposed audio on a continuum of virtual loudspeakers. As a final step, the acquired binaural parameters are superimposed on the band-limited signals to result in the binaural output signals. The complete process referred to as "phantom materialization" is shown in Fig. 4. This approach has been shown to be preferred by listeners compared to conventional methods simulating a virtual stereo loudspeaker setup [23].

Conclusions

Parametric HRTFs have been shown to provide a cost-effective method to create virtual sound sources, especially when combined with parametric audio coders. The lossy parameterization exploits well-documented perceptual irrelevancies in HRTF magnitude and phase spectra and is incorporated in the recent MPEG Surround standard for multi-channel audio and the upcoming MPEG standard for spatial audio object coding.

So far the parametric approach has been predominantly applied to anechoic impulse responses. For MPEG Surround, dedicated solutions for echoic impulse responses (BRIRs) are provided by means of a morphed-filter approach [13], but a generic parameterization for BRIRs is a more challenging task. Modeling of BRIRs using similar parameters requires an extension of the parameter set with an additional parameter (the inter-aural

coherence) to describe spatial diffuseness of reverberant sound fields, and should be extended with functionality to describe parameter variability across time. The work by Merimaa [24] to parameterize echoic room impulses is most probably a good starting point. Until such a solution is available, the parametric HRTF approach will most probably need separate (parallel) processes to simulate room acoustic properties.

References

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

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

[3] 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.

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

[5] 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.

[6] 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.

[7] D. J. Kistler and F. L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J. Acoust. Soc. Am.*, 91:1637–1647, 1992.

[8] A. Kulkarni and H. S. Colburn. Role of spectral detail in sound-source localization. *Nature*, 396:747–749, 1998.

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

[10] 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.

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

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

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

[14] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, pages 99–102, New Paltz, NY, USA, Oct 2001.

[15] R. G. Klumpp and H. R. Eady. Some measurements of interaural time difference thresholds. *J. Acoust. Soc. Am.*, 28:859–860, 1956.

[16] W. A. Yost. Tone-in-tone masking for three binaural listening conditions. *J. Acoust. Soc. Am.*, 52:1234–1237, 1972.

[17] F. L. Wightman and D. J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.*, 91:1648–1661, 1992.

[18] ISO/IEC JTC1/SC29/WG11. MPEG audio technologies - Part 1: MPEG Surround. ISO/IEC FDIS 23003-1:2006(E), 2004.

[19] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. van de Par. Background, concept, and architecture for the recent MPEG Surround standard on multichannel audio compression. *J. Audio Eng. Soc.*, 55:331–351, 2007.

[20] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K. S. Chong. MPEG Surround - The ISO/MPEG standard for efficient and compatible multichannel audio coding. *J. Audio Eng. Soc.*, 56:932–955, 2008.

[21] J. Herre and S. Disch. New concepts in parametric coding of spatial audio: from SAC to SAOC. In *Proc. ICME 2007*, pages 1894–1897, Beijing, China, 2007.

[22] J. Engdegård, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen. Spatial audio object coding (SAOC): The upcoming MPEG standard on parametric object based audio coding. In *124th AES Convention*, Amsterdam, The Netherlands, 2008.

[23] 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*, 16:1503–1511, 2008.

[24] J. Merimaa. *Analysis, Synthesis, and Perception of Spatial Sound – Binaural Localization Modeling and Multichannel Loudspeaker Reproduction.* PhD thesis, Helsinki University of Technology, Espoo, Finland, 2006.