Perceptual aspects in spatial audio processing

Karlheinz BRANDENBURG[1,2], Bernhard FIEDLER[2], Georg FISCHER[1], Florian KLEIN[1], Annika NEIDHARDT[1], Christian SCHNEIDERWIND[1], Ulrike SLOMA[1], Claudia STIRNAT[1], Stephan WERNER[1],...

[1]Technische Universität Ilmenau, Germany, karlheinz.brandenburg@tu-ilmenau.de

[2]Fraunhofer IDMT Ilmenau, Germany, bdg@idmt.fraunhofer.de

Abstract

Spatial audio processing includes recording, modification and rendering of multichannel audio. In all these fields there is the choice of either a physical representation or of perceptual approaches trying to achieve a target perceived audio quality. Classical microphone techniques on one hand and wave field synthesis, higher order ambisonics or certain methods of binaural rendering for headphone reproduction on the other hand target a good physical representation of sound. As it is known today, especially in the case of sound reproduction a faithful physical recreation of the sound wave forms ("correct signal at the ear drums") is neither necessary nor does it allow a fully authentic or even plausible reproduction of sound. 20 years ago, MPEG-4 standardized different modes for perception based versus physics based reproduction (called "Perceptual approach to modify natural source" and "Acoustic properties for physical based audio rendering"). In spatial rendering today, more and more the perceptual approach is used in state of the art systems. We give some examples of such rendering. The same distinction of physics based versus psychoacoustics (including cognitive effects) based rendering is used today for room simulation or artificial reverberation systems. Perceptual aspects are at the heart of audio signal processing today.

Keywords: Psychoacoustics, Spatial Audio, Perception

1 INTRODUCTION

Few technologies have seen such wide spread application like perceptually motivated audio technologies. When high quality audio coding was introduced some 30 years ago, it meant a paradigm shift in the way signal processing algorithms had to be designed and evaluated. In the case of classic audio coding algorithms (like mp3 and AAC, see e.g. [3]) the perceived quality improved once perceptual aspects were built into the system. At the same time the error as measured by quadratic measures (e.g. SNR) becomes worse. Nowadays many more applications of signal processing went from classic optimization for minimum quadratic error (Eucledian distance) to perception based algorithms.

Figure 1 shows the principal block diagram of a perceptual audio encoder. Perception is introduced in this case via the block "calculation of masking threshold based on psychoacoustics". This block estimates masking as it takes place because of the mechanics on the cochlea, i.e. the inner ear.

Since the days of the introduction of audio coding algorithms, perceptual aspects have been used in other fields, too. If we look at a typical audio transmission or reproduction system, we find topics concerning signal acquisition (recording using e.g. one or several microphones), signal adaptation to the actual task (mixing etc.) and rendering the sound for the intended playback situation, e.g. loudspeakers or headphones.

This includes multichannel recording and reproduction as well as binaural techniques. Overviews on these topics can be found e.g. in [20] or [10]. Another good source for more information is the book edited by Agnieszka Roginska and Paul Geloso [16].

In this paper, we will look especially at spatial rendering techniques. This includes cognitive effects like humans remembering sounds and rooms, influencing the way we recognize sound reproduction as authentic or plausible. Figure 2 emphasizes the feedback path from complex processing in the brain (auditory object analysis) back to earlier stages in the audio processing chain. To state it in a very simple way, what we hear is based on

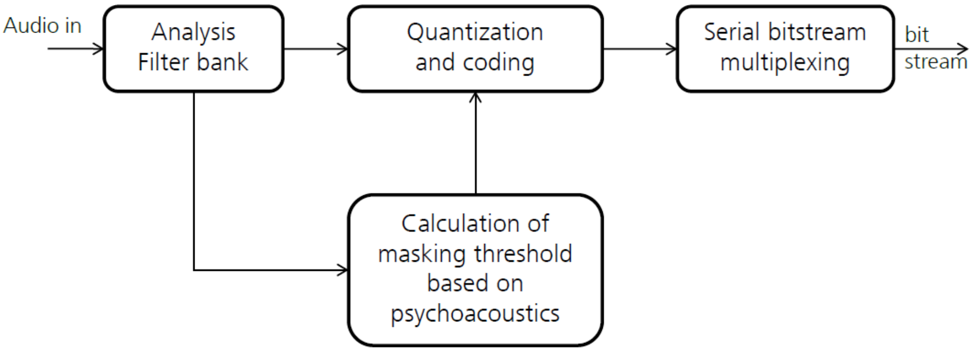
Audio in
Analysis Filter bank
Quantization and coding
Serial bitstream multiplexing
bit stream
Calculation of masking threshold based on psychoacoustics

Figure 1. Principal block diagram of a perceptual audio coder

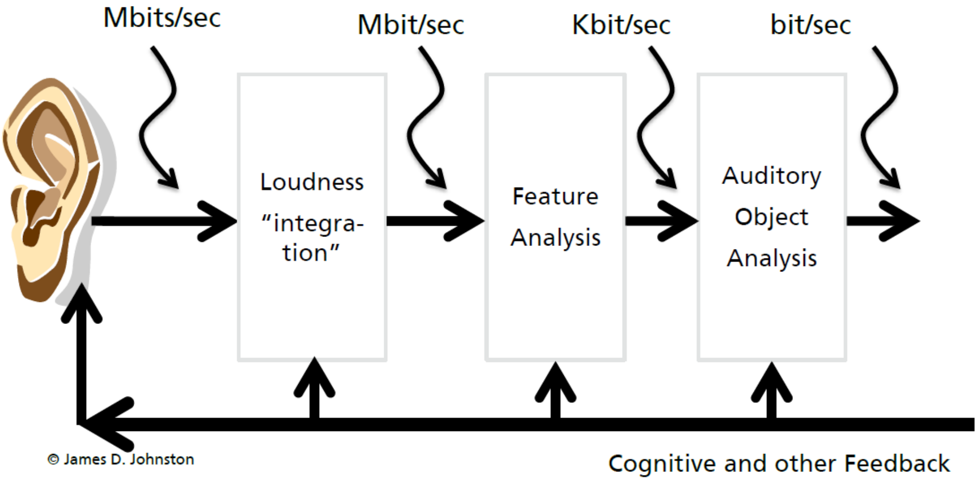

Figure 2. The Audio Processing Chain in Ear and Brain (James D. Johnston)

our expectations. This will be especially important when we look at binaural rendering of audio signals to headphones.

2 SPATIAL RENDERING TECHNIQUES

Spatial sound rendering has been around for many years using stereo or surround sound loudspeaker reproduction. However, these techniques lack real plausibility compared to the experience of sitting in a real concert hall. Newer systems use more loudspeakers to recreate plausible sound fields. These work by representing the sound as channels (as in classic stereo or surround recordings), spherical harmonics or as audio objects which are rendered e.g. using higher order ambisonics, wave field synthesis or other methods.

2.1 Wave field synthesis

Wave field synthesis started as an approach to recreate a complete sound field using many loudspeakers. There are many good papers describing the technology including [20] and chapters in [16]. The basic mathematical description of WFS is the so-called Kirchhoff-Helmholtz-Integral as formulated by Kirchhoff in 1883. Figure 3 shows an intuitive solution to this integral, the Huygen's principle as known from optics: A wave front as

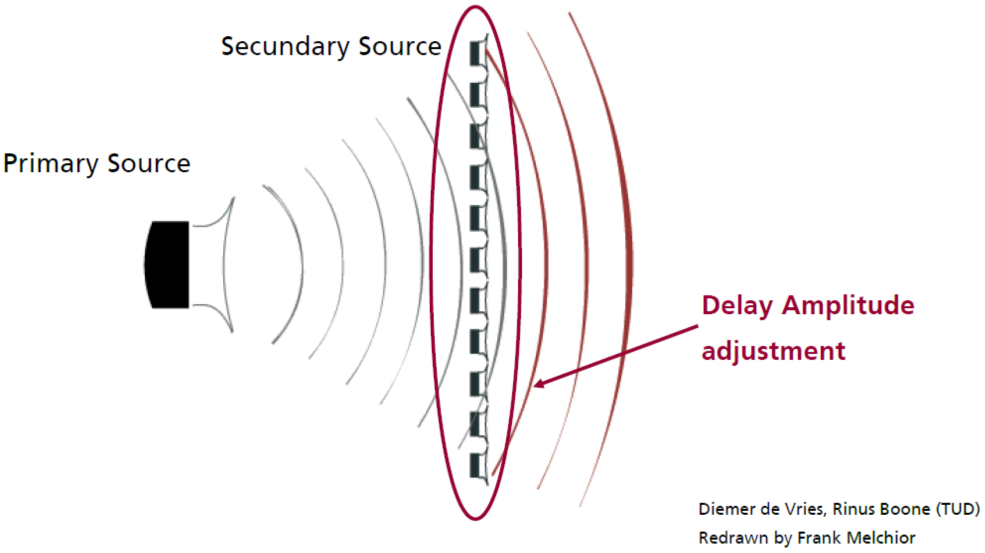
Secundary Source
Primary Source
Delay Amplitude adjustment
Diemer de Vries, Rinus Boone (TUD)
Redrawn by Frank Melchior

Figure 3. Principal of wave field synthesis Diemer de Vries, Rinus Boone (TUD) Redrawn by Frank Melchior

generated from a far away source is reproduced by many secondary sources. This reproduction of course is not precise. For example, at higher sound frequencies, there is aliasing leading to less artifacts in sound reproduction.

WFS in its basic form tries to do recreate the wave forms in the whole listening space in an accurate way. The optimization algorithms are geared toward physical correctness, not towards the best sound quality. As shown for example in [20], WFS can lead to coloration of the sound and other artifacts. Current spatial rendering techniques based on WFS use modified algorithms using a larger spacing, at the same time avoiding some of the coloration. More about this can be found in the chapter about wave field synthesis in [16].

2.2 Ambisonics

In ambisonics a sound field is represented by spherical basis functions, the spherical harmonics. Originally it was introduced by using the spherical harmonics of zeroth and first order. With the increase in computational

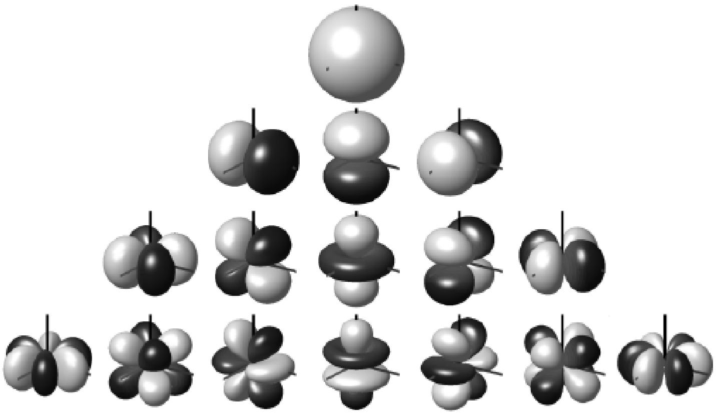

Figure 4. Picture of the elements in higher order ambisonics, from order 0 (top) to order 3

power it is now possible to use higher order spherical harmonics. This results in an increase in angular resolution and enables the reproduction of more complex sound fields.

Recording of an ambisonics sound scene can be done with a so called B-format microphone, for up to first order ambisonics, or with a higher order spherical microphone array for higher order ambisonics. Additional sources can be placed into the recorded sound scene by encoding them into the ambisonics format and specifying the source direction.

The decoding of the ambisonic signals for reproduction can be done for different loudspeaker setups or headphones, i.e. binaural listening.

For reproduction via loudspeakers, ambisonic works via the perceptual effect of phantom sources, similar to vector-base amplitude panning (VBAP), rather than through the physically exact synthesis of a sound field [8]. To get the corresponding gains for the different loudspeakers, different decoding strategies exist, such as AllRAD [23]. Here, the ambisonic signals are decoded to an optimal, virtual loudspeaker setup that is than rendered to the real loudspeaker setup with VBAP.

For a given ambisonic order and above a certain frequency, the sweet spot for reproduction will become smaller than the human head. Thus an additional weighting can be applied, that adjusts the soundfield at higher frequencies accordingly. This is called max-r_E-decoding [5] and can be seen as a deviation from the physically correct synthesis.

In terms of binaural reproduction the achieved quality is directly linked to the quality of the binaural decoding system, i.e. the used set of HRTFs and the processing respectively. More information about ambisonics can be found both in [16] and in [20].

2.3 Binaural

The usual way of rendering audio signals for a long time has still been based on physical models. As an example, in binaural playback over headphones, the task has been to recreate a signal at the eardrums which is physically very near to the actual signal audible at the recording site.

Binaural synthesis describes the reproduction of the sound pressure signals at the ears one would experience in a real acoustic scene. This allows for a headphone-based reproduction by directly playing back binaural signals. Binaural audio content is created by either recording the binaural signals, for example with the use of a head-and-torso simulator or by filtering a monaural audio signal with head related transfer functions (HRTFs). HRTFs describe the direction-dependent filtering caused by the anatomic structures of a person's head, ears and torso. These highly individual filters can be measured directly with in-ear-microphones in an anechoic environment or simulated using appropriate techniques like FEM or BEM.

While over many years, research groups improved the fidelity of binaural playback over headphones by getting closer and closer to the actual signal at eardrums, systems enabling plausible playback for all individuals and all situations are still not available.

The following techniques have been used and can improve the plausibility of binaural synthesis:

- Using HRTF e.g. by recording via a dummy head
- Using measured HRTFs (individualized)
- Using head tracking to modify the HRTF according to the head direction
- Using room simulation and synthesized Binaural room impulse responses (BRIR)
- Trying to let the visual sense and the simulated audio scene match
- Training the listeners to be adapted to the effects of binaural synthesis
- Tracking the movement of a listener and adopting the used BRIRs
- Matching the used BRIRs to the actual listening room

The list above is probably still not exhaustive. It can be noted, that the main reason to not get a plausible reproduction is the deviation from expectation. It is not necessary that every parameter is perfect compared to a real world situation, but if there is too much of a deviation, we get in-head-localization or front back confusion instead of a plausible, externalized impression of "being somewhere else".

3 PERCEPTUAL ASPECTS

Due to individual binaural cues such as Head Related Transfer Function (HRTF), Interaural Time Difference (ITD) and Interaural Level Difference (ILD) the sound fields entering the ears vary between listeners and thus lead to a different perception. In this section, we present some aspects concerning the perceptual side of spatial audio processing.

3.1 Perception

Perception is a chain of processes taking place between the sound as a physical event in the external world and its perceptual registration by the listener [18, p.21] [9]. A sound within a sound field entering the ears is transferred into electro-chemical signals that are further transmitted into brain. Thus the internal representation in the brain can differ between listeners.

As mentioned above, knowledge plays a role during the processing as well. On the one hand, Bottom-Up-Processing takes place where external information is received and activates the process, e.g. sounds arriving at the ears. On the other hand, Top-Down-Processing occurs which is based on knowledge. Then, a listener is able to label the received information by naming a sound “music” or “noise” by referring to all the knowledge the listener has learned with experiences [9].

Hearing with two ears is taken into account in spatial hearing [2]. Once the physical sound enters both ears the binaural signal processing takes place where the signals of both ears are linked together to form the internal representation. In the next stage, the hypothesis-driven stage, a certain sensation is evoked if a hypothesis formed in higher stages of the central nervous system about an appropriate sensation for an expected sound matches the internal representation of the bottom-up processing. The hypothesis-driven stage refers to top-down processing.

The auditory perception depends not only on the external sound field and the sound processing but also other senses such as visual input influence the formation of the internal representation [6, 14]. An example is Hammerschmidt’s and Wöllner’s [11] study that found an influence of the visual quality by different compression rates on the perceived audio quality when listening through headphones. Often the visual domain even dominates the auditory domain, see for example [1]. In order for the internal representation of a scene to be authentic/plausible, the visual scene needs to match the audio scene. Otherwise an observer notices this mismatch and the scene seems unrealistic or confusing.

3.2 Psychoacoustic Effects and Cognitive Aspects

Focusing on the auditory perception, psychoacoustic effects explain how physical events are perceived by a listener. Fastl and Zwicker give an overview of important facts and findings [7]. In the context of spatial audio, the precedence effect (or also Haas effect), masking effects, the Cocktailparty effect and generally auditory scene analysis are introduced.

The precedence effect describes the observation that a sound is located from the direction of the first wave front arriving at a listener and is also referred to the "law of the first wave front" [2], [22]. Masking effects occur when a signal with a low intensity (maskee) is covered by a stronger signal (masker). Two types of masking can make the maskee inaudable: simultaneous (frequency dependent, the closer the frequencies the stronger the effect) and temporal masking (intensity dependent) [7], [21, p. 222].

The cocktail party effect refers to the ability to concentrate on a certain sound event within a complex acoustic scene, e.g. a listener can focus on another person speaking at party where a high noise level with many speakers occurs [15, p. 90f.]. Auditory scene analysis means the decomposition of an acoustic scene into its individual parts [see also [4]].

The way internal representations are formed also depends on cognitive aspects such as Gestalt principles [18]. By describing how auditory or visual events are cognitively organized (either grouped or segregated) by a person the Gestalt principles have found applications in more design oriented fields such as (audio) interface design, music composition and audio mixing [17]. Since the beginning of object-based audio rendering cognitive aspects need to be considered when designing the sound field by objects instead of mixing the sound for several channels, too.

3.3 Perceptually motivated optimization of rendering

As one example of spatial rendering enabling better plausibility, we name dynamic binaural reproduction:

The limited sensitivity of humans to small room acoustical differences, e.g. spatio-temporal reflection pattern, was observed in several experiments [12, 19]. Such knowledge can be used to simplify and optimize dynamic binaural rendering for translation motion. It was found that only by adjusting the level of the direct sound of a BRIR a convincing approaching motion towards a virtual sound source could be realized [13]. The reproduction was rated as plausible by all participants of the study.

4 CONCLUSIONS

In audio signal processing, perceptual models including models of human cognition are necessary for current and future systems. They enable a much better audio quality and can help to get rid of many known artifacts (like in-head localization and front-back confusion) when listening via headphones. Work on these topics is just under way. To emphasize the main result of this paper: Just like optimization is a bad strategy for audio coding, just getting the "correct" signals to the ear drum is by far not enough to enable an authentic and/or plausible experience of spatial audio.

ACKNOWLEDGEMENTS

This work has been supported by grants of the Deutsche Forschungsgemeinschaft (Grants BR 1333/14-1, 1333/16-1 and 1333/18-1).'

REFERENCES

[1] D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257 – 262, 2004.

[2] J. Blauert. *Spatial Hearing*. MIT Press, Cambridge MA, 1997.

[3] K. Brandenburg, C. Faller, J. Herre, J. D. Johnston, and W. B. Kleijn. Perceptual coding of high-quality digital audio. *Proceedings of the IEEE*, 101(9):1905–1919, 2013.

[4] A. S. Bregman. *Auditory Scene Analysis: The perceptual organization of sound*. Cambridge, MA, USA: MIT Press, 1990.

[5] S. Clapp, A. Guthrie, J. Braasch, and N. Xiang. Evaluating the accuracy of the ambisonic reproduction of measured soundfields. *10.14279/depositonce-4103*, 2014.

[6] H. Fastl. Audio-visual interactions in loudness evaluation. In *Proceedings of the 18th International Congress on Acoustics (ICA)*, pages 1161 – 1166. Citeseer, 2004.

[7] H. Fastl and E. Zwicker. *Psychacoustics. Facts and Models*. Berlin, Heidelberg: Springer-Verlag, 3 edition, 2007.

[8] M. Frank. How to make ambisonics sound good. In *Forum Acusticum,(Krakow)*, 2014.

[9] E. B. Goldstein. In K. R. Gegenfurtner, editor, *Wahrnehmungspsychologie. Der Grundkurs*, pages 1 – 13. Berlin, Heidelberg: Springer, 9 edition, 2015.

[10] H. Hacihabiboglu, E. De Sena, Z. Cvetkovic, J. Johnston, and J. O. Smith III. Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics. *IEEE Signal Processing Magazine*, 34(3):36–54, 2017.

[11] Hammerschmidt and Wöllner. The influence of image compression rate on perceived audio quality in musicvideo-clips. In R. Timmers, N. Dibben, Z. Eitan, R. Granot, T. Metcalfe, A. Schiavio, and V. Williamson, editors, *Proceedings of ICMEM 2015. International Conference on the Multimodal Experience of Music*, 2015.

[12] F. Klein, A. Neidhardt, M. Seipel, and T. Sporer. Training on the acoustical identification of the listening position in a virtual environment. In *143rd AES Convention, New York, NY*. Audio Engineering Society (AES), 2017.

[13] A. Neidhardt, A. I. Tommy, and A. D. Pereppadan. Plausibility of an interactive approaching motion towards a virtual sound source. In *144th AES Convention, Milan, Italy*. Audio Engineering Society (AES), 2018.

[14] J. G. Neuhoff. *Ecological psychoacoustics*. Amsterdam, Netherlands: Elsevier Academic Press, 2004.

[15] J. Pierce. Hearing in time and space. In P. C. Cook, editor, *Music, Cognition, and Computerized Sound. An Introduction to Psychoacoustics*, pages 89–103. Cambridge, MA, USA: MIT Press, 2001.

[16] A. Roginska and P. Geluso. *Immersive Sound: The Art and Science of Binaural and Multi-channel Audio*. Taylor & Francis, 2017.

[17] M. T. Shelvock. Gestalt theory and mixing audio. *Innovation in Music II*, pages 1–14, 2016.

[18] R. Shepard. Cognitive psychology and music. In P. C. Cook, editor, *Music, Cognition, and Computerized Sound. An Introduction to Psychoacoustics*, pages 21 – 35. Cambridge, MA, USA: MIT Press, 2001.

[19] B. Shinn-Cunningham. Identifying where you are in the room: Sensitivity to room acoustics. In *International Conference on Auditory Display, Boston, USA*, 2003.

[20] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter. Spatial sound with loudspeakers and its perception: A review of the current state. *Proceedings of the IEEE*, 101(9):1920–38, 2013.

[21] P. Toiviainen. The psychology of electronic music. In N. Collins and J. d'Escrivan, editors, *The Cambridge companion to electronic music*, pages 218–231. Cambridge University Press, 2007.

[22] H. Wallach, E. B. Newman, and M. R. Rosenzweig. The precedence effect in sound localization. *The American Journal of Psychology*, (62):315–336, 1949.

[23] F. Zotter and M. Frank. All-round ambisonic panning and decoding. *Journal of the audio engineering society*, 60(10):807–820, 2012.