Audio Engineering Society

Convention Paper

Presented at the 113th Convention
2002 October 5–8 Los Angeles, CA, USA

Design and Evaluation of Binaural Cue Coding Schemes

Frank Baumgarte[1], Christof Faller[1]

[1] *Media Signal Processing Research, Agere Systems, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A.*

Correspondence should be addressed to Frank Baumgarte (`fb@agere.com`)

ABSTRACT

Binaural Cue Coding (BCC) offers a compact parametric representation of auditory spatial information such as localization cues inherent in multi-channel audio signals. BCC allows to reconstruct the spatial image given a mono signal and spatial cues that require a very low rate of a few kbit/s. This paper reviews relevant auditory perception phenomena exploited by BCC. The BCC core processing scheme design is discussed from a psychoacoustic point of view. This approach leads to a BCC implementation based on binaural perception models. The audio quality of this implementation is compared with low-complexity FFT-based BCC schemes presented earlier. Furthermore, spatial equalization schemes are introduced to optimize the auditory spatial image of loudspeaker or headphone presentation.

1 INTRODUCTION

One important factor that determines perceived audio quality is the capability to produce a "realistic" spatial auditory image. The evolution from mono audio to two-channel stereophonic playback was a major first step towards increased spatial reproduction quality. To-

day, multi-channel systems further enhance the auditory space dimensions. With this background, it is desirable to provide users with more than only one mono channel, even if the channel capacity is very low. However, traditional audio coding tools for joint-channel coding are often insufficient to enable two or multi-channel audio at very low bit rates.

It was shown that Binaural Cue Coding (BCC) allows to deliver two and more audio channels at significantly lower rates than with traditional coders [1]. The concept and applications of BCC were introduced in a series of past publications, e.g. [2][3][4]. In Fig. 1 the generic scheme of a BCC-enhanced mono audio coder is shown. This scheme is efficient for multi-channel audio coding since the BCC parameters require only a few kbit/s while the larger fraction of the total bit rate is used for the mono audio bitstream which does not increase with the number of channels.

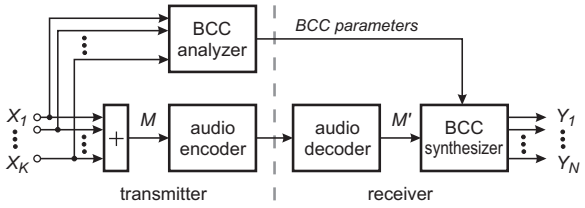

Fig. 1: Generic multi-channel coding scheme based on a mono audio coder.

We distinguish two types of BCC. BCC type II is applied to audio coding with the aim of reconstructing the spatial image of the input multi-channel signal at the receiver (cf. Fig. 1). BCC type I is applied to a number of separate monophonic audio signals with the intention of placing each corresponding phantom source within the auditory scene controlled by the receiver [4][5]. Applications such as tele-conferencing or gaming can be efficiently addressed by BCC type I.

Beyond improved compression, BCC's parametric representation adds enhanced features and functionalities with respect to other audio coders. These enhancements include the flexibility of the receiver to adapt to custom playback systems for optimized spatial image rendering. For instance, the BCC synthesizer inserts different spatial cues depending on whether loudspeaker or headphone playback is used. There is also flexibility in the number playback channels. Moreover, head-related transfer functions can be applied with BCC type I to create 3D images.

The contribution of this paper has two components. The first component is intended to better understand and explain the impact of BCC on the perceived audio quality and auditory image. While some psychoacoustic background was given in [3] based on "classical" binaural cues, a review of spatial perception as part of the auditory scene analysis is given here. This review points out that binaural localization cues such as time and level differences only partly determine the auditory scene. Since BCC is based on binaural spatial cues, it is important to understand the effective perceptual mechanisms that can help to improve the spatial image quality if the cues are encoded with very low resolution. The same mechanisms can help to mitigate cue estimation errors of a BCC analyzer.

The second component addresses two modifications to the BCC synthesizer aiming at improving the obtained spatial image quality. The first proposed modification concerns the spectral representation used for spatial cue insertion. Previously presented BCC synthesizers were based on a DFT representation [1][2]. The uniform time/frequency resolution of a DFT can only roughly approximate the properties of the nonuniform critical bands found in the auditory system. A Cochlear Filter Bank (CFB), however, as used in the BCC analyzer provides an adequate solution. This paper presents an implementation of the inverse CFB and its application to BCC synthesis. This allows to synthesize multi-channel signals from the critical band representation of the mono input signal. The subjective audio quality from both, the CFB-based and the FFT-based synthesizer, is compared by means of listening-test results.

The second proposed modification of the BCC synthesizer includes simple enhancements to improve the spatial image quality specifically for loudspeaker or headphone presentation. The parameterization of the spatial information in BCC permits these enhancements at low complexity. In this paper, the image rendering at low frequencies is improved by equalizing the binaural time-difference cues of a two-channel loudspeaker or headphone presentation with those of a physical source in free field. A spherical head model serves as simulation tool for the acoustics of a presentation via loudspeakers. The simulation results are discussed and verified by data from measured individual head-related transfer functions and psychoacoustic data.

For the purpose of this study, the spatial information used in the BCC scheme is restricted to level-difference and time-difference cues. The cues available to the auditory system, interaural time difference (ITD) and interaural level difference (ILD), must be clearly distinguished from the cues present in the audio signal, inter-channel level difference (ICLD) and inter-channel time difference (ICTD). The binaural cues, ILD and ITD are

also referred to as "classical" localization cues.

2 HOW AUDITORY SCENE ANALYSIS SUPPORTS BINAURAL CUE CODING

This section discusses and summarizes several known perceptual factors and phenomena that contribute to the perceived spatial auditory image quality of a BCC-synthesized audio signal. BCC is based on classical binaural localization cues such as ILDs and ITDs. These cues can only be perfectly synthesized, if the audio signals of the different sources are given separately. Since BCC generally synthesizes the auditory scene based on one mono audio signal that contains all source signals, the synthesized localization cues will be misaligned to a certain degree. Among other factors, the amount of misalignment depends on the degree of overlap in the time/frequency plane of the different sound sources. In the following, we argue that the perceived spatial distortion of the BCC-synthesized sound sources is smaller than would be predicted by taking into account the misalignment of the classical localization cues only.

The perception of an auditory scene (which includes the localization of individual sound sources) is not only based on classical localization cues. From psychology and psychoacoustics it is known that higher levels of the auditory perception process have major impact on the eventually derived image. Most of these processes assist in correcting potentially misaligned binaural cues in order to form a consistent auditory scene. Some of these processes are briefly mentioned in the following. For a more complete treatment of this topic the interested reader is referred to [6].

The auditory system fuses sound components that have similar attributes into one stream. An auditory stream corresponds to an object in vision. Similar attributes of frequency components can consist for instance of a common fundamental frequency of different partials or a similar modulation structure of spectral components. Even if the binaural cues of the different fused components are contradicting, the object will be perceived in one specific location only. Sound components that are received at the same time, are fused on the basis of factors like pitch, timbre, loudness, and spatial origin.

Objects of an auditory scene usually exist over a relatively long time span. The auditory system keeps track of these objects over time in auditory streams. If sound components are assigned to an existing stream, their location depends mostly on the properties of that stream and only to a limited extend on the binaural localization cues associated with the stream. The streaming process is controlled by "organizational cues" that determine how components are fused, grouped and assigned to streams. These cues can override classical binaural cues.

Some factors that promote sequential grouping of sound components into an auditory stream are features that define the similarity and continuity of successive sounds. These include their fundamental frequency, their temporal proximity, the shape of their spectra, their intensity, and their apparent spatial origin.

A factor that influences spectral integration is for instance the similarity of a spectrum that was present earlier and might continue but it may be overlapped with a new spectrum. The old and new spectrum are usually not integrated by the auditory system, i.e. they are not assigned to the same auditory stream. This example shows that fusion as well as source segregation play a major role in auditory streaming.

A very powerful strategy of grouping spectral components is based on their spatial direction. However, even natural spatial cues are often unreliable, thus the auditory system averages among different spatial estimates. The scene analysis process uses the history of a signal to correct momentary spatial estimates. The auditory system uses the fact that sound producing events tend to persist over time, to move only slowly in space, and to give rise to sounds that have a coherent inner structure.

For the reproduction of "natural" sounds like speech and music, many of the perceptual streaming effects are active at the same time. Thus, it can be concluded that the spatial perception of these sounds is only partly controlled by classical binaural cues such as ILDs and ITDs. The auditory streaming processes are in place to make most sense of the given audio input. These processes are able to partially compensate misaligned or contradicting binaural cues, thus, they can mitigate spatial distortions of a BCC-synthesized audio signal with potentially misaligned cues. Moreover, the processes involved in auditory scene analysis contribute to the robustness of the BCC-synthesized audio quality with respect to modifications for minimizing the data rate. These modifications include lowering of the time, frequency, and amplitude resolution of the spatial parameters.

3 DESIGN OF BCC SCHEMES

A BCC synthesizer must be able to generate the proper spatial localization cues in a multi-channel signal. The desired cues are either estimated in a BCC-type-II analyzer or they are assigned according to a source index transmitted from a BCC-type-I analyzer [4]. A straightforward approach for the design and evaluation of a BCC synthesizer is to use a suitable binaural perception model as a reference system. Such a model can be used to esti-

mate the spatial cues of a reference signal and the corresponding BCC-synthesized signal. The difference of the estimated spatial cues is a measure of BCC synthesizer performance. From that perspective, a BCC synthesizer would perform optimally if it would be realized as the inverse binaural perception model. However, more sophisticated binaural models are not perfectly invertible. Moreover, the inversion can result in a BCC synthesizer with unreasonably high complexity.

For these reasons, we propose a "moderate" solution to the inverse problem. This solution is based on the perceptually motivated BCC analyzer introduced earlier [3]. The core of that analyzer is a Cochlear Filter Bank [7] that approximates the time and frequency resolution of the peripheral auditory system. A BCC synthesizer is presented here that approximates the inverse analyzer. This is an ideal scheme with respect to the design goals formulated. Its performance will be compared with previously described schemes based on the FFT. The main motivation to use filter banks different from the CFB is a reduction of computational complexity.

3.1 Analysis and Synthesis based on a Cochlear Filter Bank

Suitable binaural models [8][9] apply a filter bank as first processing stage that has similar properties as the frequency decomposition found in the auditory system. A filter bank with equivalent properties but with a particularly efficient implementation is given in [7]. This CFB is used for the BCC analysis. The corresponding inverse CFB is described here and is applied together with the forward CFB for the BCC synthesis. Figure 2 shows a block diagram of the forward and inverse filter-bank structure for this application.

The forward structure [7] consists of a low-pass filter (LPF) cascade with down-samplers. Each low-pass output is processed by a high-pass filter (HPF) to generate the band-pass signals at the CFB output. These outputs represent "critical-band" signals that overlap spectrally. The input audio signal can be approximately reconstructed from the critical-band signals by applying the inverse filter bank. The inverse CFB includes the reverse structure ($\mathrm{CFB_R}$) and time reversal operations. It is derived by reversing the signal flow, replacing down-samplers by up-samplers, and by time-reversing the filter impulse responses of the forward CFB. The time-reversal of the impulse responses, however, is substituted by applying time-reversal to the input and output signals of the inverse CFB. This substitution allows a less complex implementation than the reversal of the IIR-type filter responses. For signals that are not time limited,

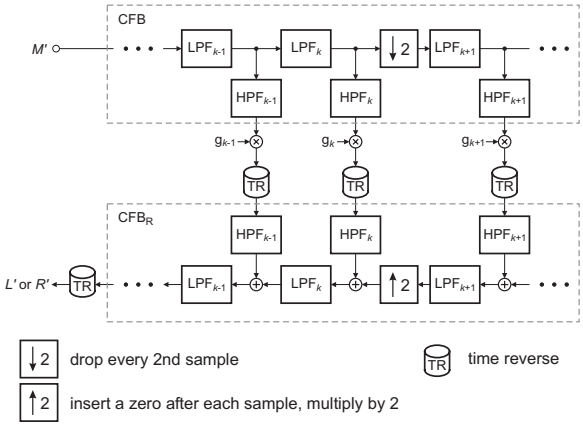

Fig. 2: Structure of the forward Cochlear Filter Bank (CFB) and its inverse. The inverse CFB includes time reversals (TRs) and the reverse CFB ($\mathrm{CFB_R}$). The low-pass filters (LPFs) and high-pass filters (HPFs) are identical for the CFB and $\mathrm{CFB_R}$. The coefficients g_k are needed for equalization.

the time-reversal can be implemented by block-wise processing with temporal overlap, as described in [10]. The signal processing for the experiments reported in this paper was done by time-reversing all ($\mathrm{CFB_R}$) full-length band-pass input signals at once and time-reversal of the output signal after filtering as outlined in Fig. 2. The gain factors, g_k, to the band-pass signals are needed to compensate for the energy increase due to the band overlap of neighboring filters.

For the BCC analysis, only the forward CFB is necessary and complemented by a simple inner hair cell (IHC) model in each band. This analyzers is described in more detail in [3] and it is shown in Fig. 3. The IHC model includes a half-wave rectifier and low-pass filter.

The ICTD, τ, is estimated by locating the maximum of the coherence function (normalized cross-correlation [11]). The ICLD estimation is based on the ratio of the estimated band powers. The power estimation uses a recursive low-pass filter applied to the squared inner hair cell model outputs. The ICLD, ΔL, is the ratio of the delay-compensated power estimates from both channels and converted to the logarithmic (dB) domain.

The coherence function is computed for a limited range of delays because auditory localization based on ITDs "saturates" at the extreme left or right of the auditory space for delays larger than approximately 1 ms.

An overview of the CFB-based BCC synthesis scheme is

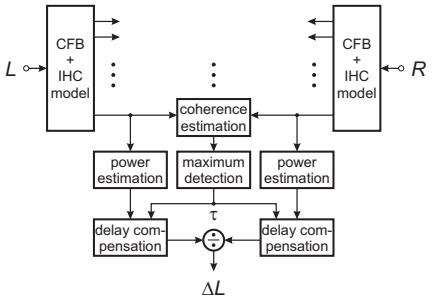

Fig. 3: Block diagram of CFB-based BCC analyzer.

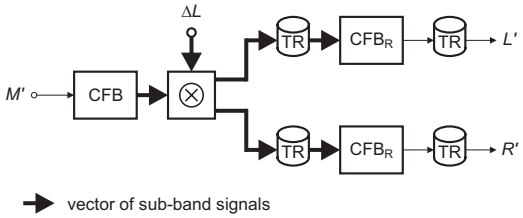

Fig. 4: Block diagram of CFB-based BCC synthesizer (see Fig. 2 for legend).

given in Fig. 4. The mono audio signal is decomposed into critical bands by the forward CFB. The estimated cues are applied to the band-pass signals. Currently, only the ICLD synthesis is implemented. This is done by modifying the gains, g_k, in Fig. 2 according to the estimated ICLDs, ΔL, for the left and right channel. In principle, the synthesis of ICTDs and coherence modifications can also be done in the band-pass signal domain.

3.2 CFB versus FFT-based synthesis

BCC synthesizer schemes based on the FFT were introduced in [1][12]. This section evaluates the subjective performance of those FFT-based synthesizers with respect to the CFB-based synthesizer introduced here. According to the BCC design goals, we refer to the CFB-based scheme as the reference scheme that is expected to achieve optimum quality. The evaluation includes FFT-based synthesizers with different time/frequency resolution. Thus, the impact of the effective block size as a major design parameter of FFT-based systems is also considered. The excerpts were processed with the CFB-based analyzer to estimate the ICLDs. These ICLDs were resampled to match the time/frequency resolution of the FFT-based synthesis. The synthesizer introduces the ICLD in each band when generating the reconstructed stereophonic signal according to Fig. 5. The

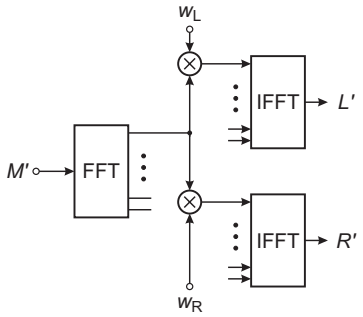

Fig. 5: FFT-based BCC synthesis scheme for interchannel level differences (ICLDs).

sub-band representation of the mono signal M' is computed in the synthesizer by applying a forward FFT of the same size as the inverse FFT. The weighting factors w_L and w_R are derived from the ICLDs.

Table 1 lists the five synthesizer configurations used in the test. The filter bank (FB) type is either CFB or FFT. Four different FFT sizes were used to evaluate the impact of different time/frequency resolutions on the audio quality.

label	*FB*	*size*
A	CFB	98 non-uniform bands
B	FFT	2048
C	FFT	1024
D	FFT	512
E	FFT	256

Table 1: Filter bank parameters.

Four different stereophonic audio excerpts, each with a duration of approximately 10 s sampled at 32 kHz, were used in the test. Table 2 summarizes their contents. The first three excerpts were generated by mixing two mono signals of the same category, e.g. male and female speech. One of the sources was amplitude-panned to the left and the other to the right side by a level difference of +10 or −10 dB. Sources of the same category were mixed because they are most likely to have an impact on each other's phantom image due to their similar time-frequency characteristics. The audio excerpts were selected from a collection of critical stereophonic signals with the objective of having different types of content and the most critical material for ICLD imaging in the test. The fourth excerpt is a stereophonic recording of applause. It is known as a critical signal for joint-stereo coding since the spatial image is very dynamic.

excerpt	*category*	*left*	*right*
1	speech	male	female
2	singing	tenor	soprano
3	percussions	castanets	drums
4	applause	(stereo recording)	

Table 2: Audio excerpts. The last two columns contain the sources of excerpt 1, 2, and 3, that are placed to the left or right side of the spatial image by imposing a level difference (amplitude panning).

The test items, including the reference excerpt with its differently processed versions, were presented over loudspeakers. The test was performed by each subject sitting at the standard listening position ("sweetspot") for conventional stereophonic playback. The test items were played back from a computer under the subject's control and with comfortable volume. The five participating subjects were asked to grade different specific distortions and the overall audio quality of the processed excerpts with respect to the known reference, the original excerpt. The four different grading tasks of this test are summarized in Table 3. Task 1 and 2 assess the two properties of the reproduced spatial image that are thought to determine the spatial image quality, width and stability. Task 3 evaluates distortions introduced by the stereophonic synthesis that do not result in image artifacts. For example, aliasing and blocking artifacts should be detected here. Task 4 is an important measure for global optimization of BCC.

	task	*scale*
1	image width	stereo...mono
2	image stability	stable...unstable
3	audio quality disregarding spatial image distortions	ITU-R 5-grade impairment
4	overall audio quality	ITU-R 5-grade impairment

Table 3: Tasks and scales of the subjective test.

During the test, each subject was able to randomly access each test item processed by the five different BCC synthesizers and the reference by using the corresponding "Play" button of a graphical interface. This play function stops a possibly active audio output at any time, such that the subject can do quick initial listening through all items before proceeding with a more thorough evaluation. The gradings were entered via graphical sliders that are permanently visible for all test items and can be adjusted at any time to reflect the proper

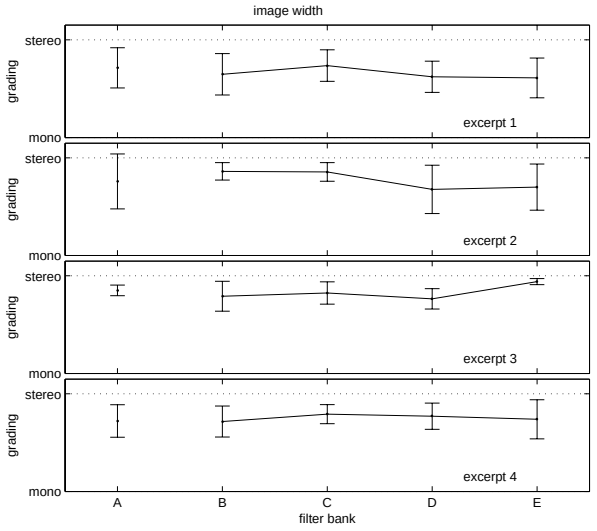

Fig. 6: Perceived image width and 95% confidence intervals.

grading and ranking. It is important to note that subjects were specifically asked to pay attention to the rank order of the test items. The feature of being able to play the items according to their rank order greatly facilitates this task as opposed to other testing schemes that allow to listen only once to each item in a pre-defined order. The ordering of the synthesizers was randomly chosen for each subject and each excerpt but not changed during the four different tasks performed for each excerpt.

The experimental results are shown for the individual excerpts only. Averaging over the gradings of different excerpts cannot be justified due to substantially deviating ratings. The gradings of each task will be discussed in the following sub-sections.

Image width

The gradings for image width are shown in Fig. 6 for each excerpt and each synthesizer with respect to the reference. Apparently, all synthesizers reduce the image width for all test items. For excerpt 2 there is a trend toward a smaller image width with reduced FFT size. This trend is reverse for excerpt 3. This result can be explained by the more stationary character of excerpt 2 (singing) in contrast to the non-stationary excerpt 3 (percussions) which requires a higher time resolution for a proper image reproduction. The overall performance of the 256-point FFT is similar to the CFB performance.

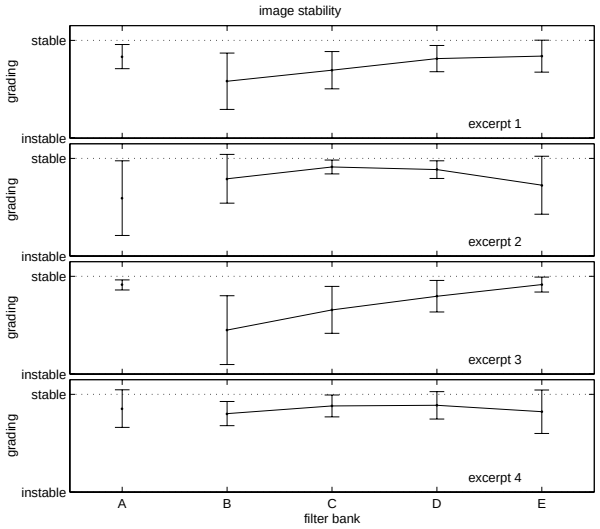

Fig. 7: Perceived image stability and 95% confidence intervals.

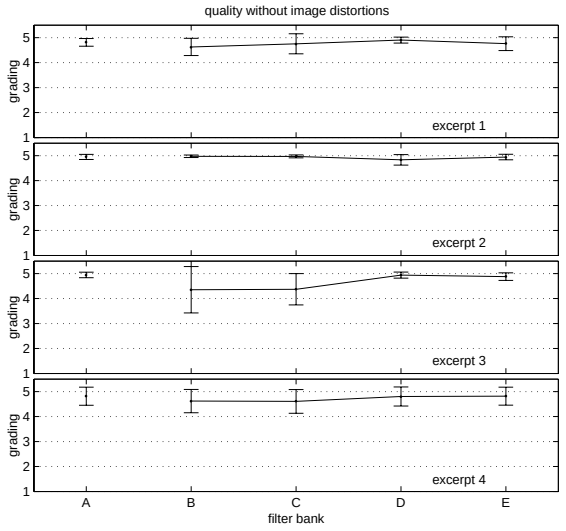

Fig. 8: Perceived audio quality disregarding spatial image distortions and 95% confidence intervals.

Image stability

Gradings for image stability are given in Fig. 7. The image stability is best if the virtual sound source location is stationary. Source locations are well defined for the reference excerpts 1, 2, and 3. However, for excerpt 4 (applause) each source is only active for a short time so that a moving source cannot be detected. That is why excerpt 4 appears close to "stable" for all synthesizers. From the remaining excerpts, 1 and 3 are more critical than 2. For excerpt 1 and 3, the stability increases consistently with time resolution of the FFT-based synthesizer. For excerpt 2 an FFT with medium time resolution shows best gradings. The CFB-based synthesis performs equally or better than the FFT-based schemes except for excerpt 2.

Quality, disregarding image distortions

In task 3 the audio quality is assessed without considering image degradations. The results in Fig. 8 show no significant degradations except for excerpt 3 which appears critical for the size 2048 and 1024 FFT. The time resolution is apparently insufficient for this excerpt (percussions) containing many transients.

Overall quality

The overall quality gradings in Fig. 9 show the integral impact of all noticeable distortions on audio quality to facilitate the selection of the synthesizer with best overall performance. Obviously, the overall quality reflects the influence of the degradations assessed in task 1, 2, and 3 and it combines these individual components into a perceptually meaningful global measure.

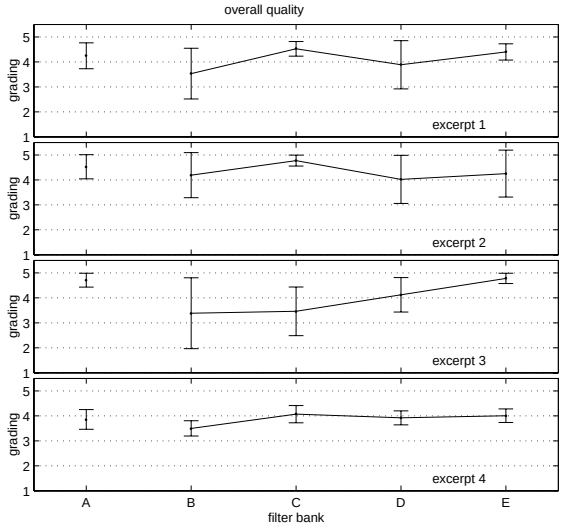

Fig. 9: Perceived audio quality and 95% confidence intervals.

From visual inspection it is concluded that among the FFT-based schemes the 256-point synthesis has best performance for the test excerpts followed by the 512-point FFT. The 1024 and 2048 size FFT show significantly reduced quality for at least one excerpt. The 256-point

FFT has a clear advantage over longer FFTs for excerpt 3 (percussions) which requires a high time resolution. For the more stationary excerpts 1 and 2, an FFT length between 256 and 1024 reaches about the same quality. On average the CFB-based synthesis has the same performance as the 256-size FFT.

4 SPATIAL EQUALIZATION

The parameterization of spatial information in BCC allows for simple manipulations to improve the spatial image quality. Specifically, the acoustical properties of the playback system can be taken into account when rendering the multi-channel audio signal. This is done with the aim of controlling the binaural cues such that the optimum spatial image quality is achieved.

In this paper we address two specific playback scenarios that are considered highly relevant for consumer applications. The first scenario is a standard two-channel (stereo) loudspeaker playback system. The second is playback via headphones. In both cases we assume that only ICLDs are transmitted for BCC type II to minimize the data rate.

For BCC type II the approach for cue modification must be based on the acoustics of the recording and playback setups as well as the psychoacoustics of sound localization. Most two-channel recordings are produced with microphone and mixing techniques that "encode" the spatial information in the ICLD of each sound source. These techniques commonly create a virtually frequency-independent ICLD for each source. For example, a coincidence microphone or a spaced microphone pair are approximately conform to this rule. Obviously, mixing of a mono source by amplitude panning introduces a frequency-independent ICLD. Therefore, we assume that the auditory scene "intended" by the producer, e.g. a sound engineer, is mostly encoded in the ICLDs of the recording. The spatial cue manipulations of BCC do not have the purpose to change the intended auditory scene but to enhance its reproduction quality for a given reproduction system.

For type I BCC, the image rendering can be optimized without taking into account the recording technique. Since conceptually the rendered auditory image is intended to be controlled by the user, there is no need to adjust the image according to a reference. Thus, spatial equalization is applicable to type I BCC with less restrictions.

According to the Duplex Theory well established in psychoacoustics [13][14], ILDs are the most salient spatial cues above ca. 1.5 kHz. At lower frequencies, ITDs are the most relevant cues for the auditory system for determining the sound source azimuth in the horizontal plane. Thus, it is important to provide suitable ITD cues at low frequencies by the playback systems.

Loudspeaker Playback

In the following we use a spherical rigid head model to simulate ITDs and ILDs in free field. The model provides us with reference data for one sound source in the horizontal plane. It is also used to simulate the ITDs and ILDs generated by a standard stereophonic playback system in free field. It will be shown, how the ICLDs can be equalized, such that the ITDs at low frequencies closely match those for the "natural" case of a single sound source. This equalization can be considered as a generalized panning-law which is frequency dependent. For playback via headphones, we derive a mapping of ICLDs to ICTDs such that the ICTDs approximate "natural" ITDs.

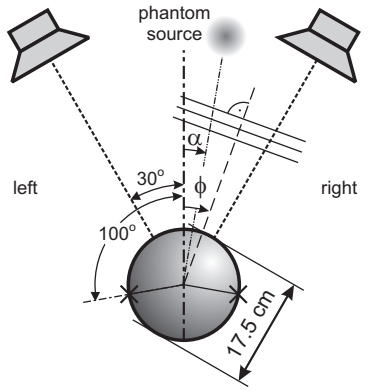

Fig. 10: Parameters for the spherical head simulation.

The spherical head model used in the simulations is shown in Fig. 10. The sphere diameter is 17.5 cm with the ear-canal entrances at $\pm 100^\circ$ with respect to the forward direction. The sphere is assumed to be rigid. The standard two-channel loudspeaker configuration with transducers at $\pm 30^\circ$ azimuth is also shown. However, for single source simulations sound impinges from the azimuth angle ϕ. The angle α indicates a phantom source azimuth. Similar spherical head simulations were published in the past (for an overview see [11]). Therefore, a description of the detailed simulation method is omitted here. The model can be expected to yield accurate simulation results in the framework of this study for frequencies up to 3 to 4 kHz. This is partially verified by the measurements described below.

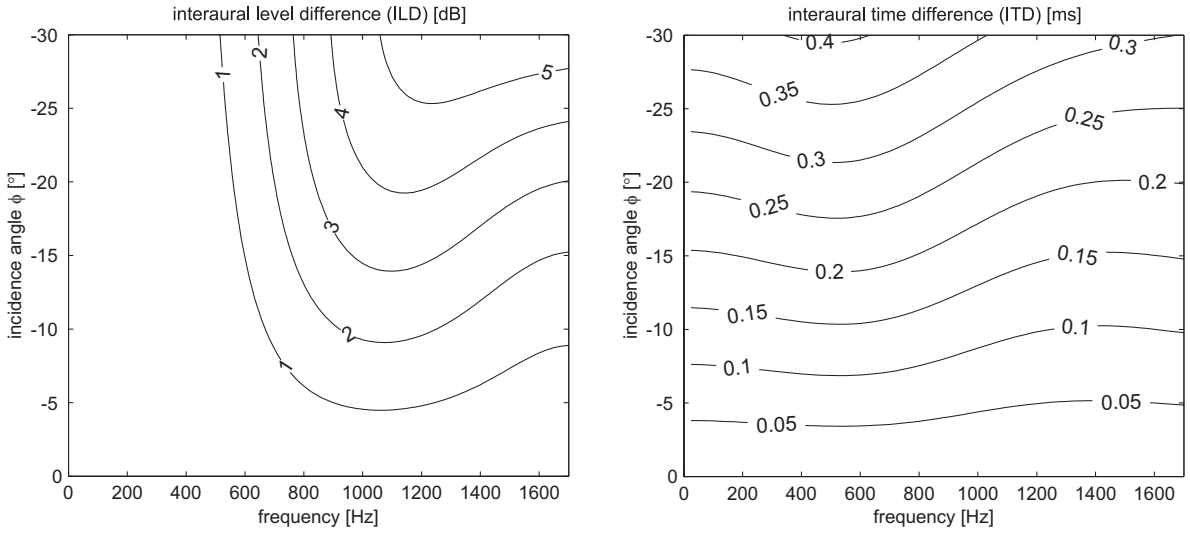

Fig. 11: Simulated ILD and ITD contours in free field for a single sound source at the azimuth angle ϕ in the horizontal plane. A positive ILD indicates a higher sound level at the left ear. A positive ITD indicates that sound arriving at the right ear is delayed with respect to the left ear.

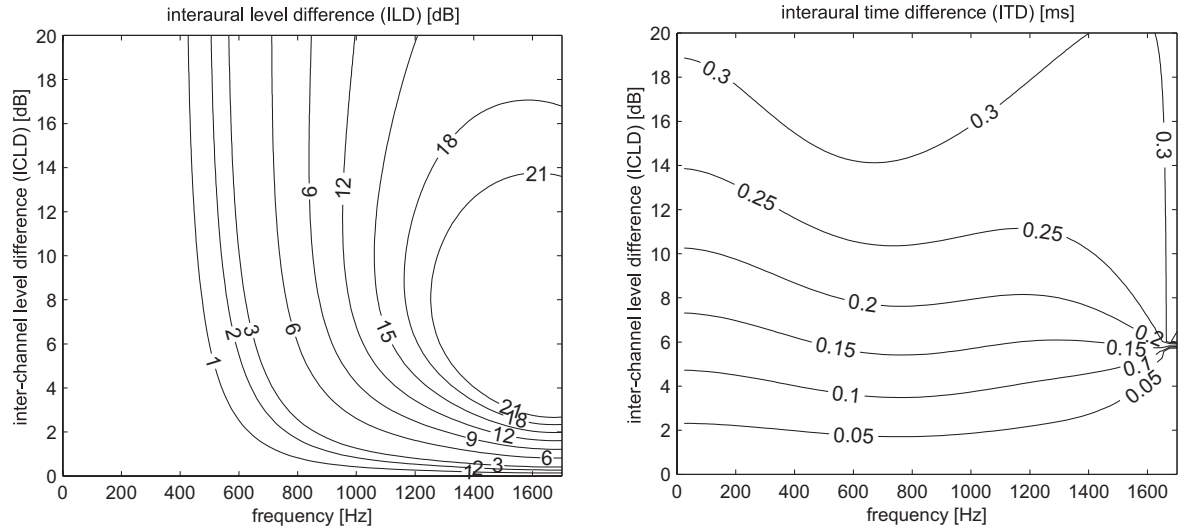

Fig. 12: Simulated ILD and ITD contours for a standard stereo loudspeaker configuration as shown in Fig. 10 in free field. The two simulated loudspeaker signals differ only by a level difference (ICLD).

Simulation results for a single sound source in free field are shown in Fig. 11. Due to the sphere dimension, ILDs gradually disappear below 1 kHz since the wavelength becomes large compared to the sphere. The ITDs are mainly determined by the different path lengths to the two ears. Thus, they vary only slightly with frequency. It is assumed that the azimuth angle of a rendered phantom source is controlled more accurate the closer the ILD and

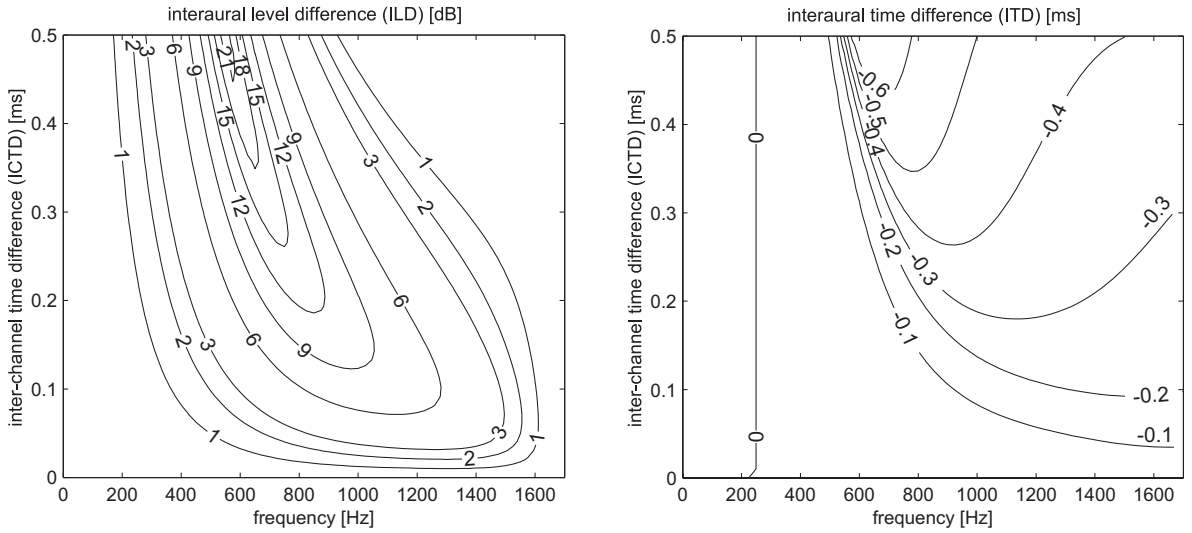

Fig. 13: Simulated ILD and ITD contours for a standard stereo loudspeaker configuration as shown in Fig. 10 in free field. The two simulated loudspeaker signals differ only by a time difference (ICTD).

ITD matches the natural profile shown in Fig. 11.

The ILDs and ITDs simulated for a two loudspeaker standard playback situation are shown in Fig. 12. Only level differences (ICLDs) were used, which is equivalent to mixing by amplitude panning. The ILD shows a high gradient around 1.6 kHz for small changes of the ICLD. The simulated ITD shows that the ICLDs are basically transformed to ITDs at frequencies below 1.6 kHz, i.e. the level difference at the loudspeakers translates into a time difference at the two ears. This effect is well known and widely exploited in two channel recording and mixing.

For completeness and comparison, simulated ILDs and ITDs caused by ICTDs only are shown in Fig. 13. Though less relevant for the focus of this paper, the right graph indicates that ICTDs are not suitable to control the ITDs for loudspeaker playback at frequencies below 1 kHz.

For the purpose of generating ITDs that closely approximate the natural time-difference cues at low frequencies we need to relate the ITD profile in Fig. 11 to the ITD profile in Fig. 12. We can approach this problem by first estimating the phantom source azimuth α for frequencies above 1.5 kHz. Since in that frequency range the localization will be dominated by the ILDs that largely depend on the ICLDs. In a second step, the ICLDs below 1.5 kHz are adjusted such that the resulting ITDs approximate the ITDs of a single sound source at the azimuth $\phi = \alpha$. With the assumption that the phantom source azimuth α is determined by only the ICLD ΔL at frequencies above 1.5 kHz, we can construct a model describing the dependency of α from ΔL. Conceptually, the model can be derived either from the physics of the acoustical system, e.g. a spherical head model, or from empirical psychoacoustic measurements. However, the data from the physical spherical head model does not show enough resemblance of the ILD contours for a single source with the standard stereo playback to justify a reasonably simple model for this application. Therefore, the dependency is derived from psychoacoustic data [11][15]. For a standard stereo loudspeaker configuration and averaged over different wide-band material, e.g. speech and noise, the phantom source azimuth can be approximated by (1).

$$\alpha = \Delta L \left(\left[\frac{1}{S} \right]^t + \left| \frac{\Delta L}{\delta} \right|^t \right)^{-1/t} \qquad (1)$$

The angle α indicates the phantom source azimuth in degrees. The slope parameter of the model was chosen as $S = 2.4°/\mathrm{dB}$. The transition into saturation is controlled by the parameter t which is chosen as $t = 4$. The loudspeaker azimuth is $\delta = 30°$. The resulting characteristic is shown in Fig. 14.

Given the model (1) the ITD contours at frequencies

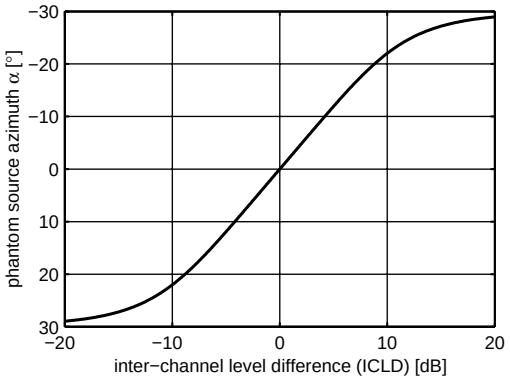

Fig. 14: Model characteristic used for predicting the phantom source azimuth from a given inter-channel level difference in a standard stereo loudspeaker configuration.

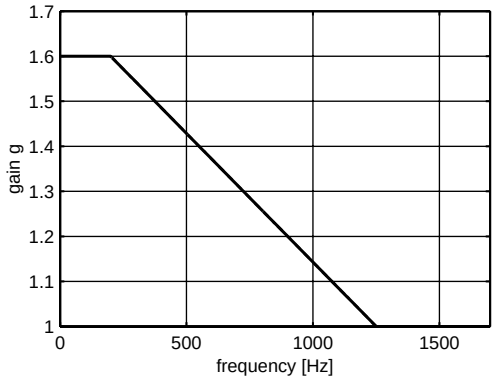

Fig. 15: Gain function for spatial equalization.

below 1.5 kHz can be modified such that they correspond to the same phantom source azimuth as induced by the ICLDs at frequencies above 1.5 kHz. For that purpose we used a frequency dependent amplification of the ICLDs at low frequencies. The amount of amplification was adjusted to obtain the best match with the simulated "natural" ITD contours in Fig. 11. The gain g is applied to modify the ICLDs ΔL according to (2).

$$\Delta L'(f) = g(f)\ \Delta L(f) \tag{2}$$

Figure 15 shows the derived gain function. It has a constant slope between 200 and 1250 Hz. ICLDs above 1250 Hz are not modified.

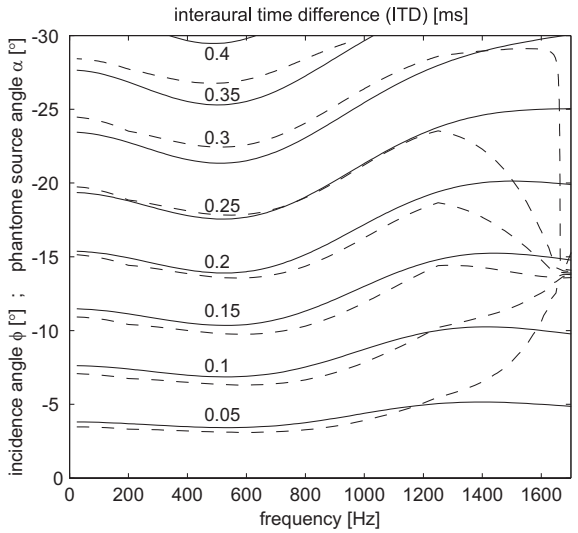

Fig. 16: Equalized ITD contours (dashed) simulated for a standard stereo configuration and simulated ITD contours (solid) for a single sound source with an azimuth of ϕ. For the equalized ITDs the ordinate refers to the predicted phantom source azimuth α.

The resulting close match of the ITDs when applying the modified ICLDs according to (2) is shown in Fig. 16. A more precise match would be possible with a nonlinear gain function. However, the increased complexity of such a function cannot be justified if we consider the variance of the data the model (1) is based on. The non-equalized ITD contours are shown in Fig. 17 for comparison.

For verification, the simulation results are compared with measured head-related transfer functions (HRTFs). The averaged low-frequency ITDs of 45 subjects from data base [16] were calculated for different azimuths and 0° elevation. The data is plotted in Fig. 18 together with the simulation results. The ITDs of the model appear to be slightly larger than for the HRTFs. This might be due to the not exactly matching head dimensions, which are 17.5 cm for the model sphere compared to an average subject head width of 14.5 cm and depth of 20 cm. Moreover, the sound source distance was only 1 m in the HRTF measurements. However, the simulation results approximate the measurements closely enough for the purpose of this study.

To further explore and verify the effect of the stereophonic playback situation on the ITDs, HRTFs for the stereophonic loudspeaker azimuths of $\pm 30^\circ$ were super-

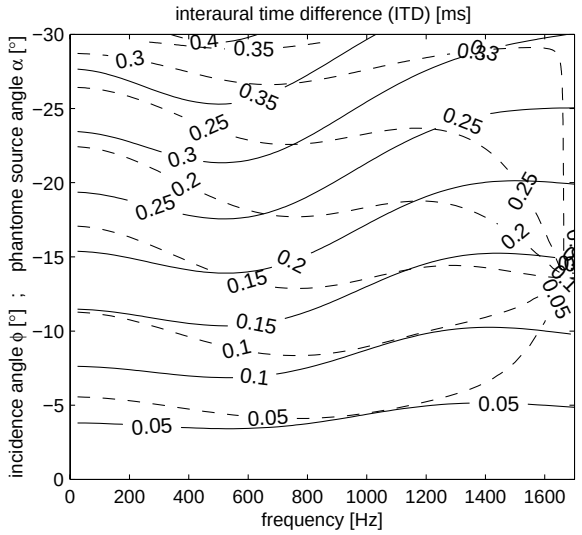

Fig. 17: Non-equalized ITD contours (dashed) simulated for a standard stereo configuration and simulated ITD contours (solid) for a single sound source with an azimuth of ϕ. For the equalized ITDs the ordinate refers to the predicted phantom source azimuth α.

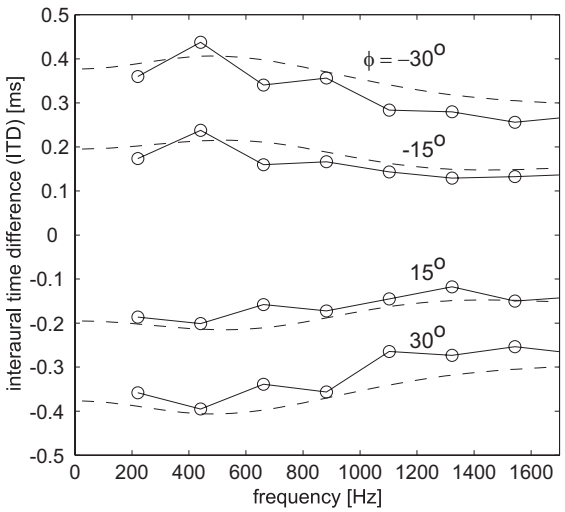

Fig. 18: Average interaural time difference of 45 subjects calculated from their HRTF for different azimuths ϕ (solid). The spherical head model data is included for comparison (dashed).

imposed using an ICLD of +9 or −9 dB. The resulting ITDs for the stereophonic case are shown in Fig. 19 together with the ITDs for a source at either +20° or −20° which are the reference conditions. The dotted line shows the ITDs obtained with spatial equalization according to (2). The results are in agreement with the spherical head simulations shown in Fig. 17 for frequencies below 1.2 kHz. This comparison shows that the spherical head model is suitable for ITD simulations at low frequencies.

The proposed spatial equalization for loudspeaker playback can be applied, in principle, to any two-channel playback system independently of whether BCC is applied or not. In fact, for generic two-channel loudspeaker systems similar equalization schemes have been suggested by different authors before, e.g. [17][18]. However, these schemes amplify the difference signal of the two channels at low frequencies or attenuate it at high frequencies. Thus, they apply a different gain characteristic than the one described here. Moreover, our proposed technique ensures that the power sum of the left and right signal is not modified by the equalization. It thus allows to better preserve the loudness level.

Further evidence that the proposed equalization can enhance the spatial image quality can be drawn from [19]. In that study, subjects were asked to adjust the phantom source azimuth as close as possible to a fixed physical reference source position. Different narrow-band stimuli were used and adjusted using a (wide-band) ICLD. The results show consistently, that for decreasing stimulus frequencies below 1.7 kHz the magnitude ICLD must be increased to keep the phantom source at the same azimuth position (see Figs. 7–9 in [19]). The necessary ICLD change with frequency depends on the signal category and it is in the same order of magnitude as the proposed gain function (2) for low frequencies.

Headphone Playback

For headphone playback it is important to create both, ICLDs and ICTDs, consistently. In this case, ICLDs and ICTDs are virtually identical to the ILDs and ITDs, respectively. Standard two-channel recording techniques do usually not provide ICTDs that are consistent with either the recorded ICLDs or "natural" ITDs of the recorded scene. However, for most recordings the ICLDs are closely related to the "intended" auditory scene as explained above. Thus, consistent ICTDs can be generated from the known ICLDs. The criterion for consistency used here is purely based on empiric psychoacoustic lateralization data ([11] Fig. 2.68 and 2.80). The amount of ICTD introduced in the two-channel output

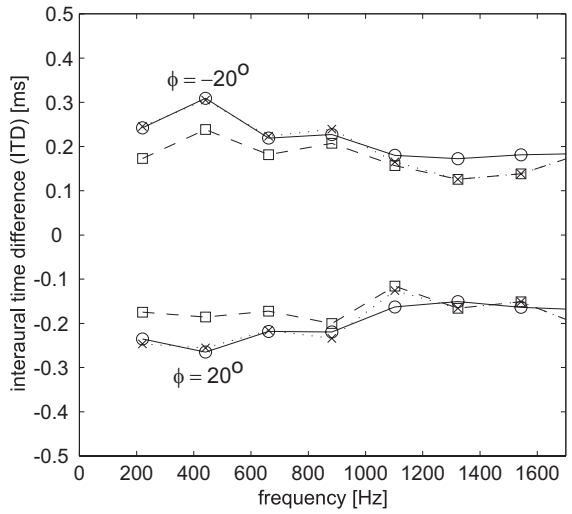

Fig. 19: Average interaural time difference of 45 subjects calculated from their HRTF for an azimuth of ±20° (solid). Average interaural time difference calculated form their HRTF for the standard stereo configuration with ±9 dB ICLD (dashed) and with spatial equalization (dotted).

is adjusted such that the resulting lateralization is equal to the lateralization caused by the corresponding ICLD. For this mapping of ICLD to ICTD a linear, frequency-independent function can be fitted to the data. The resulting mapping function is

$$\tau = \lambda \, \Delta L \tag{3}$$

with $\lambda = 80\mu$s/dB. Based on the arguments given above that ITDs are the most important localization cues at low frequencies, the generation of ICTDs for headphone playback can be reduced to frequencies below 1.5 kHz without significant perceptual impact. Such a reduction leads to lower complexity and it can also avoid potential audible distortions of synthesized time-varying delays at higher frequencies.

4.1 Subjective evaluation of spatial equalization

Informal listening tests were done to evaluate the spatial equalization for loudspeaker playback and the ITD generation for headphone playback. While the headphone playback with ITDs shows a major improved spatial image, the spatial equalization for loudspeakers had only a weak effect on the spatial image in our tests.

For the loudspeaker presentation, we used an anechoic chamber and two different listening rooms, each with different professional equipment and two subjects. The listening room walls were treated with sound absorbing material to reduce reverberation. The size of both rectangular rooms was about 6 m X 6 m and 2.4 m height. The test signals used were a click train of 1.3 s duration containing 10 clicks and 10 s of white noise. Different band-limited versions were generated with a bandwidth of 1.2, 2, or 4 kHz. Moreover, 1/3rd-octave noise with a center frequency of 200, 400, 700, 1100, and 1700 Hz was used additionally in the anechoic chamber in an attempt to recreate the test condition of [19] (Fig. 7). The test signals were processed by a BCC type I synthesizer with an ICLD of 9 dB. The synthesizer used an FFT size of 256 and effective window length of 256. The spectral resolution was equivalent to 1 ERB [1]. For each test signal a BCC-processed version with and without spatial equalization was generated. The two versions were played back alternately until the subject was sure how to judge the difference between the two versions.

In the anechoic chamber, the two subjects detected a smaller variation of the spatial location for the 1/3rd-octave noise signals at the different frequencies. This observation is in line with the simulations. However, the difference between equalized and non-equalized signals appeared small. The noise centered at 1700 Hz was excluded from this evaluation, since it was always localized outside the range between the speakers. This can be explained by the large ILD at this frequency (see Fig. 12). Similar results were obtained with the white noise and the clicks, i.e. the spatially equalized signals appeared to be slightly more focused on one spot as opposed to the non-equalized signals.

The results from the listeneing room presentations, however, did not show a clear preference for the spatially equalized signals. In one listening room one subject reported a more focused image with the spatial equalization. However, the other subject reported contradicting results. In the other listening room, an image widening was detected by the subjects for signals with spatial equalization. Most subjects also detected an image shift away from the center for the equalized signal in both listening rooms.

While the measurements from the anechoic chamber show trends that are in line with the simulations, the observations from the listening test in rooms do not agree well with the spherical head simulations, the HRTF data, and the subjective test results of [19]. This discrepancy can be related to the different acoustic environments. While the listening test was done in limited-size rooms, the simulation is based on free field conditions and the

measurements in [19] were done in an anechoic chamber. However, considering the remarks given in [18], spatial equalization should improve the spatial image especially in rooms.

In preliminary informal listening tests for the spatial equalization with headphones we used BCC type II with an FFT size of 1024. The time-window length was 896 samples which results in a frame length of 448 samples due to the 50% window overlap. The mapping of ICLDs to ICTDs according to (3) over the full frequency range shows a large improvement of lateralization. The auditory image without ICTDs appears narrow while with equalization it extends considerably to the sides.

5 SUMMARY

The psychoacoustics of source localization and knowledge about auditory scene analysis indicates that the perceived spatial image distortions of a BCC-synthesized auditory scene can be considerably smaller than predicted from the misalignment of "classical" binaural localization cues.

A systematic BCC design approach takes advantage of existing binaural perception models. A design example is the BCC reference scheme based on a Cochlear Filter Bank (CFB). The perceived quality of this BCC analysis/synthesis scheme using level-difference cues only is investigated for loudspeaker playback. The results show that the perceived degradation is mainly caused by a reduced auditory image width and stability. Other distortions are negligible. A low-complexity FFT-based BCC implementation is evaluated. The best performing FFT-based BCC synthesizer has an FFT size of 256 at 32 kHz sampling rate and shows equal performance to the reference, CFB-based BCC synthesizer. This implementation is suitable for low-cost real-time systems.

Two simple schemes for spatial parameter modification were introduced to improve the auditory spatial image quality. The first scheme is intended for two-channel loudspeaker playback. It amplifies level differences at low frequencies such that the interaural time differences are consistent with the cues created by a physical sound source. This spatial equalization is motivated by spherical head simulations that show a systematic deviation of time-delay cues at low frequencies. The simulation results are verified by an evaluation of an HRTF database and other published psychoacoustic results. Despite the promising agreement of the simulation with measurements, we were only able to confirm the image improvements in listening tests in an anechoic chamber. For loudspeaker presentation in a regular room, the equalization might not reduce the spatial image spread depending on the specific acoustic situation. Spatial equalization was not perceived to improve the spatial image in the two acoustically treated rooms used in our tests. Thus, we still need to better understand the influence of room acoustics on the spatial image perception and the consequences for equalization. Thus, we can not universally claim benefits from spatial equalization for loudspeaker playback in arbitrary rooms. Most likely, the spatial equalization needs to be adapted to the room acoustics in critical rooms.

The second scheme is intended to enhance the auditory image for headphone playback. A simple mapping of level differences to time differences is derived to create the important time-difference localization cues at low frequencies. Informal listening tests confirm that a considerable auditory image enhancement is achieved. Thus, the insertion of time-difference cues for headphone playback is important for achieving a sufficiently large lateralization.

Possible extensions of these studies include the enhancement of spatial equalization to the full audio bandwidth and to more than two channels. Moreover, as mentioned above, we need to understand better in which acoustical situations spatial equalization improves the image quality.

6 ACKNOWLEDGMENTS

We thank Jens Meyer for providing the spherical head model software and for many inspiring discussions. Peter Kroon contributed helpful comments on previous drafts. We appreciate the availability of the HRTF database [16].

REFERENCES

[1] C. Faller and F. Baumgarte, "Binaural Cue Coding applied to stereo and multi-channel audio compression," *112th AES Conv., Munich, preprint 5574*, May 2002.

[2] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust., New Paltz, NY*, Oct. 2001, pp. 199–202.

[3] F. Baumgarte and C. Faller, "Estimation of auditory spatial cues for Binaural Cue Coding," in *Proc. ICASSP 2002, Orlando, Florida*, May 2002.

[4] C. Faller and F. Baumgarte, "Binaural Cue Coding: A novel and efficient representation of spatial audio," in *Proc. ICASSP 2002, Orlando, Florida*, May 2002.

[5] C. Faller and F. Baumgarte, "Binaural Cue Coding applied to audio compression with flexible rendering," *113th AES Conv., Los Angeles*, Oct. 2002.

[6] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, 1994.

[7] F. Baumgarte, "Improved audio coding using a psychoacoustic model based on a cochlear filter bank," *IEEE Trans. Speech Audio Proc.*, (accepted).

[8] R. M. Stern and C. Trahiotis, *Binaural and spatial hearing in real and virtual environments*, chapter 24: Models of binaural perception, Lawrence Erlbaum Associates, New Jersey, 1997.

[9] J. Breebart, S. v. d. Par, and A. Kohlrausch, "Binaural processing model based on contralateral inhibition. I. Model structure," *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1074–1088, Aug. 2001.

[10] L. Lin, W. H. Holmes, and E. Ambikairajah, "Auditory filter bank inversion," *Proc. IEEE ISCAS 2001*, vol. II, pp. 537–540, May 2001.

[11] J. Blauert, *Spatial Hearing. The Psychophysics of Human Sound Localization*, MIT Press, 1983.

[12] F. Baumgarte and C. Faller, "Why Binaural Cue Coding is better than Intensity Stereo Coding," *112th AES Conv., Munich, preprint 5575*, May 2002.

[13] F. L. Wightman and D. J. Kistler, *Binaural and spatial hearing in real and virtual environments*, chapter 1: Factors Affecting the Relative Salience of Sound Localization Cues, pp. 1–23, Lawrence Erlbaum Associates, New Jersey, 1997.

[14] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Acoust. Soc. Am.*, vol. 111, no. 5(1), pp. 2219–2236, May 2002.

[15] C. Hugonnet and P. Walder, *Stereophonic Sound Recording: Theory and Practice*, Wiley and Sons, Chichester, 1995.

[16] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendo, "The CIPIC HRTF database," in *IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust., New Paltz, NY*, Oct. 2001, pp. 99–102.

[17] M. Gerzon, "Stereo shuffling: new approach, old technique," *Studio Sound*, Jul. 1986.

[18] D. Griesinger, "Spaciousness and localization in listening rooms and their effects on the recording technique.," *J. Aud. Eng. Soc.*, vol. 34, no. 4, pp. 225–268, 1986.

[19] V. Pulkki and M. Karjalainen, "Localization of amplitude-panned virtual sources I: stereophonic panning," *J. Aud. Eng. Soc.*, vol. 49, no. 9, pp. 739–752, Sep. 2001.