Binaural Cue Coding—Part II: Schemes and Applications

Christof Faller and Frank Baumgarte

***Abstract*—Binaural Cue Coding (BCC) is a method for multichannel spatial rendering based on one down-mixed audio channel and side information. The companion paper (Part I) covers the psychoacoustic fundamentals of this method and outlines principles for the design of BCC schemes. The BCC analysis and synthesis methods of Part I are motivated and presented in the framework of stereophonic audio coding.**

This paper, Part II, generalizes the basic BCC schemes presented in Part I. It includes BCC for multichannel signals and employs an enhanced set of perceptual spatial cues for BCC synthesis. A scheme for multichannel audio coding is presented. Moreover, a modified scheme is derived that allows flexible rendering of the spatial image at the receiver supporting dynamic control.

All aspects of complete BCC encoder and decoder implementations are discussed, such as down-mixing of the input signals, low complexity estimation of the spatial cues, and quantization and coding of the side information. Application examples are given and the performance of the coder implementations are evaluated and discussed based on subjective listening test results.

***Index Terms*—Audio coding, auralization, binaural signal, HRTF, multichannel audio, spatial image, spatial rendering, stereo audio, surround sound.**

I. Introduction

BINAURAL Cue Coding (BCC) is a technique for low-bit-rate coding of a multitude of audio signals or audio channels. Specifically, it addresses the following two application scenarios.

- *Transmission of a number of separate source signals for the purpose of rendering at the receiver*: When a number of source signals are transmitted separately, a receiver has the flexibility of rendering these source signals spatially as desired. For example, if music is transmitted in the form of separate tracks, a user may create his own favorite mix. Also, the mixing at the receiver can be carried out to match the given playback setup (e.g., stereo[1] or multichannel audio system).
- *Transmission of a number of audio channels of a stereo or multichannel signal*: A number of audio channels are transmitted for a specific playback setup, i.e., stereo or multichannel audio system.

Manuscript received May 25, 2002; revised August 6, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerald Schuller.

The authors are with the Media Signal Processing Research Department, Agere Systems, Allentown, PA 18109 USA (e-mail: cfaller@agere.com; fb@agere.com).

Digital Object Identifier 10.1109/TSA.2003.818108

[1]The term "stereo" always refers to two-channel stereophonics only.

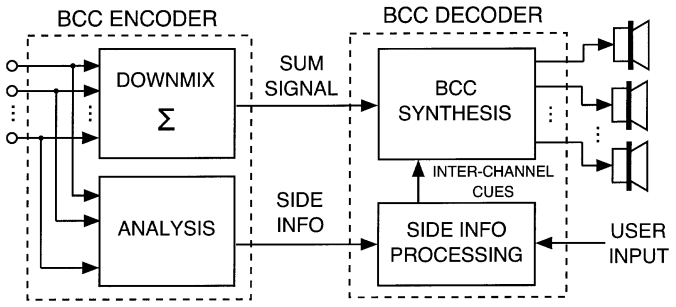

Fig. 1. Generic BCC scheme. The audio input signals are analyzed and down-mixed. The sum signal plus side information is transmitted to the decoder. The inter-channel cues are generated from the side information and local user input. BCC synthesis generates the multichannel audio output signal.

With conventional coding techniques, both of these scenarios are rather expensive in terms of bit rate, compared to the transmission of a single audio channel. The bit rate of BCC schemes is only slightly higher than the bit rate required for the transmission of one audio channel. A generic BCC scheme is shown in Fig. 1. BCC schemes jointly transmit a number of audio signals as one single channel, denoted sum signal, plus low-bit-rate side information, enabling low-bit-rate transmission of such signals. BCC is a lossy technique and can not recover the original signals. It aims at recovering the signals perceptually. The BCC schemes addressing the two previously described scenarios are denoted BCC for *Flexible Rendering* (*type I* in [1]) and BCC for *Natural Rendering* (*type II* in [1]), respectively.

Multichannel panning techniques [2] and wavefield synthesis techniques [3] generate a number of audio channels given a number of audio source signals. The major difference between these techniques and BCC is, that BCC operates in subbands and is able to spatialize a number of source signals given only the respective sum signal (with the aid of side information). *Intensity Stereo* (IS) [4] has a similar functionality, but is only useful at frequencies above about 1–2 KHz and has some other disadvantages [5]. For the specific case of a number of signals from a microphone array, different approaches for reducing the number of audio channels were studied in [6]. A technique for efficient joint coding of multiple concurrent sound sources was presented in [7]. As opposed to considering binaural perception, this approach considers perceptual effects such as the continuity illusion.

The companion paper, Part I [8], motivates BCC from a psychoacoustic point of view. Also, a number of filter banks are assessed with respect to their performance in BCC schemes and the quality of the spatial image of BCC synthesized signals is discussed based on a series of subjective test results.

In this paper, Part II, we present in detail complete BCC schemes, including quantization and coding of the side information. The presented schemes are based on a low complexity DFT-based filter bank with support of an arbitrary number of audio channels. Multichannel signal analysis and synthesis and the generation of binaural signals with *Head Related Transfer Functions* (HRTFs) [9] is described in detail. The previously presented BCC schemes [10]–[12] are put into a common framework. Novel additions for enhanced quality and functionality are the consideration of cross-correlation cues and a frequency-domain algorithm for improved down-mixing of stereo and multichannel signals.

The paper is organized as follows. Section II motivates BCC synthesis and describes its implementation in detail. The two types of BCC schemes are presented in Section III. Section IV describes how to code the BCC side information, to significantly reduce its data rate. Computational complexity and audio delay issues are discussed in Section V. Section VI explains how BCC can be combined with conventional audio or speech coders and proposes various applications for the combined schemes. Section VII describes the results of various subjective tests that were conducted to assess the usefulness and quality of the proposed implementations. Conclusions are drawn in Section VIII.

II. BCC Synthesis

The main purpose of the BCC synthesis technique is to generate stereo or multichannel signals with certain inter-channel cues in subbands between pairs of channels such that a desired binaural perception results. It is assumed that such inter-channel cues are the determining factor for the perception of a spatial image. The bandwidth of each of these subbands corresponds conceptually to the "critical bandwidth" for binaural perception. For the purpose of synthesizing inter-channel cues, we found that a subband width equal to twice the *Equivalent Rectangular Bandwidth* (ERB) [13] provides a sufficiently high frequency resolution. Smaller bandwidths can have disadvantages in terms of time resolution and side information bit rate.

Commonly used definitions of stationary inter-channel cues for corresponding subband signals $x_1(k)$ and $x_2(k)$ of two audio channels with time index k are:

- *Inter-Channel Level Difference* (ICLD) [dB]:

$$\Delta\bar{L} = \lim_{\ell\to\infty} 10\log_{10}\left(\frac{\sum_{k=-\ell}^{\ell} x_2^2(k)}{\sum_{k=-\ell}^{\ell} x_1^2(k)}\right). \tag{1}$$

- *Inter-Channel Time Difference* (ICTD) [samples]:

$$\bar{\tau} = \arg\max_d\{\bar{\Phi}_{12}(d)\} \tag{2}$$

with the normalized cross-correlation defined as

$$\bar{\Phi}_{12}(d) = \lim_{\ell\to\infty} \frac{\sum_{k=-\ell}^{\ell} x_1(k)x_2(k+d)}{\sqrt{\sum_{k=-\ell}^{\ell} x_1^2(k)\sum_{k=-\ell}^{\ell} x_2^2(k)}}. \tag{3}$$

- *Inter-Channel Correlation* (ICC):

$$\bar{\Gamma} = \max_d|\bar{\Phi}_{12}(d)|. \tag{4}$$

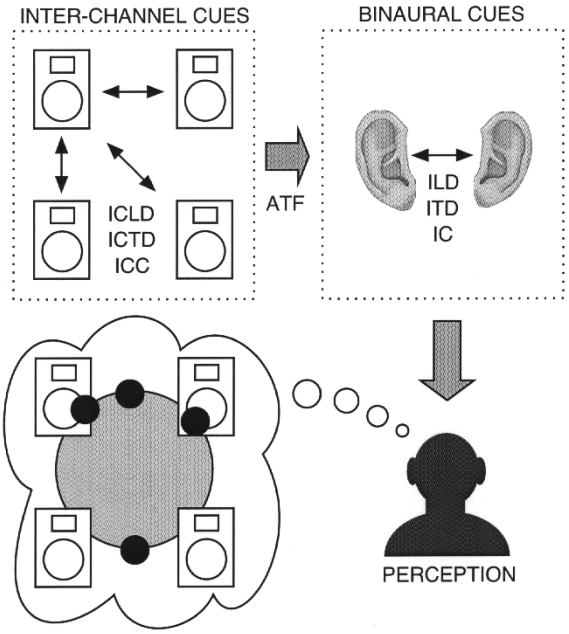

Fig. 2. Chain from inter-channel cues to binaural cues to perception. The inter-channel cues between pairs of loudspeaker signals (ICLDs, ICTDs, ICCs) determine together with the ATF the binaural cues (ILDs, ITDs, ICs), which determine the perception of the spatial image.

The measure for the ICC (4) is denoted degree of correlation or just correlation in the remaining part of the paper. Since usually inter-channel cues are nonstationary, we use short-time estimates of (1)–(4) as is described in detail later.

When the two audio channels contain a signal with a 180° phase difference, the degree of correlation is 1. Since we are only considering the magnitude of the normalized cross-correlation (4), this case is treated the same as if the signals were in phase. Thus, in this case BCC synthesis generates two audio channels with identical signals, which are in general perceptually different than the original signals with a 180° phase difference. This is a current limitation.

Additional signal cues may be added for controlling other aspects such as reverberance and distance more explicitly, e.g., parameters of a reverberation model.

Fig. 2 illustrates the chain from inter-channel cues to perception.

- The inter-channel cues and the *Acoustic Transfer Function* (ATF) from each speaker to the eardrum are the determining factors for the binaural cues [*Interaural Level Difference* (ILD), *Interaural Time Difference* (ITD), and *Interaural Correlation* (IC)].
- These binaural cues are the determining factors for the perception of the spatial image [9], [14].
  - ILD and ITD cues are determining factors for the lateralization of auditory objects. According to the duplex theory [15], ILD are more important than ITD at frequencies above about 1–1.5 kHz. At lower frequencies ITD are more important.
  - IC cues determine the width or diffuseness of auditory objects.

In the general case of C playback channels ICLDs and ICTDs are considered in each subband between pairs of channels, i.e.,

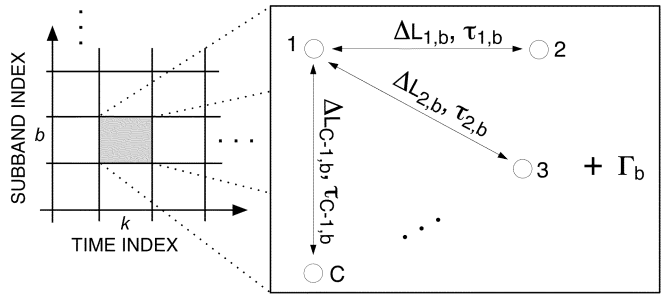

Fig. 3. For each subband b at each time index k the ICLD and ICTD are considered between pairs of channels as shown. The ICC is considered once for all channels in each subband b.

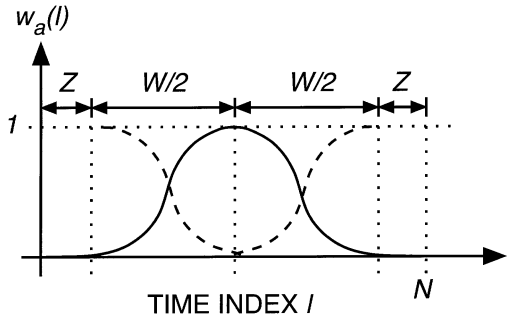

Fig. 4. Analysis window. The time-span of the window W is shorter than the DFT length N such that noncircular time-shifts within the range $[-Z, Z]$ are possible. The window is advanced by $W/2$ samples.

for each channel relative to a reference channel. Without loss of generality, channel number 1 is defined as the reference channel. With the ICLD and ICTD cues it is possible to render a source to any direction between one of the loudspeaker pairs of the playback setup that is used. For determining the width (or diffuseness) of a rendered source it is enough to consider one parameter per subband for all audio channels, i.e., one ICC. The width of the rendered source is controlled by modifying the subband signals such that all possible channel pairs have the same ICC.

Fig. 3 illustrates how the inter-channel cues are considered in each subband b at each time index k. For example, $\{\Delta L_{ib},\ \tau_{ib}\}$ are the ICLD and ICTD between channel 1 and channel $i+1$ for the bth subband. The ICC Γ_b determines the correlation of all channel pairs in subband b. Note that for simplicity of notation we ignore the time index k.

Sections II-A and B give the implementation details of BCC synthesis.

A. Time-Frequency Transform

BCC synthesis needs to be able to modify the level of audio signals and introduce delays adaptively in frequency and time to generate ICLDs and ICTDs between pairs of channels. Therefore, the aim is to use a spectral representation that supports level modifications, positive and negative time shifts, or more generally speaking, filtering in a time-varying fashion of the underlying audio signal.

It is shown in the following how a DFT can meet these requirements. As time-frequency transform, a DFT is applied frame-wise. The frame number that enumerates the applied DFT transforms in time corresponds to the time index k. When W sum-signal samples $s(0), \ldots, s(W-1)$ are transformed by a DFT into a complex spectral representation S_n $(0 \le n < W)$, a circular time shift of d time-domain samples is obtained by modifying the phase according to $\hat{S}_n = S_n \exp(-j(2\pi nd/W))$. However, artifacts occur due to the circular time shift within the frame. To achieve a noncircular time shift, the samples $s(0), \ldots, s(W-1)$ are padded with Z zeros at the beginning and the end of each frame and a DFT of size $N = 2Z + W$ is used. The sum signal is processed frame-wise with such a zero-padded DFT with W samples time advance (hop size) between the subsequent DFTs such that perfect reconstruction with an inverse DFT is achieved if the spectrum is not modified. A time shift within the range $d \in [-Z, Z]$ is implemented by modifying the N spectral coefficients according to

$$\hat{S}_n = S_n \exp\left(-j\frac{2\pi nd}{N}\right). \tag{5}$$

More generally speaking, the underlying signal can be filtered by multiplication of its spectrum S_n with the frequency response H_n of a filter

$$\hat{S}_n = S_n H_n. \tag{6}$$

As long as the filter's impulse response satisfies

$$h(l) = 0 \text{ for } |l| > Z \tag{7}$$

no artifacts occur due to circular convolution. Obviously, not only time shifts can be implemented by filtering but also level modifications. The described procedure is similar to the overlap-add convolution method using the DFT [16].

The described scheme works perfectly as long as the spectral modification is not varied over time k. When d varies over time, the transitions have to be smoothed. Therefore, overlapping windows are used for the forward transform. A frame of N samples is multiplied with a window before an N-point DFT is applied. We use a Hann window with zero padding at both sides

$$w_a(l) = \begin{cases} 0, & \text{for } 0 \le l < Z \text{ or} \\ & N - Z \le l < N \\ \sin^2\left(\frac{(l-Z)\pi}{W}\right), & \text{for } Z \le l < Z + W \end{cases} \tag{8}$$

where Z is the width of the zero region before and after the nonzero part of the window. Fig. 4 shows the described window schematically. The nonzero window span is W and the size of the transform is $N = 2Z + W$. Adjacent windows are overlapping and are shifted by $W/2$ samples (hop size). The window was chosen such that the overlapping windows add up to a constant value of 1. Therefore, for the inverse transform there is no need for additional windowing. A plain inverse DFT of size N with time advance of successive frames of $W/2$ samples is used. If the spectrum is not modified, perfect reconstruction is achieved by overlap add. A time-invariant (integer) delay d results in perfect reconstruction with a delay. Audible distortions resulting from spectral modifications (level and delay) are avoided by changing the spectral modification only smoothly over frequency and time.

TABLE I
PARTITION BOUNDARIES A_b $(0 \leq b \leq B = 20)$ FOR THE CASE OF PARTITION BANDWIDTHS OF 2 ERB, $N = 1024$, AND A SAMPLING RATE OF $f_s = 32$ kHz

A_0	A_1	A_2	A_3	A_4	A_5	A_6
0	2	4	7	11	15	20
A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
26	34	44	56	71	90	113
A_{14}	A_{15}	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}
142	178	222	277	345	430	513

Commonly used time-frequency transforms, such as the MDCT [17], use windowing for analysis and synthesis, e.g., a sine window. The condition for perfect reconstruction is, that the overlapping analysis windows multiplied by the synthesis windows add up to a constant value of 1. When modifying the spectrum of the sum signal in the decoder, such as to impose a time shift, not only the underlying signal is shifted but also the imposed analysis window. Therefore, the synthesis window does not match anymore the analysis window resulting in distortions in the reconstructed signal. We are avoiding these distortions by not using a synthesis window.

The uniform spectral resolution of the DFT is not well adapted to human perception. Therefore, the uniformly spaced spectral coefficients S_n $(0 \leq n \leq N/2)$ are grouped into B nonoverlapping partitions with bandwidths better adapted to perception. Only the first $N/2+1$ spectral coefficients of the spectrum are considered because the spectrum is symmetric. The indices of the DFT coefficients which belong to the partition with index b $(1 \leq b \leq B)$ are $n \in \{A_{b-1}, A_{b-1}+1, \ldots, A_b - 1\}$ with $A_0 = 0$. The signals represented by the spectral coefficients of the partitions correspond to the perceptually motivated subband decomposition used by BCC. Thus, within each such partition only one set of inter-channel cues (ICLD, ICTD, ICC) is synthesized for each channel pair.

For our experiments we used $W = 896$, $Z = 64$, and $N = 1024$ for a sampling rate of $f_s = 32$ kHz. We used $B = 20$ partitions, each having a bandwidth of approximately 2 ERB. The resulting partition boundaries A_b are shown in Table I.

With the chosen Z, ICTDs of up to 4 ms ($2Z = 128$ samples) can be synthesized. This is well above the range of ICTDs we are using (Section IV-B). For BCC synthesis with HRTFs according to (17), the chosen Z is also large enough. A motivation for choosing Z larger than necessary is to increase the time-resolution of the spectral representation, i.e., use a shorter window and window hop size.

B. Spectral Modifications for Each Output Channel

Fig. 5 illustrates the processing of BCC synthesis. The given sum signal is converted to a spectral representation and as a function of the given ICLDs, ICTDs, and ICCs the spectral coefficients are modified for generating the spectra of the multiple output channels. These spectra are converted back to the time-domain resulting in the multichannel output. An FFT is used as time-frequency transform (TF).

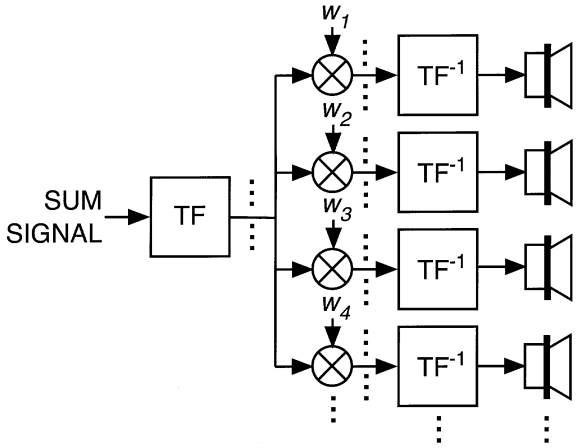

Fig. 5. The sum signal is converted to a spectral representation (time-frequency transform, TF). Then as a function of the ICLDs, ICTDs, and ICCs the spectral coefficients are modified by scaling. The processing for one spectral coefficient is shown. These modified spectra are converted back to the time-domain with the inverse transform (TF^{-1}).

Given the spectral coefficients S_n of the mono sum signal, the spectral coefficients $S_{c,n}$ for each channel c are obtained by

$$S_{c,n} = F_{c,n} G_{c,n} S_n \tag{9}$$

where $F_{c,n}$ is a positive real number determining a level modification for each spectral coefficient. $G_{c,n}$ is a complex number of magnitude one determining a phase modification for each spectral coefficient. The following three paragraphs describe how $F_{c,n}$ and $G_{c,n}$ are obtained given $\{\Delta L_{c,b},\ \tau_{c,b}, \Gamma_b\}$.

1) Determining the Level Modification for Each Channel: The factors $F_{c,n}$ for channel $c > 1$ are computed for each spectral coefficient within a partition b (indices: $A_{b-1} \leq n < A_b$) given $\Delta L_{c-1,b}$

$$F_{c,n} = 10^{(\Delta L_{c-1,b} + r_{c-1,n})/20} F_{1,n} \tag{10}$$

where $r_{c-1,n}$ are random numbers for controlling the degree of correlation between the channel pairs. Section II-B-III describes how $r_{c-1,n}$ is computed as a function of Γ_b. The factors $F_{c,n}$ for the reference channel $(c = 1)$ are computed such that for each partition the sum of the power of all channels is the same as the power of the sum signal

$$F_{1,n} = \frac{1}{\sqrt{1 + \sum_{i=1}^{C-1} 10^{(\Delta L_{i,b} + r_{i,n})/10}}} \tag{11}$$

resulting in a loudness which is approximately independent of the level differences $\Delta L_{i,b}$. This normalization (11) is usually used for amplitude panning [2]. Before applying (9), $F_{c,n}$ is smoothed at the partition boundaries by interpolating between the different $\Delta L_{i,b}$ to reduce aliasing artifacts. We use a simple smoothing rule which uses one averaged scale factor $F_{c,n}$ at the partition boundaries.

2) Determining the Phase Modification for Each Channel: The complex factors $G_{c,n}$ for channel $c > 1$ are computed for each spectral coefficient within a partition b (indices: $A_{b-1} \leq n < A_b$) given $\tau_{c-1,b}$ in sampling intervals

$$G_{c,n} = \exp\left(-j \frac{2\pi n (\tau_{c-1,b} - \tau_b)}{N}\right) \tag{12}$$

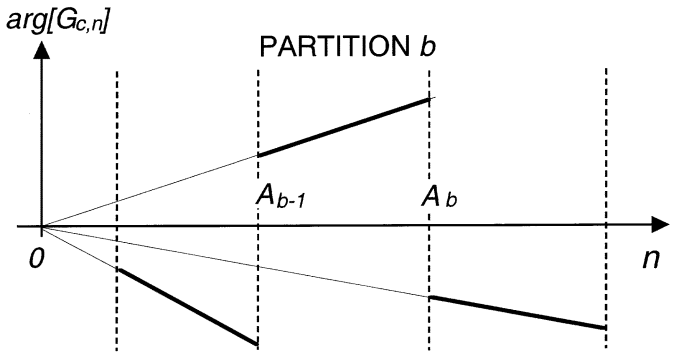

Fig. 6. Phase of $G_{c,n}$ for the synthesis of ICTDs as delays.

where τ_b is the delay which is introduced into reference channel 1

$$\tau_b = \frac{(\max_{1\le i<C} \tau_{ib} + \min_{1\le i<C} \tau_{ib})}{2}$$
$$G_{1,n} = \exp\left(-j\frac{2\pi n \tau_b}{N}\right). \quad (13)$$

The delay for the reference channel τ_b is computed in (13) such that the maximum absolute delay introduced to any channel in a specific partition is minimal. By minimizing the maximum phase modification, the maximum audio signal distortion resulting from (12) is minimized. Fig. 6 shows an example of $\arg\{G_{c,n}\}$ for synthesizing three different delays in 3 partitions of one audio channel c.

3) Determining the Correlation Modification for Each Channel: A perceptually meaningful way to reduce the correlation between a channel pair is to modify the ICLDs within partitions. The modification is done by adding a random sequence to the ICLDs along the frequency axis. The random sequence $\bar{r}_{c,n}$ (c = channel pair index, n = frequency index) for each channel pair is chosen such that the variance is approximately constant for all partitions in all channels and the average is zero in each partition. Moreover, clustering of large or small values in the sequence should be avoided. For each channel pair a different independent random sequence is used. The random sequences are not varied in time, i.e., the same sequences are applied to each frame. The correlation is controlled by modifying the variance of the random sequence. The variance modification is done in each partition b (indices: $A_{b-1} \le n < A_b$) as a function of Γ_b

$$r_{c,n} = (1 - \Gamma_b)\bar{r}_{c,n}. \quad (14)$$

A suitable amplitude distribution for the random sequence is a uniform distribution. The range of the distribution of $\bar{r}_{c,n}$ determines the strength of the correlation reduction. We use a range of ± 5 dB.

III. Binaural Cue Coding Schemes

As described in the introduction, the generic BCC scheme shown in Fig. 1 has different functionality, depending on the type of input signals. In this section we describe the functionality of both schemes, BCC for Flexible Rendering and BCC for Natural Rendering, in detail. Both schemes use different kinds of signal analysis, side information, and summation operations.

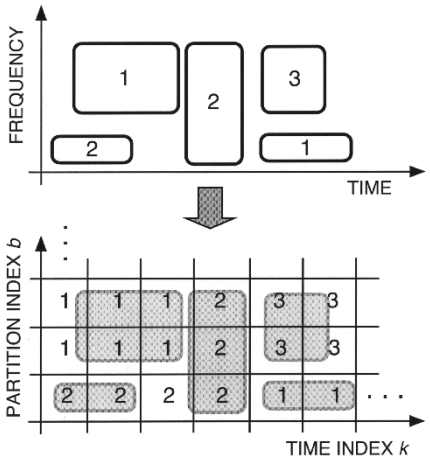

Fig. 7. Different source signals dominate in different regions of the time-frequency plane of the sum signal (top). For each partition b at each time k the source index of the strongest source (bottom) is transmitted to the decoder.

A. BCC for Flexible Rendering

1) Encoder Processing: BCC for Flexible Rendering generates the sum signal by simple addition of all M input signals. In different regions of the time-frequency representation of this sum signal, different sources dominate as illustrated in the top of Fig. 7. For each partition b at each time k the source index of the strongest source is transmitted to the decoder as shown in the bottom of Fig. 7. Given these source indices, the decoder has enough information for synthesizing the inter-channel cues of the strongest source in each partition at each time. It is assumed that the inter-channel cues of the strongest sources are the most important cues for the perception of the spatial image.

This is implemented by assigning to each partition b a source index I_b ($1 \le I_b \le M$)

$$I_b = \arg \max_{1\le m\le M} P_{m,b} \quad (15)$$

with partition power estimates $P_{m,b}$

$$P_{m,b} = \sum_{n=A_{b-1}}^{A_b-1} |S_{m,n}|^2 \quad (16)$$

where $S_{m,n}$ are the spectral coefficients of audio source m and M is the number of sources. The same time-frequency transform is used as for BCC synthesis (Section II-A).

2) Decoder Processing: For each partition b the inter-channel cues are obtained from a local table which stores one set of inter-channel cues for each source m ($1 \le m \le M$). (This corresponds to the "User Input" in Fig. 1). For each partition the inter-channel cues are chosen according to the transmitted source index I_b. Then the multichannel output signal is generated by applying BCC synthesis. The ICLD and ICTD stored in the table for each source determine the direction, whereas the ICC determines the diffuseness or width of the source. Time adaptive flexible rendering is implemented by (smoothly) modifying the inter-channel cues in the table in real-time. The decoder is simply adapted to the number of desired playback channels C by employing a table with $C - 1$ sets of inter-channel cues for each source.

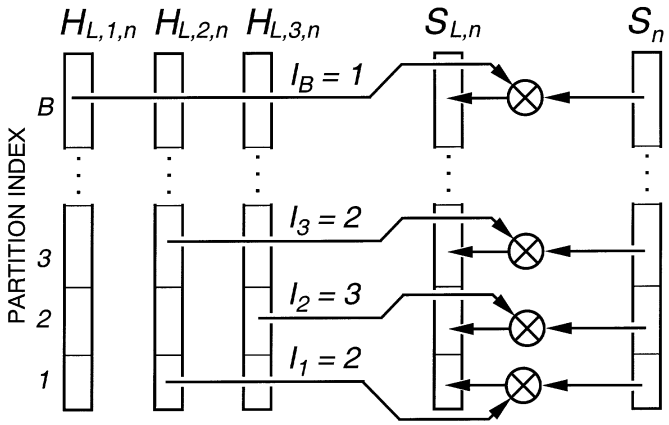

Fig. 8. Process of generating the left channel of a binaural signal with HRTFs. As a function of the source index I_b portions of different HRTFs are applied in the different partitions.

As an alternative to synthesizing ICLDs and ICTDs, binaural signals can be generated with HRTFs. In this case, the local table in the BCC synthesizer stores for each source m a left and right HRTF frequency response, $H_{L,m,n}$ and $H_{R,m,n}$. For obtaining the left and right binaural signals, each partition b is multiplied with the coefficients corresponding to the left and right HRTF associated with source I_b

$$S_{L,n} = H_{L,I_b,n} S_n \text{ and } S_{R,n} = H_{R,I_b,n} S_n \tag{17}$$

where $n \in \{A_{b-1}, A_{b-1}+1, \ldots, A_b - 1\}$ are the indices of spectral coefficients within partition b. It must be made sure that the impulse response of each HRTF satisfies the condition of (7). An example of this process of "mixing" HRTFs is shown for the left channel in Fig. 8. To prevent aliasing artifacts, the transitions between partitions need to be smoothed. This is done by using overlapping spectral windows for the different portions of HRTFs between the partitions.

B. BCC for Natural Rendering

1) Encoder Processing: For this scenario, the encoder estimates the inter-channel cues in each partition b at each time index k. These estimated signal cues are transmitted as side information to the decoder. In the following, the estimation of the inter-channel cues and the generation of the sum signal is described in detail.

An estimation algorithm for ICLD, ICTD, and ICC based on a cochlear filter bank is presented in Part I [8]. In the following it is described how the cue estimation can be implemented with lower computational complexity by using the time-frequency transform described in Section II-A.

ICLD estimation: First, for each of the audio input channels $1 \le m \le C$ ($M = C$) the power within each partition $1 \le b \le B$, $P_{m,b}$, is estimated similarly to (16). The estimated ICLD in dB between channel c and the reference channel 1 for partition b is

$$\Delta L_{c-1,b} = 10 \log_{10} \left(\frac{P_{c,b}}{P_{1,b}} \right). \tag{18}$$

ICTD estimation: At frequencies below about 1.5 kHz the phase delay between a pair of channels is relevant for spatial perception [9] and the ICTDs are estimated by averaging the phase delay within each partition between a pair of channels. At frequencies above about 1.5 kHz the group delay ("envelope delay") is relevant for spatial perception [9]. To estimate the group delay, the phase difference between a pair of channels is computed. Then for each partition b linear regression is used to compute the slope of the phase difference of the spectral coefficients within the partition (indices: $A_{b-1} \le n < A_b$). The group delay between a pair of channels is proportional to the slope of the regression line.

ICC estimation: At each time for each partition the two strongest input channels are selected for the correlation estimation. The motivation for selecting the two strongest channels is that in many cases sources are amplitude panned between two adjacent loudspeakers and the diffuseness of the source is then also determined by the correlation between these two channels.

If the spectra of the selected two input channels are denoted $\tilde{S}_{i,n}$ and $\tilde{S}_{j,n}$ (in this paragraph the letters i and j are used independently of the other parts of the paper). The magnitude-squared coherence between these two channels is

$$|\Gamma(n,k)|^2 = \frac{\Phi_{ij}(n,k)\Phi_{ji}^*(n,k)}{\Phi_{ii}(n,k)\Phi_{jj}(n,k)} \tag{19}$$

with

$$\Phi_{ij}(n,k) = E\{\tilde{S}_{i,n}(k)\tilde{S}_{j,n}^*(k)\} \tag{20}$$

where $E\{\ldots\}$ denotes mathematical expectation and * is the complex conjugate of a complex number. A short-time estimate of $\Gamma(n,k)$ is obtained by computing $\Phi_{ij}(n,k)$ according to

$$\Phi_{ij}(n,k) = \alpha \tilde{S}_{i,n}(k)\tilde{S}_{j,n}^*(k) + (1-\alpha)\Phi_{ij}(n,k-1). \tag{21}$$

The factor α determines the degree of smoothing of the estimation over time. For the specific parameter settings used here we set $\alpha = 0.1$.

For obtaining a measure for the degree of correlation, the coherence estimates are averaged in each partition. For the averaging it is meaningful to apply a weighting function to the coherence before averaging. The weighting can be made proportional to the product of power estimates, $\Phi_{ii}(n,k)\Phi_{jj}(n,k)$, which eliminates the denominator in (19). Since we are interested in the average degree of correlation in each partition, we average the weighted magnitude coherence in each partition and normalize it by the sum of power estimate products. Thus, the final degree of correlation is

$$|\Gamma_b(k)|^2 = \frac{\sum_{n=A_{b-1}}^{A_b-1} |\Gamma(n,k)|^2 \Phi_{ii}(n,k)\Phi_{jj}(n,k)}{\sum_{n=A_{b-1}}^{A_b-1} \Phi_{ii}(n,k)\Phi_{jj}(n,k)}. \tag{22}$$

Summation of the input signals: When independent audio input signals are given, the the sum signal is generated by simply adding up the signals, as we do in the case of BCC for Flexible Rendering. However, in the case of BCC for Natural Rendering a number of audio channels are given which are often not independent. Many stereo signals are "mono compatible" and simple addition of the audio channels is feasible. However, some stereo

signals and many multichannel signals are not "mono compatible," i.e., ordinary addition results in cancellation and/or undesired amplification of certain signal components. Therefore, we use a simple scheme to reduce the negative effects of the summation. Firstly, the sum signal is computed in the spectral domain. Secondly, for each partition b the power of the sum signal, P_b, and the power of each input signal, $P_{c,b}$ $(1 \leq c \leq C)$, is computed similarly to (16). Then the spectral coefficients of each partition b (indices: $A_{b-1} \leq n < A_b$) of the sum signal are multiplied with a gain factor

$$g_b = \sqrt{\frac{\sum_{c=1}^{C} P_{c,b}}{P_b}} \tag{23}$$

such that the power of the sum signal is equal to the sum of all input signal powers in each partition. Cancellation and amplification effects are reduced by this processing. Additionally, scaling with (23) has the positive side effect that the output signal has in each partition of each channel the same power as the original input signal (assuming that the ICLD estimates are accurate). This follows from the normalization used by the ICLD synthesis [(9)–(11)].

The presented algorithm is simple and gives good results. Further improvements may be possible by applying more sophisticated algorithms incorporating phase modifications prior to adding the spectral coefficients.

2) Decoder Processing: The estimated inter-channel cues (BCC side information) are directly used to generate the output multichannel audio signal by applying BCC synthesis.

IV. Coding of BCC Side Information

For coding applications it is desirable that the BCC side information requires only a low bit rate. In the following, we are describing the quantization and coding techniques that are applied to achieve this.

A. BCC for Flexible Rendering

A run-length coding algorithm [18] is applied to the source indices I_b over frequency (independently for each frame, a more sophisticated algorithm may consider also dependencies between frames). For the chosen parameters (Section II-A), this results in a bit rate of about 2 kb/s for the case of having three simultaneously active talkers as source signals. If not all talkers are active simultaneously, the run-length coding is more efficient and results in lower bit rates.

B. BCC for Natural Rendering

The estimated ICLDs and ICTDs ($\Delta L_{i,b}$, $\tau_{i,b}$) are limited to a range of $\pm\Delta L_{\max}$ and $\pm\tau_{\max}$ with $\Delta L_{\max} = 18$ dB and $\tau_{\max} = 800$ μs. For binaural perception, these limits correspond to a phantom image at the far right or far left. In case of loudspeaker playback, the ICLD limit corresponds to a phantom image at the far side between a pair of speakers. The ICC is already limited to $0 \leq \Gamma_b \leq 1$. After limiting, uniform quantizers are used for quantizing the estimated inter-channel cues. For the experiments reported here, we used 7 quantizer values for ICLDs and ICTDs and 4 quantizer values for ICCs. More specifics about the quantization of ICLDs and ICTDs are given in [12].

For a lower bit rate, the quantizer indices are coded as follows. For each frame, two sets of quantizer index differences are computed. Firstly, for each partition b the difference between the current (time index k) and the previous $(k-1)$ quantizer index is computed. Secondly, the difference of the quantizer index of each partition b compared to the index of the partition with the next higher index, $b+1$, is computed. Both of these sets of index differences are coded with a Huffman code. We use one single Huffman codebook for all cases. For transmission, the set is chosen which uses less bits. One additional bit is transmitted indicating which set is transmitted.

The resulting average bit rate for ICLDs or ICTDs for one channel pair for the entropy coded quantizer indices is approximately 2 kb/s. The average bit rate for the ICC is about 1.5 kb/s.

The ICTDs with the range used here $(\pm 800\ \mu\text{s})$ are perceptually not relevant for loudspeaker playback in reverberant environments [8]. Therefore, for loudspeaker playback one may decide to only transmit ICLDs resulting in a lower bit rate for the BCC side information. However, for headphone playback ICTDs are important at frequencies below about 1–1.5 kHz [8]. One may still decide to only transmit ICLDs. But for improved headphone performance, the BCC for Natural Rendering decoder can alternatively generate ICTDs between pairs of channels that are proportional to the corresponding ICLDs

$$\tau_{i,b} = -q f_s \Delta L_{i,b} \tag{24}$$

with a scaling factor of q. Positive values of q result in ICTDs shifting the phantom image in the same direction as the given ICLDs. Informal listening revealed that for headphone playback $q = 25 \cdot 10^{-6}$ seconds/dB delivers improved quality over the case of only synthesizing ICLDs $(q = 0)$.

V. Complexity and Delay

The complexity of the presented BCC implementation is reasonably low. The most demanding operations are the FFT and IFFT. these are of size 1024 for the specific parameters chosen (Section II-A). Implementations of BCC encoder and decoder are both running in real-time on a laptop computer (500 MHz PowerPC G4 processor) with each having about 5% processor load for stereo audio input and output. Also, a fixed-point version of the BCC for Natural Rendering decoder was implemented, running in real-time together with a PAC [19] decoder on a general purpose DSP.

The algorithmic delay is defined as the delay for encoding and decoding of a signal excluding the time needed for calculations and data transmission. When the sum signal is computed by simple addition of the input signals in the time domain, the sum signal can be fed to the BCC decoder without any delay. For this case, the windowed FFTs of encoder and decoder operate synchronously and no delay compensation for the side information needs to be considered. Thus, the algorithmic delay is determined by just the decoder window size $W = 896$ which corresponds to 28 ms.

If the sum signal is computed in the frequency domain (as explained in the last paragraph of Section III-B-I), then the ad-

ditional delay of the transmitted sum signal is $W/2$ samples. The resulting total algorithmic delay is $3W/2 = 1344$ samples or 42 ms.

It should be noted that the value of W was chosen to be a compromise between quality, delay, and bit rate. Choosing smaller values of W will reduce the delay and will improve the perceived quality (as was shown in Part I [8]). However, smaller values of W will increase the bit rate, assuming the coding scheme described in Section IV.

VI. Applications of BCC

A. Stereo and Multichannel Audio Coding

BCC for Natural Rendering together with a conventional mono audio coder can be used for coding of stereo and multichannel audio signals. Multichannel surround can be provided with BCC at bit rates at which previously only mono or stereo could be delivered. Not only the low bit rate but also the backward compatibility aspect is of interest. For example, existing mono audio radio broadcasting systems can be enhanced for stereo or multichannel playback if the BCC side information can be embedded into the existing transmission channel. BCC can also be combined with analog broadcasting by transmitting the BCC side information in the same frequency band as the analog transmission. The digital data for the BCC side information could be transmitted with similar techniques as are used for *In-Band-On-Channel* (IBOC) hybrid broadcasting [20], [21]. Due to the low bit rate of the BCC side information, potentially simpler techniques can be used, such as embedding data directly into the audio signal.

B. Tele-Conferencing

Fig. 9 shows a scheme for a stereo tele-conferencing client incorporating BCC for Flexible Rendering. It consists of (A) speech decoder, (B) BCC decoder, (C) speech encoder, and (D) stereo echo canceler [22], [23]. In the server, each client is connected to the other clients with a scheme as shown in Fig. 10 (A) and (C) are speech decoders and encoders, respectively. The BCC encoder is the same as the left part in Fig. 1.

There are several reasons why BCC is very interesting for applications in tele-conferencing.

- BCC for Flexible Rendering has rendering capabilities at the decoder: Each client (decoder) can not only determine the spatialization of each of the other conference participants but also the number of playback channels. The same side information supports different types of client systems (e.g., mono, stereo, or multichannel).
- Low bit rate: The bit rate is as low as in a mono system with a small overhead for the BCC side information.
- Backward compatibility: If the BCC side information can be embedded into the transmission channel between the server and the clients of a mono system, this system can be upgraded for stereo or multichannel tele-conferencing while maintaining backward compatibility. For example, one could use LSB "bit-flipping" in μLaw [24] to embed the BCC side information.

We tried two speech coders, G.722 [25] and G.729 [26], which operate at 24 kb/s and 8 kB/s, respectively. Since these

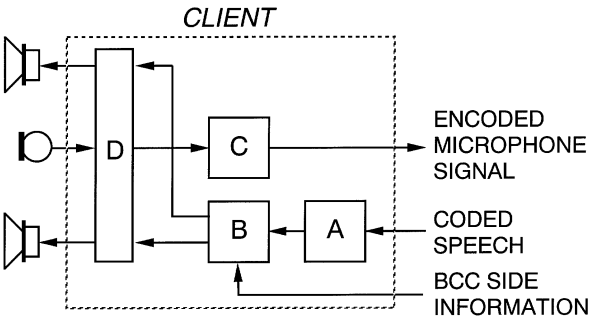

Fig. 9. Scheme for a stereo tele-conferencing client with one microphone based on BCC: (A) speech decoder, (B) BCC decoder, (C) speech encoder, and (D) stereo echo canceler.

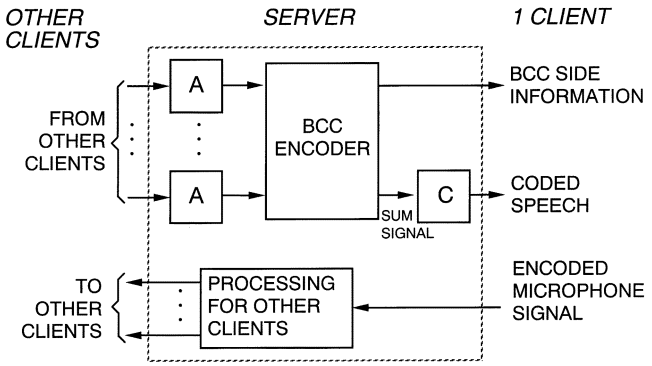

Fig. 10. Each client is connected to the other clients with a scheme as shown: (A) speech decoder and (C) speech encoder.

speech coders model a single human vocal tract, it is not obvious that they perform well for multiple simultaneously active speech signals. We found that the speech quality degrades gradually as the number of simultaneous speech signals increases. When rendered to different directions with BCC, speech can be well understood with as many as five simultaneous talkers. In contrast, without spatialization the speech intelligibility is poor [27], [28].

1) Virtual Reality: In a virtual reality system users interact with the visual and auditory scene which is presented to them, e.g., they can move around within the virtual scene. It is desirable that in such a system the spatial audio (stereo or multichannel) adapts to the visual scene. For example, depending on the position of the person in the virtual scene, sound sources are perceived in different directions. The flexible rendering capability of BCC can be used for placing the different audio sources in different directions within the virtual scene. For a flexible rendering capability without BCC, each source signal would need to be stored or transmitted separately. Therefore, with BCC, such systems can be implemented with a lower bit rate for the audio. Similarly, BCC can be used for interactive computer games. Especially games which are played over a network (e.g., Internet) benefit from a low bit rate for the audio transmission.

VII. Results

Several subjective tests were conducted to assess the performance of some of the presented schemes. The performance in a multitalker communication environment of the presented BCC for Flexible Rendering scheme is assessed in the first subjec-

tive test. Another set of subjective tests compares audio coding schemes based on BCC for Natural Rendering to conventional transform perceptual audio coders.

A. Subjective Test: Multitalker Communication

It is a well known fact that a listener's performance of responding to simultaneous talkers is better when these are perceptually spatially separated than without spatial separation [27], [29]. This improvement may be explained by informational masking that is reduced by differences in perceived locations of the talkers [28]. The aim of this test was to assess whether this is also the case for talkers spatialized with BCC for Flexible Rendering. Also, we were examining how close the performance of the BCC-based scheme comes to the performance of ideal separation. These results were previously reported [10] and are repeated here for completeness.

Twelve subjects were given a task which required responding to one of two simultaneous voice messages. This is a variation of the "cocktail party effect" of attending to one voice in the presence of others. The signals were presented to the subjects with high quality headphones (*Sennheiser HD 600*) in an acoustically isolated room. Five tests with five different signal types were conducted to assess the ability of a listener to respond to one of two simultaneous messages:

- *diotic*: sum signal of the talkers to both ears;
- $ICLD_{sep}$: separate talker signals rendered with ICLDs;
- $ICTD_{sep}$: separate talker signals rendered with ICTDs;
- $ICLD_{BCC}$: signals generated with BCC and ICLDs;
- $ICTD_{BCC}$: signals generated with BCC and ICTDs.

For the diotic signals, both talkers are localized in the center. For the other signals, the talkers are localized at the sides (left and right) with ICLDs of ± 16 dB and ICTDs of ± 500 μs. The ICC was chosen equal to one for all samples.

Each of the subjects took all of the tests in a randomized order. For the test we used the speech corpus introduced in [30]. A typical sentence of the corpus is "READY LAKER, GO TO BLUE FIVE NOW," where LAKER is the call sign and BLUE FIVE is a color-number combination. There are eight different call signs, four colors, and eight numbers. All these sentences are synchronized, i.e., the call sign and color-number combination occur at approximately the same time. One out of four female voices was randomly chosen for each of the two talkers in each item of each test. We chose the call sign, color, and number randomly with the restriction that the call sign assigned to the subject occurred in 50% of the cases. The subjects were instructed to respond when their call sign (always LAKER) was called by indicating the color-number combination of the sentence in which the call sign appeared. Each of the five tests consisted of 10 training items followed by 20 test items.

Table II shows the results for the case when the listener was called (50% of the items for each signal type). As expected these results suggest that the percentage of correct identification of the call sign and of the color-number combination significantly improve for the signals generated with separated source signals ($ICLD_{sep}$, $ICTD_{sep}$). The signals generated with BCC for Flexible Rendering ($ICLD_{BCC}$ and $ICTD_{BCC}$) are almost as good. For the case when the listener was not called, the percentages of the listeners responding was below 2% for all tests.

TABLE II
RESULTS FOR THE CASE WHEN THE LISTENERS WERE CALLED BY THEIR CALL SIGN. THE MIDDLE COLUMN SHOWS THE PERCENTAGE OF CORRECT IDENTIFICATION OF THE CALL SIGN AND THE RIGHT COLUMN SHOWS THE CONDITIONAL PERCENTAGE OF THE CORRECT COLOR-NUMBER COMBINATION GIVEN THAT THE LISTENER'S CALL SIGN WAS CORRECTLY IDENTIFIED

	call sign	*color-number*
diotic	70 %	64 %
$ICLD_{sep}$	78 %	98 %
$ICTD_{sep}$	85 %	88 %
$ICLD_{BCC}$	77 %	96 %
$ICTD_{BCC}$	78 %	91 %

TABLE III
CODERS AND BIT RATES FOR THE THREE SUBJECTIVE TESTS CONDUCTED. THE NUMBERS (1–5) DENOTE THE FIVE DIFFERENT CODING CONFIGURATIONS USED

Coder	*Bitrate for Stereo*	*Bitrate for BCC-based Coder*
PAC	(1) 64 kb/s	(2) $52 + 2$ kb/s
PAC	(3) 56 kb/s	(2) $52 + 2$ kb/s
MP3	(4) 40 kb/s	(5) $32 + 2$ kb/s

B. Subjective Test: Stereo Audio Coding

We compared audio coding schemes based on BCC for Natural Rendering with conventional stereo audio coders at various bit rates. Each conventional stereo audio coder was compared with the BCC-based scheme using the same audio coder for encoding the mono sum signal. For that purpose we used PAC [19] and MPEG-1 Layer 3 ("MP3") [31]. The MP3 encoder used is incorporated into Apple's iTunes program (by Fraunhofer IIS). These results were previously reported [11] and are repeated here for completeness.

The rows in Table III show the bit rates of the audio coders for encoding the stereo signals and the sum signals when used with BCC. The bit rate of the BCC schemes is shown as the sum of the mono audio coder bit rate and the BCC side information bit rate. The mono audio coder bit rate is chosen lower than the bit rate for stereo because for the same level of distortion and audio bandwidth, less bits are needed to encode the mono. For PAC we chose a sampling rate of 32 kHz and an audio bandwidth of 13.5 kHz. The parameters for the MP3 encoder were a bit rate of 40 kb/s for stereo and 32 kb/s for mono, joint-stereo coding enabled, and a sampling rate of 24 kHz. For both, PAC and MP3, we chose the fixed bit rate encoding mode.

The tests were carried out in a quiet (but not sound proof) reverberant listening room with a two-loudspeaker setup using high-end audio equipment with the listener's head located in the sweet-spot. To achieve a lower bit rate of the BCC side information, only ICLDs were used as inter-channel cues. We used an ICLD-to-ICTD scaling factor of $q = 25 \cdot 10^{-6}$ s/dB (24), with an ICLD range of ± 18 dB. This results in an ICTD range of ± 180 μs. We synthesized ICTDs despite of the fact that they are irrelevant for loudspeaker playback. This was done to show

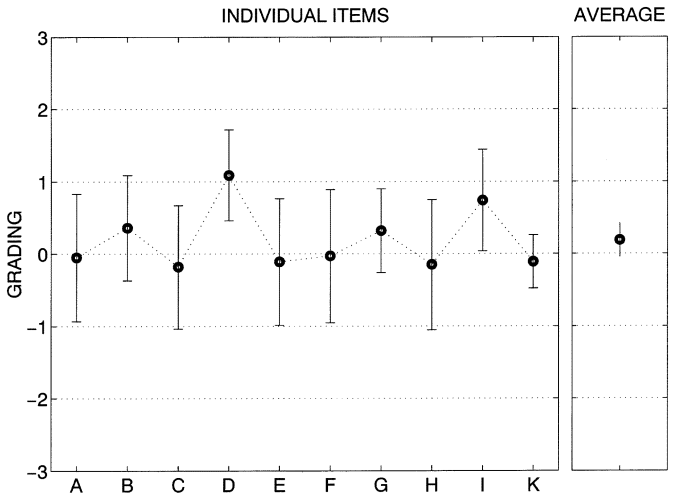

Fig. 11. Relative grading of the BCC-based coder operating at a bit rate of 52 + 2 kb/s versus stereo PAC at 64 kb/s (BCC is better than PAC for positive gradings, 1: slightly better, 2: better, and 3: much better).

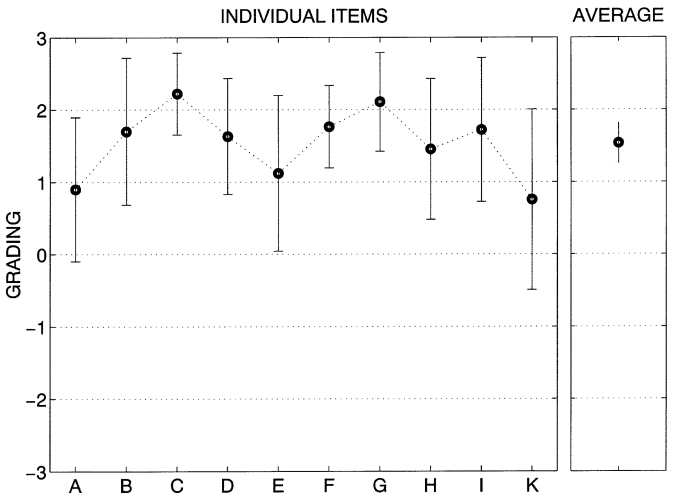

Fig. 12. Relative grading of the BCC-based coder operating at a bit rate of 52 + 2 kb/s versus stereo PAC at 56 kb/s (same grading scale as Fig. 11).

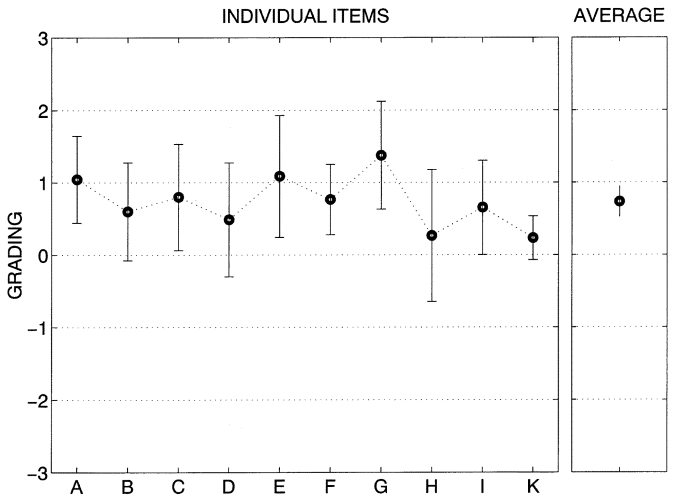

Fig. 13. Relative grading of the BCC-based coder operating at a bit rate of 32 + 2 kb/s versus stereo MP3 at 40 kb/s (same grading scale as Fig. 11).

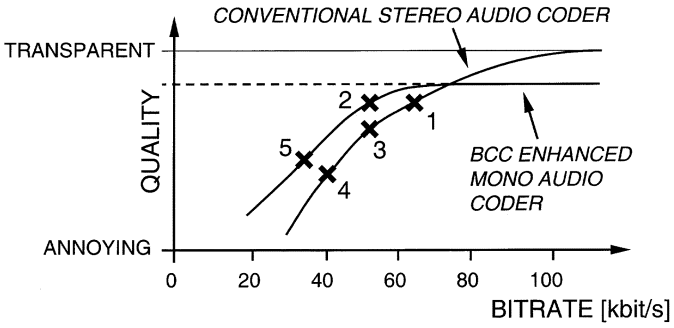

Fig. 14. At bitrates at which conventional perceptual transform audio coders operate near a transparent quality these are better than BCC-based schemes. At lower bitrates BCC-based schemes are better on average.

that the ICTD synthesis does not introduce major artifacts, and also because a signal with both, ICLDs and ICTDs, is preferred for headphone playback.

For each of the tests we chose the same 14 music clips. Each of these clips has a pronounced wide spatial image. BCC is challenged by a wide spatial image in the sense that it needs to perceptually separate audio sources. Also, for the conventional stereo audio coder a wide spatial image is challenging because the redundancy between the channels is small in that case resulting in a high bit demand. Different kinds of music signals such as Jazz, Rock, and percussive music were selected. Four of the clips were used as training items and 10 as test items. The type of test was a blind triple-stimulus test (ITU-R Rec. BS.562.3 [32]) to grade the quality difference of two processed versions with respect to a reference using a seven-grade comparison scale. For each test, 10 listeners were asked to participate. The 10 listeners were presented with triples of signals, each of 12 s length for each trial. The uncoded source signal (reference) was presented first followed by the coded clips of the conventional stereo audio coders and BCC-based coders in random order.

Figs. 11–13 show the results of the three subjective tests. Positive gradings correspond to preference for the BCC-based schemes. For every test the total bit rate of the BCC-based scheme was lower than the bit rate of the stereo audio coder. For the average of each test, the BCC-based scheme outperforms the stereo audio coder despite its lower bit rate. It has to be noted that the artifacts of the two coding schemes are quite different. The BCC-based coder generally modifies the spatial image more while the conventional stereo audio coders introduce more quantization distortions. From the test results one can conclude that the listeners preferred spatial image modification (as introduced primarily by BCC) over quantization distortions (introduced primarily by conventional coders).

Derived from the test results (Figs. 11–13), Fig. 14 shows qualitatively the subjective quality of each coding configuration that was used for the tests (the same numbering as in Table III is used): (1) Stereo PAC 64 kb/s, (2) Stereo PAC 56 kb/s, (3) BCC with mono PAC 52 + 2 kb/s, (4) Stereo MP 3 40 kb/s, and (5) BCC with mono MP3 32 + 2 kb/s. At bitrates high enough for transparent or nearly transparent coding the conventional coder is better since BCC can generally not achieve transparency. The test results give an indication that for bitrates in the range of about 24–64 kb/s the BCC-based coding scheme

has better quality than conventional perceptual transform audio coders for stereo. The lower the bit rate the more is the BCC-based coding scheme at an advantage.

For the kinds of music signals chosen for the tests (Jazz, Rock, and percussive music), BCC generally provides a good quality of the spatial image using only ICLDs and ICTDs as was done in the subjective test. For recordings with a high amount of uncorrelated reverberation in the audio channels, such as classical recordings, it is desirable to also use ICC cues in order to restore the diffuseness of the reverberation. Informal listening revealed that the ICC synthesis does not only restore some of the diffuse reverberation, but also seems to improve the stability of the spatial image in many cases.

We also did an informal listening test rendering multichannel audio signals with BCC. BCC for Flexible Rendering seems to perform about equally well for four or five loudspeakers as it does for stereo. BCC for Natural Rendering performs also well for multichannel audio signals and nonreverberant audio material. However in the case of only low energy coming from the rear speakers mimicking a concert hall, the realistic "concert hall" impression is lost. We hope to address this in the future with improved ICC analysis and synthesis.

VIII. Conclusions

We reviewed Binaural Cue Coding (BCC) and presented several additions for improved performance and versatility. BCC is a method for multichannel spatial rendering based on one down-mixed audio channel plus side information. One BCC scheme is applicable to low bitrate stereo and multichannel audio coding. The other scheme allows flexible rendering of multiple auditory objects at the receiver.

BCC operates in a short-time spectral domain. The short-time spectra of each of the output channels is obtained by modifying the corresponding spectra of the down-mixed audio channel. The spectra are modified such that desired level difference, time difference, and cross-correlation cues appear in subbands with perceptually motivated bandwidths.

Novel additions include improved down-mixing reducing undesired cancellation/amplification of signal components and synthesis of cross-correlation cues for improved synthesis of diffuse spatial images (e.g., reverberant recordings).

The described implementations have a computationally low complexity and a delay low enough for two-way communication. It is explained how the two types of BCC schemes can be combined with conventional mono audio and speech coders. Proposed applications for these systems include tele-conferencing, virtual reality systems, gaming, and low-bit-rate stereo and multichannel audio coding. It is an important aspect of schemes based on BCC that existing mono communication systems can be enhanced for spatial audio in a backward compatible manner if the BCC side information can be embedded into the existing transmission link.

A subjective test with simultaneous talker signals presented to a listener suggests that the perceptual separation resulting from BCC for Flexible Rendering is nearly as good as ideal separation with respect to detecting and responding to these talkers. Another set of subjective tests comparing a BCC for Natural Rendering based scheme for coding of stereo audio signals to conventional perceptual transform coders shows that for bit rates in the range of about 24–64 kb/s the BCC-based scheme is better.

Acknowledgment

The authors thank J. Chen, T. Gänsler, A. Härmä, P. Kroon, Y. Shoham, F. Soong, and M. Vetterli for valuable suggestions on the draft manuscript. They thank the anonymous reviewers for their valuable comments.

References

[1] C. Faller and F. Baumgarte, "Binaural cue coding: A novel and efficient representation of spatial audio," in *Proc. ICASSP 2002*, Orlando, FL, May 2002.

[2] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, pp. 456–466, June 1997.

[3] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Amer.*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[4] J. Herre, K. Brandenburg, and D. Lederer, "Intensity stereo coding," in *Proc. AES 96th Convention*, Feb. 1994.

[5] F. Baumgarte and C. Faller, "Why binaural cue coding is better than intensity stereo coding," in *Preprint 112th Conv. Aud. Eng. Soc.*, May 2002.

[6] A. Härmä, "Coding principles for virtual acoustic openings," in *Proc. AES 22nd Int. Conf. Virtual, Synthetic, and Entertainment Audio*, Espoo, Finland, June 2002.

[7] M. C. Kelly and A. I. Tew, "The continuity illusion revisited: Coding of multiple concurrent sound sources," in *Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, 2002.

[8] F. Baumgarte and C. Faller, "Binaural cue coding—Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 509–519, Nov. 2003.

[9] J. Blauert, *Spatial Hearing. The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1983.

[10] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *Proc. IEEE Workshop Applications Signal Processing to Audio and Acoust.*, Oct. 2001.

[11] ——, "Binaural cue coding applied to stereo and multi-channel audio compression," in *Preprint 112th Conv. Aud. Eng. Soc.*, May 2002.

[12] ——, "Binaural cue coding applied to audio compression with flexible rendering," in *Preprint 113th Conv. Aud. Eng. Soc.*, Oct. 2002.

[13] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.

[14] F. Baumgarte and C. Faller, "Estimation of auditory spatial cues for binaural cue coding (BCC)," in *Proc. ICASSP 2002*, Orlando, FL, May 2002.

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

[16] A. V. Oppenheim and R. W. Schaefer, *Discrete-Time Signal Processing, Signal Processing Series*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[17] H. S. Malvar, *Signal Processing With Lapped Transforms*. Norwood, MA: Artech House, 1992.

[18] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[19] D. Sinha, J. D. Johnston, S. Dorward, and S. Quackenbush, "The perceptual audio coder (PAC)," in *The Digital Signal Processing Handbook*, V. Madisetti and D. B. Williams, Eds. Boca Raton, FL: CRC, 1997, ch. 42.

[20] S.-Y. Chung and H. Lou, "Multilevel rs/convolutional concatenated coded QAM for hybrid IBOC-AM broadcasting," *IEEE Trans. Broadcast.*, vol. 46, pp. 49–59, Mar.h 2000.

[21] R. L. Cupo, M. Sarraf, M. Shariat, and M. Zarrabizadeh, "An OFDM all-digital in-band-on-channel (IBOC) AM & FM radio solution using the pac encoder," *IEEE Trans. Broadcast.*, vol. 44, no. 1, 1998.

[22] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation—An overview of the fundamental problem," *IEEE Signal Processing Lett.*, vol. 2, pp. 148–151, Aug. 1995.

[23] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. New York: Springer, 2001.

[24] "Pulse Code Modulation (PCM) of Voice Frequencies," 1988 Blue Book, ITU-T, Rec. G.711, 1993.

[25] "Coding at 24 and 32 kbit/s for Hands-Free Operation in Systems With Low Frame Loss,", ITU-T, Rec. G.722.1, 1999.
[26] "Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP),", ITU-T, Rec. G.729, 1996.
[27] W. Spieth, J. F. Curtis, and J. C. Webster, "Responding to one of two simultaneous messages," *J. Acoust. Soc. Amer.*, vol. 26, no. 3, pp. 391–396, 1954.
[28] R. L. Freyman, K. S. Helfer, D. D. McCall, and R. K. Clifton, "The role of perceived spatial separation in the unmasking of speech," *J. Acoust. Soc. Amer.*, vol. 106, no. 6, pp. 3578–3588, Dec. 1999.
[29] R. S. Bolia, M. A. Ericson, W. T. McKinley, and B. D. Simpson, "A cocktail party effect in the median plane?," *J. Acoust. Soc. Amer.*, vol. 105, pp. 1390–1391, 1999.
[30] R. S. Bolia, W. T. Nelson, M. A. Ericson, and B. D. Simpson, "A speech corpus for multitalker communications research," *J. Acoust. Soc. Amer.*, vol. 107, no. 2, pp. 1065–1066, Feb. 2000.
[31] K. Brandenburg and G. Stoll, "ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio," *J. Audio Eng. Soc.*, pp. 780–792, Oct. 1994.
[32] (1990) ITU-R Rec. BS.562.3. [Online]. Available: http://www.itu.org

Christof Faller received the M.S. (Ing.) degree in electrical engineering from ETH Zurich, Switzerland, in 2000. During his studies, he worked as an independent consultant for Swiss Federal Labs, applying neural networks to process parameter optimization of sputtering processes and spent one year at the Czech Technical University (CVUT), Prague.

In 2000, he became a Consultant for the Speech and Acoustics Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ. After one and a half year consulting, partially from Europe, he became a Member of Technical Staff, focusing on new techniques for audio coding applied to digital satellite radio broadcasting. Recently, he moved to the newly formed Media Signal Processing Research Department of Agere Systems, a Lucent spin-off. His research interests include generic signal processing, specifically audio coding, control engineering, and neural networks.

Mr. Faller won first prize in the Swiss national ABB (Asea Brown Boveri) Youth Science Contest organized in honor of the 100-year existence of ABB (formerly BBC) in 1991.

Frank Baumgarte received the M.S. and Ph.D. (Dr.-Ing.) degrees in electrical engineering from the University of Hannover, Germany, in 1989 and 2000, respectively. During the studies and as independent consultant he implemented real-time signal processing algorithms on a variety of DSPs including a speech coder and an MPEG-1 Layer 3 decoder. His dissertation includes a nonlinear physiological auditory model for application in audio coding.

In 1999 he joined the Acoustics and Speech Research Department, Bell Labs, Lucent Technologies, Murray Hill, NJ, where he was engaged in objective quality assessment and psychoacoustic modeling for audio coding. He became a Member of Technical Staff of the Media Signal Processing Research Department, Agere Systems, a Lucent spin-off, in 2001, focusing on advanced perceptual models for multichannel audio coding, auditory scene analysis and music synthesis. His main research interests in the area of acoustic communication include the understanding and modeling of the human auditory system physiology, psychophysics, audio and speech coding, and quality assessment.