BINAURAL CUE CODING: A NOVEL AND EFFICIENT REPRESENTATION OF SPATIAL AUDIO

Christof Faller and Frank Baumgarte

Media Signal Processing Research
Agere Systems, Murray Hill, NJ, U.S.A.

ABSTRACT

We present a novel concept for representing multi-channel audio signals: Binaural Cue Coding (BCC). BCC aims at separating the basic audio content and the information relevant for spatial perception. A multi-channel audio signal is represented as a mono signal and BCC parameters. We present two types of applications of BCC. Firstly, a number of separate sound source signals are reduced to a mono signal and BCC parameters. In this case, the decoder has control over the location of each source in auditory space. In other words, the decoder can render spatial images as if the separate source signals were given. Secondly, a multi-channel audio signal is reduced to a mono signal and BCC parameters. In this case the decoder generates a multi-channel signal with a spatial image similar to the spatial image of the input signal of the encoder. Results from a subjective test suggest that BCC, combined with existing mono audio coders, offers better quality than conventional stereo and multi-channel perceptual transform audio coders for a wide range of bitrates.

1. INTRODUCTION

With conventional audio coders such as PAC [1] or MPEG-2 AAC [2] the bitrate scales as the number of channels increases. Two techniques are commonly used for reducing the bitrate for encoding of stereo and multi-channel audio signals: *(1) Mid/Side (M/S) coding* [3] is used to reduce the redundancy between pairs of channels (e.g. left and right). With M/S coding the sum and difference of left and right are encoded instead of the left and right audio signals. Given the decoded sum and difference signals, the decoder can recover the left and right signals. *(2) Intensity Stereo* [4] is related to the approach we propose in this paper. However, it is less general and has several other drawbacks. Intensity stereo as used in MPEG-2 AAC [2] transmits for each coding band of the high frequencies only the sum signal along with a scalar representing the energy distribution among channels. The time-frequency resolution of intensity stereo is the same as for the audio coder's coding bands and therefore can not be optimized for spatial perception. Additionally, the filterbanks used in audio coders are critically sampled. Therefore, spectral modifications that are carried out for intensity stereo can lead to aliasing artifacts. Because of these limitations, intensity stereo is mostly suitable for non-transparent audio coding and is applied mainly at high frequencies.

The concept of Binaural Cue Coding (BCC) is the separation of the information relevant for the spatial perception of multi-channel audio signals and the basic audio content. BCC represents multi-channel signals as a mono audio signal and *BCC parameters*. The mono audio signal is just the sum signal derived from all sound sources which are to be part of the spatial image of the multi-channel signal. In this context, a spatial image is the perception of the sound source locations of a human listener. The spatial image of a stereo or multi-channel audio signal is the spatial image which is perceived when the signal is played back over loudspeakers or headphones.

BCC has potential for many applications. We propose two types of schemes applying BCC, type I and II. Type I BCC encoder is shown in Fig. 1. It takes M separate sound source signals as input. These are mono signals without any spatial information. If a group of several sound sources are to be placed at the same spatial location then one input signal can be the sum signal of that group. As will be shown, with this type of scheme the BCC decoder has control over the locations of the sound sources in the spatial image independently of the encoder. Also, the number of playback channels C is determined at the decoder.

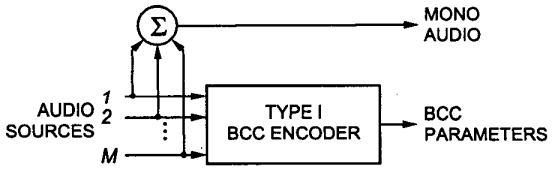

Fig. 1. Type I BCC encoder with M mono sound source signals as input.

Type II BCC encoder is shown in Fig. 2. The input signal is a multi-channel audio signal. With this type of scheme the goal is that the BCC decoder generates a multi-channel audio signal with a spatial image similar to that of the encoder input signal.

Applications for type I BCC schemes are tele-conferencing with stereo or multi-channel audio reproduction [5]. The mono sum signal can be encoded with a mono audio or speech coder as desired. Such a system can be enhanced for stereo or multi-channel audio in a backwards compatible manner, if the data link

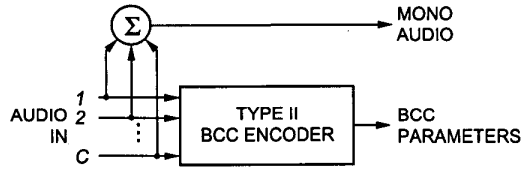

Fig. 2. Type II BCC encoder with a multi-channel audio signal as input.

0-7803-7402-9/02/$17.00 ©2002 IEEE

for an existing mono conferencing system allows inclusion of the BCC parameters. Another application is rendering of spatial images for virtual reality at low bitrates since the BCC decoder can adapt the rendering of the spatial image according to the visual display of the virtual reality.

Applications for type II BCC schemes are stereo and multi-channel audio coding. The bitrate for encoding stereo or multi-channel signals with conventional perceptual transform audio coders is significantly higher than the bitrate necessary for encoding a mono audio signal. BCC combined with a conventional mono audio coder allows encoding of stereo and multi-channel signals at the rate of the mono audio coder with a small additional overhead for the BCC parameters. Also, existing mono broadcasting systems can be enhanced for stereo or multi-channel audio in a backwards compatible manner if inclusion of the BCC parameters into the existing data link is possible.

The rationale for BCC is presented in Section 2. Section 3 shows how the different types of BCC schemes are implemented. Subjective test results of a comparison of a BCC enhanced mono audio coder compared with the same audio coder for stereo are given in Section 4. In Section 5 conclusions are drawn.

2. RATIONALE FOR BCC

Sound from a single freefield source reaches the two ears of a listener with an interaural level difference (ILD) and an interaural time difference (ITD). The ILD and ITD determine the perceived azimuth of a sound source in the horizontal plane. A more precise description of binaural directional cues is the direction-dependent transfer function of sound to the eardrum (head related transfer function HRTF [6]).

For headphone listening the ILD and ITD correspond to the inter-channel level difference (ICLD) and inter-channel time difference (ICTD) of the left and right signal. For loudspeaker playback the ICLD and ICTD determine the location of a source between a pair of loudspeakers by indirectly determining binaural localization cues. Spatial cues in this paper always refer to the more general inter-channel cues ICLD and ICTD and possibly HRTFs for generating binaural signals for headphone playback.

A mono signal of a sound source can be processed such that the sound source is spatially placed by directly (headphone playback) or indirectly (loudspeaker playback) providing binaural localization cues to the ear [6]. For example, a stereo signal with a single source, placed somewhere between left and right, can be generated by imposing an ICLD and ICTD between the left and right channel.

A stereo signal with several sources individually placed in auditory space as desired can be generated as follows. For each source a separate stereo signal is first generated with spatial cues determining the location of the source. Then all resulting stereo signals are added to one stereo signal.

If the several source signals occupy non-overlapping regions in the time-frequency plane, then it is possible to render a spatial image as desired given only the sum signal of the sources. This is achieved by taking the sum signal and by applying to each region in the time-frequency plane the spatial cues corresponding to the source to which the region belongs to. Figure 3 shows an example of the time-frequency plane in which three sources occupy non-overlapping regions.

We assume, that even in the case when different sources occupy overlapping regions in the time-frequency plane, it is still possible to render spatial images as desired, with a high degree of flexibility when taking into account the properties of the auditory system [5]. This forms the basis of BCC. The regions are formed such as to approximate the ideal case of complete separation in the time-frequency plane. As in the ideal case only the spatial cues corresponding to one sound source are associated with each region.

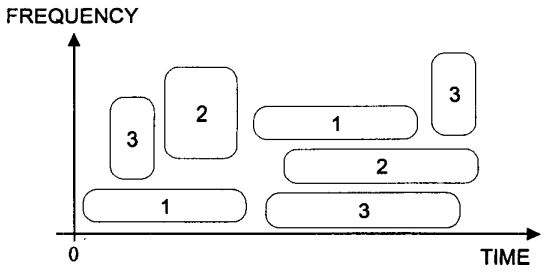

Fig. 3. Three sources which occupy non-overlapping regions in the time-frequency plane.

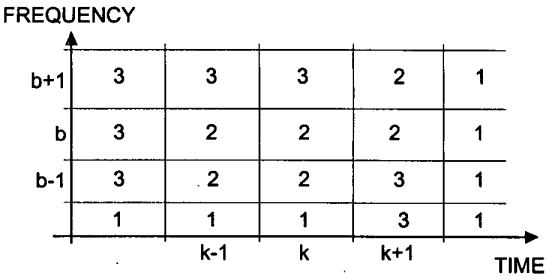

Fig. 4. Each of the partitions associated with one of the sound sources contained in the mono sum signal.

3. IMPLEMENTATION OF BCC

For practical reasons the time-frequency plane is divided into a grid of non-overlapping partitions as shown in Fig. 4. Each of these partitions is associated with one of the sound sources which are contained in the mono sum signal. To the b^{th} partition at time k the index of the source which has most energy within the partition is assigned, $s_{b,k}$. Figure 4 shows an example of how partitions are associated with three sources. For brevity the time index k will be dropped in the remaining part of the paper.

3.1. Definition of Spatial Cues for Multiple Channels

In the general case of C playback channels the spatial cues are given for each channel relative to a reference channel. Without loss of generality, channel number 1 is defined as the reference channel. Figure 5 shows how the spatial cues are defined between the reference channel and each other channel for the b^{th} partition. For example, $\{\Delta L_{ib},\ \tau_{ib}\}$ are the level difference and time difference between channel 1 and channel $i+1$ for the b^{th} partition.

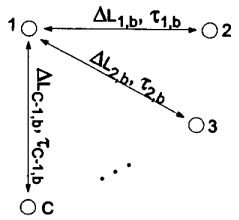

Fig. 5. For the general case of C channels the spatial cues are defined relative to a reference channel for the b^{th} partition.

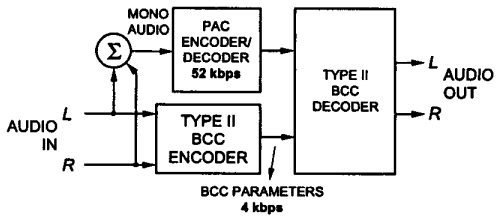

Fig. 6. BCC combined with PAC for a total bitrate of 56 kbit/s.

3.2. Type I BCC

A type I encoder (Fig. 1) encodes and transmits the source indices s_b to the BCC decoder as BCC parameters. The BCC decoder has a table in which it stores the spatial cues for each source. This table in the decoder determines the location of each of the M sources in the spatial image. Also the number of playback channels C is determined by the decoder. The spatial cues stored in the table for the m^{th} source between channel number $i+1$ and the reference channel 1 are: ICLD_{im}, ICTD_{im}, and/or left and right HRTF frequency responses H_m^L, H_m^R. At time k the BCC decoder generates a multi-channel signal by applying for the b^{th} partition the spatial cues for source s_b. In the case of synthesizing a binaural signal these are $H_{s_b}^L$ and $H_{s_b}^R$ and in the case of synthesizing a multi-channel audio signal these are $\Delta L_{ib} = \mathrm{ICLD}_{i,s_b}$ and $\tau_{ib} = \mathrm{ICTD}_{i,s_b}$.

3.3. Type II BCC

For a type II encoder (Fig. 2) a multi-channel audio signal is given without explicit knowledge about the separate sound source signals. In this case at time k the spatial cues $\{\Delta L_{ib}, \tau_{ib}\}$ are estimated for the b^{th} partition by inter-channel analysis. These cues are quantized, encoded, and transmitted as BCC parameters to the decoder. In this case the BCC decoder uses the quantized spatial cues to generate a multi-channel signal with the same number of channels C and a similar spatial image as the encoder input signal. A scheme for extracting spatial cues for a type II BCC encoder is presented in a separate paper submitted to this conference [7].

3.4. Synthesis of Spatial Cues

For both types of BCC decoders the last processing step is to generate the multi-channel audio signal given the mono signal and spatial cues. The signals of the C output channels are computed in the frequency domain and are obtained by modification of the complex spectrum of the mono signal by synthesizing spatial cues between pairs of channels as described in [5].

4. RESULTS

An experiment is designed to evaluate the subjective quality of an audio coding scheme based on BCC type II compared to a conventional stereo perceptual transform audio coder. As shown in Fig. 6 the mono audio output signal of the BCC encoder is encoded with mono PAC at 52 kbit/s. Therefore, the total bitrate of this BCC based audio coder is 56 kbit/s. We compared this coder with stereo PAC operating with the same bitrate.

To reduce the bitrate of the BCC parameters only level differences are used as spatial cues. The level differences are quantized with a uniform quantizer with 15 steps. A Huffman coder is used to reduce the bitrate of the quantized level differences. The resulting bitrate is approximately 4 kbit/s and is fairly constant. For the BCC based scheme no binaural masking level difference (BMLD, [6]) needs to be considered because the signal and the quantization noise introduced by the mono audio coder are always directionally located at the same place for each frequency.

For the test we chose 14 music clips. Each of these clips has a pronounced wide spatial image. 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 test was carried out with a two loudspeaker setup using high end audio equipment with the listener's head located in the sweet spot. A blind triple stimulus test [8] was conducted to grade the quality difference with respect to the reference using a seven-grade comparison scale. The 9 listeners were presented with a triple 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 BCC based coder and stereo PAC in random order. The average subjective comparison scores of both schemes are shown in Fig. 7. An inspection of Fig. 7 reveals that at the same bitrate the BCC coder offers better quality.

The test was repeated with 10 listeners with the same items for the BCC scheme at 56 kbit/s but stereo PAC operating at 64 kbit/s. The results are presented in Fig. 8. One can see that both, the BCC based coder and stereo PAC, give about the same quality while the BCC scheme saves about 8 kbit/s. Even when for this testing scenario the BCC based coder and PAC operate at about the same subjective quality it has to be noted that the artifacts of these two coders are quite different. The BCC based coder generally modifies the spatial image more while stereo PAC introduces more distortions.

At bitrates high enough for transparent coding of stereo signals with PAC, the BCC based coder will be worse because it does not aim at transparency. Similar results can be expected for other subband coders. The test results give an indication that for bitrates lower than about 64 kbit/s the BCC based coder is better than conventional perceptual transform audio coders for stereo. The lower the bitrate the more is the BCC based coder at an advantage.

For a possible conferencing application we informally tested several speech coders (G.729 [9], G.722.1 [10]) for coding of the mono sum signal. These speech coders work well for encoding of several simultaneous voices which are to be rendered to a spatial image with BCC. A different subjective test for identification of messages in a multi-talker environment with headphone listening was conducted with signals generated by a BCC type I scheme with ICLD and ICTD. The perceived separation of the talkers with

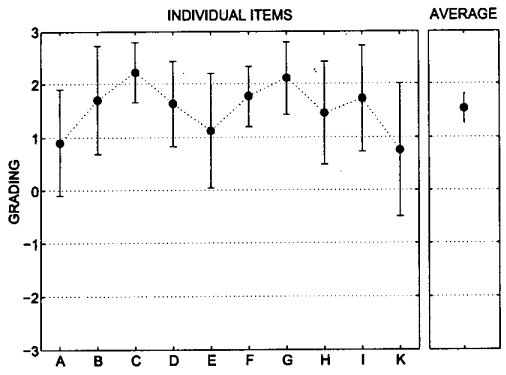

Fig. 7. Relative grading of the BCC based coder versus stereo PAC with both operating at a bitrate of 56 kbit/s (BCC is better than PAC for positive gradings, 1: slightly better, 2: better, 3: much better).

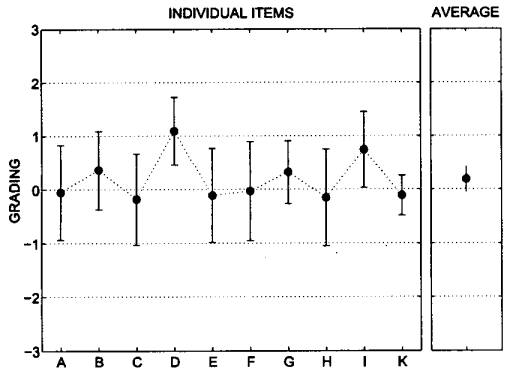

Fig. 8. Relative grading of the BCC based coder with a bitrate of 56 kbit/s versus stereo PAC at 64 kbit/s (same grading scale as Fig. 7).

BCC and true separation yielded about the same amount of improvement for the identification rate compared to a diotic presentation of the talkers. Details can be found in [5].

5. CONCLUSIONS

The concept of Binaural Cue Coding (BCC) which represents the spatial information and the basic audio content separately was presented. Multi-channel signals can be generated given one mono signal and BCC parameters. A promising feature of BCC is that existing mono systems can be upgraded for stereo or multi-channel playback in a backwards compatible manner, if the inclusion of the additional BCC parameters is possible within the given communications channel. This may enable BCC to have great potential for applications such as tele-conferencing and audio broadcasting.

Two types of schemes based on BCC have been designed. The first type of BCC scheme reduces a number of mono sound source signals to a mono signal and BCC parameters. The decoder can render spatial images by individually controlling the locations of the input sources that were supplied to the encoder. Such a decoder can be viewed as a scheme for rendering spatial images which only requires one mono signal as input as opposed to conventional schemes [11] that require all individual sound source signals as input. This is especially interesting for applications in which a remote client is to be given the flexibility of rendering custom spatial images.

The second type of BCC scheme reduces a given multi-channel signal to a mono signal and BCC parameters. In this case the decoder generates a multi-channel audio signal with the aim of reproducing the spatial image of the encoder input. In combination with existing audio coders this type of BCC scheme can be used to encode spatial audio for stereo or multi-channel playback with only a small additional overhead. The results from a subjective test suggest, that for bitrates below 64 kbit/s, a BCC based coder achieves improved perceived quality compared to conventional perceptual transform stereo audio coders.

Acknowlegements

We would like to thank Jiashu Chen, Jingdong Chen, Aki Härmä, Peter Kroon, Yair Shoham, and Martin Vetterli for their suggestions and contributions.

6. REFERENCES

[1] 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., chapter 42. CRC Press, IEEE Press, Boca Raton, Florida, 1997.

[2] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 advanced audio coding," *J. Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–814, 1997.

[3] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *ICASSP-92 Conference Record*, 1992, pp. 569–572.

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

[5] 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.*, Oct. 2001.

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

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

[8] ITU-R Rec. BS.562.3, 1990, http://www.itu.org.

[9] ITU-T, Rec. G.729, *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP)*, March 1996.

[10] ITU-T, Rec. G.722.1, *Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss*, Sept. 1999.

[11] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*, Academic Press, Cambridge, MA, 1994.