Audio Engineering Society

Convention Paper

Presented at the 116th Convention
2004 May 8–11 Berlin, Germany

MP3 Surround: Efficient and Compatible Coding of Multi-Channel Audio

Juergen Herre[1], Christof Faller[2], Christian Ertel[1], Johannes Hilpert[1], Andreas Hoelzer[1], and Claus Spenger[1]

[1] Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany

[2] Agere Systems, Allentown, PA 18109, USA

ABSTRACT

Finalized in 1992, the MP3 compression format has become a synonym for personalized music enjoyment for millions of users. The paper presents a novel extension of this popular format which adds support for the coding of multi-channel signals, including the widely used 5.1 surround sound. As a prominent feature of the extended format, complete backward compatibility with existing stereo MP3 decoders is retained, i.e. standard decoders reproduce a full stereo downmix of the multi-channel sound image. The paper discusses the underlying advanced technology enabling the representation of multi-channel sound at bitrates that are comparable to what is currently used to encode stereo material. Results for subjective sound quality are presented and related activities of the MPEG standardization group are reported.

1. INTRODUCTION

With the broad availability of the Internet and modern computer technology, a palette of perceptual audio coding schemes has found widespread use in multimedia applications. Among these codecs, the popular MP3 compression format is one of the most frequently used coding schemes for both hardware-based playback devices (such as portable audio players, PDAs, and CD/DVD players) and software-based applications (media players, music jukebox software etc.). Initially mainly associated with Internet audio, MP3 has turned into a synonym for personalized music enjoyment for millions of users worldwide.

Formally speaking, the name MP3 refers to the Layer 3 coding scheme of the ISO/MPEG-1 and ISO/MPEG-2 Audio specifications [21] [22] that were finalized in 1992 and 1994, respectively[1]. Compliant to these standards, MP3 capability usually supports the

[1] MPEG-2 Audio also specifies (matrixed) backward compatible multi-channel audio coding schemes. These are not well represented in the marketplace and are not included in MP3 implementations.

encoding/decoding of mono or stereo[2] audio at a number of common sampling rates (16, 22.05, 24, 32, 44.1, 48kHz) and bit-rates up to 320 kbit/s. Thus, MP3 technology has been synonymous to stereo (non-multi-channel) audio storage and transmission for a period of more than 10 years.

More recently, however, multi-channel sound reproduction setups in home environment have become more popular. This trend is driven mostly by multi-channel sound that comes together with movie content, be it as discrete 5.1 channel sound (DVD Video/AC-3) or as matrixed analog multi-channel sound (e.g. Prologic, Logic7). Although market penetration in Europe is lagging behind the US, the trend towards owning a "home theater" setup is clearly detectable. A further growing application area in this context is multi-channel for automotive applications. More and more high-end cars will be equipped with multi-channel audio capabilities by default.

The consequence of this trend will be a clear demand for multi-channel audio-only material that is played on home theater setups without a video component. In the high-end segment, this desire will be served by emerging media, such as DVD-Audio and SA-CD. For bandwidth-constrained multi-channel applications, the MPEG-2/4 Advanced Audio Coding (AAC) scheme has been developed as an efficient multi-channel audio format providing excellent sound quality at bit-rates in the range of 256-320kbit/s for 5-channel material [20] [5].

The general increase in significance of multi-channel sound in consumer audio raises the question about a multi-channel MP3 audio format that could serve the MP3 user community with efficient representation of "surround sound". Naturally, it is important for such an advanced format to maintain some degree of compatibility with existing MP3 systems in order to leverage the existing base of deployed systems. As can be expected for any transition to a new generation of technology, there will always be a vast number of consumers who will stick to their existing stereo reproduction setups for the foreseeable future. To accommodate a smooth transition towards multi-channel transmission and reproduction technology, an ideal advanced MP3 coding format needs to support equally well both user populations (i.e. owners of traditional stereo setups and multi-channel enabled reproduction setups).

This paper introduces the concept behind a new format that extends current MP3 capabilities towards the efficient representation of multi-channel audio. Based on recent advances in multi-channel audio coding technology, it permits the representation of 5.1 sound at bit-rates that are comparable to those currently used for representing 2-channel material. Due to its underlying structure, this "MP3 Surround" format is fully backward compatible with existing MP3 decoders in the sense that these will decode a stereo downmix of the multi-channel sound image. Enhanced decoders make use of additional aspects of the extended bitstream format in order to reproduce the full multi-channel sound image.

An alternative non-backward compatible approach for extending MP3 technology towards multi-channel in the context of MPEG-4 has been discussed in [18] and will not be elaborated in this paper.

The subsequent sections briefly review state of the art methods in perceptual coding of stereo and multi-channel audio and give a short overview of Binaural Cue Coding (BCC), a coding method for spatial audio, which forms part of the technological basis of the MP3 Surround format. Next, it discusses how the traditional BCC approach is extended to permit the expansion of stereo signals into a multi-channel sound image. Then, first results of subjective listening tests are presented and the status of recent related standardization activities within the MPEG group is described. This will be rounded up with a discussion of several applications in multi-channel sound enabled by the MP3 Surround format.

2. PERCEPTUAL CODING OF STEREO AND MULTI-CHANNEL AUDIO

The basic approach of monophonic perceptual audio coding relies on two general principles:

- The reduction of the signal's redundancy by means of entropy coding, utilization of the "nonflatness" of its spectral envelope or predictive approaches facilitate a more compact representation of the signal without irreversible loss of information.

- The exploitation of the signal's irrelevance by means of shaping the coder quantization noise in frequency and time allows a signal representation for which the

[2] In this paper the term "stereo" always refers to two-channel stereophony.

coding distortion is not perceptible to the human listener ("masking") while consuming only a fraction of the signal's original data rate.

When coding audio material which is represented by several audio channels (i.e. stereo or multi-channel material), state of the art coding schemes generally make use of a number of additional techniques that enable the efficient joint coding of these audio channels and account for the spatial aspects of human auditory perception. The two most well known techniques for joint stereo coding of signals are Mid/Side (M/S) stereo coding (also known as "Sum/Difference coding") and intensity stereo coding. Both joint coding strategies may be combined by selectively applying them to different frequency regions. A brief review will be given next.

2.1. M/S Stereo Coding

In M/S stereo coding [16] the sum and the difference signals of the two stereo channels are quantized and coded rather than processing the left and right channel signals separately (L/R coding). Switching between L/R and M/S coding effectively controls the imaging of coding noise, in relation to the imaging of the original signal and in this way prevents spatial unmasking. Specifically, this technique addresses the issue of "Binaural Masking Level Difference" (BMLD) [4], where a signal at lower frequencies (below 2 kHz) can result in up to 20 dB difference in the masking threshold depending on the phase and correlation of the signal and probe noise present. In such cases, proper coding of the stereo signal in L/R mode can require more bits than two transparently coded monophonic signals whereas M/S coding provides both proper spatial masking and – very frequently – further bitrate savings.

M/S stereo coding has been used extensively in the MP3 and MPEG-2/4 AAC audio coders. In both cases, the switching state (L/R or M/S coding) is determined selectively in time on a block-by-block basis and transmitted to the decoder. As a refinement, AAC also allows for frequency selective switching (for each scale factor band the L/R vs. M/S decision is made individually).

In order to code multi-channel audio, M/S stereo coding is applied to channel pairs of the multi-channel signal, i.e. between pairs of channels that are arranged symmetrically on the left/right listener axis. In this way, imaging problems due to spatial unmasking are avoided to a large degree [17].

2.2. Intensity Stereo Coding

Intensity stereo coding is a joint stereo coding technique applied to reduce "perceptually irrelevant information" between audio channels and is described in [29]. Specifically, intensity stereo coding exploits the fact that the perception of high frequency sound components mainly relies on the analysis of their energy-time envelopes [4]. When applied to each coder frequency band, the corresponding spectral coefficients (or subband samples) of both signal channels are replaced by a single sum signal and a direction angle (azimuth or equivalent information). The azimuth controls the intensity stereo position of the auditory event created at the decoder. Only one azimuth is transmitted for a coder frequency band. Using this information, the original energy-time envelopes of the coded channels are preserved approximately by means of a scaling operation of the transmitted sum signal such that each channel signal is reconstructed with its original level after decoding. Intensity stereo coding is capable of significantly reducing the bitrate for stereo and multi-channel audio due to the reduction in the effective number of transmitted spectral coefficients. However, its application is limited since intolerable distortions can occur if it is applied to the full audio bandwidth or for use with signals with a highly dynamic and wide spatial image [15] [28]. Potential improvements of the approach are constrained because the time-frequency resolution is given by the core audio coder and cannot be modified without considerable complexity being added.

Intensity stereo coding has been used widely under various names in both MPEG audio coders (MPEG-1, MPEG-2 [27] and MPEG-2/4 AAC [17]) and other schemes (e.g. AC-3 [8]).

For the coding of multi-channel audio, intensity stereo coding can be applied either to pairs of channels that are arranged symmetrically on the left/right listener axis (similar to M/S stereo coding) or to arbitrary groups of audio channels.

3. BINAURAL CUE CODING (BCC)

Binaural cue coding (BCC) [11] [13] is a coding scheme based on a compact and perceptually motivated representation of multi-channel audio signals. BCC has

different applications in scenarios where such signals need to be represented compactly. Specifically, BCC for *natural rendering* [12] [13] is a parametric multi-channel audio coding technique, similar to intensity stereo coding.

Some of the limitations of intensity stereo coding are overcome by BCC in the way that different filterbanks are employed for coding of the audio waveform and for parametric stereo [1]. Most audio coders use a *Modified Discrete Cosine transform* (MDCT) for coding of audio waveforms. The advantages of using a different and more suitable filterbank for parametric stereo are reduced aliasing [1] and more flexibility, such as the ability to efficiently synthesize not only intensities (ICLD), but also time delays and coherence between the audio channels. Another notable conceptual difference between intensity stereo coding and BCC is that the latter is able to operate on the full-band audio signal and transmits a single monophonic time domain signal whereas intensity stereo coding is usually applied to high frequency components of a signal and thus transmits only spectral portions as a mono signal component.

Figure 1 shows a scheme for BCC for natural rendering. The input audio channels are downmixed to one single channel, denoted *sum signal*. From psychoacoustics it is known that the *level difference* (ICLD), *time difference* (ICTD), and *coherence* (ICC) between audio channels are relevant parameters for the perception of the auditory spatial image [2]. (The "IC" in the abbreviations denote *inter-channel*). Given its multi-channel input, the BCC encoder estimates ICLD, ICTD, and ICC between channel pairs as a function of frequency (i.e. in different subbands) and time and transmits these estimates to the decoder as side information. Given the sum signal, the decoder generates a multi-channel signal such that ICLD, ICTD, and ICC of the synthesized signal approximate those of the original multi-channel signal.

Figure 2 illustrates the scheme for ICTD, ICLD, and ICC synthesis for generating a multi-channel audio signal given the transmitted sum signal. By simply imposing different delays and applying different gain factors in subbands, a multi-channel signal with specific ICTD and ICLD can be generated [12] [13]. It is less obvious how to synthesize ICC. Several approaches for synthesizing ICC for parametric stereo coding were proposed previously [13] [26].

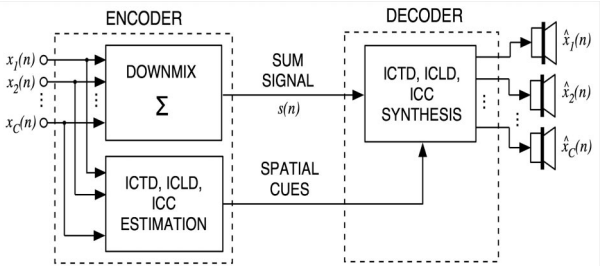

Figure 1: Generic BCC scheme. A number of input signals are downmixed to one channel and transmitted to the decoder together with side information.

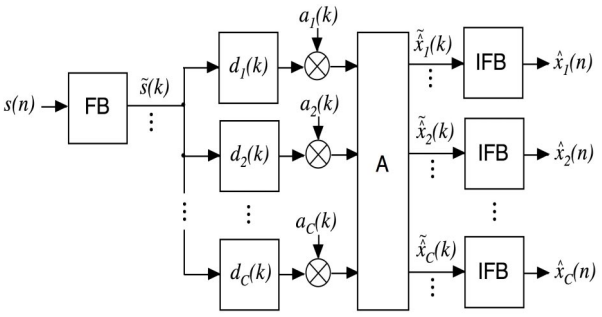

Figure 2: ICTD, ICLD, and ICC synthesis scheme. Processing is carried out individually for each frequency band. ICTD and ICLD are synthesized by imposing different delays and scale factors, respectively.

4. MP3 SURROUND CODING

The Binaural Cue Coding approach, as described in the preceding section, may be combined with any type of low bitrate audio coder to form an efficient system for transmission and storage of multi-channel sound providing two main functional aspects:

- Firstly (and probably most importantly), it enables a bitrate-efficient representation of multi-channel audio signals. Compared to a transmission of C discrete audio channel signals, only one audio signal has to be sent to the decoder together with a compact set of spatial side information which results in impressive bitrate savings. As an example, the usual 5 channel (3/2) format is reduced into a single sum audio

channel corresponding to an overall data reduction of about 80% (i.e. 4 out of 5 channels are dropped, neglecting the compact BCC side information).

- Secondly, the transmitted sum signal corresponds to a mono downmix of the multi-channel signal. For receivers that do not support multi-channel sound reproduction, listening to the transmitted sum signal is thus a valid method of presenting the audio material on low-profile monophonic reproduction setups. Conversely, BCC can therefore also be used to enhance existing services involving the delivery of monophonic audio material towards multi-channel audio.

The latter aspect of BCC can be regarded as a bridging function between monophonic and multi-channel representation. When looking at today's consumer electronics world, however, the dominant sound format is clearly 2-channel stereophony rather than a monophonic presentation. This motivates the use of a stereo sound representation as the basis for a BCC-type algorithm which then could scale up the information contained in these channels towards a multi-channel sound image. This is exactly the core idea behind the MP3 Surround approach, to be summarized as follows:

- Two audio channels are transmitted from the encoder to the decoder side forming a compatible stereo downmix of the multi-channel sound to be represented.

- A BCC-type algorithm produces multi-channel sound at the decoder end by making best possible use of the information contained in the transmitted stereo downmix signal.

- The compact spatial side information is embedded into the basic stereo MP3 bitstream in a compatible way, such that a standard MP3 decoder is not affected.

The proposed algorithm adds scalability to the basic BCC scheme in terms of transmitting more than one audio channel. Due to the increase in information available to the decoder, it is expected that the proposed scheme achieves higher quality than conventional BCC with one transmission channel. Another way of adding scalability to BCC is to use a regular multi-channel coder at lower frequencies and a mono audio coder with BCC at higher frequencies, as presented in [3].

4.1. Basic Scheme

Figure 3 illustrates the general structure of an MP3 Surround encoder for the case of encoding a 3/2 multi-channel signal (L, R, C, Ls, Rs). As a first step, a two-channel compatible stereo downmix (Lc, Rc) is generated from the multi-channel material by a downmixing processor or other suitable means. The resulting stereo signal is encoded by a conventional MP3 encoder in a fully standards compliant way. At the same time, a set of spatial parameters (ICLD, ICTD, ICC) are extracted from the multi-channel signal, possibly considering the stereo downmix signals. These spatial parameters are encoded and embedded as surround enhancement data into the ancillary data field of the MP3 bitstream within a suitable data container that unambiguously identifies the presence of such data for decoders with corresponding extended capabilities (i.e. MP3 Surround decoding). This mechanism of extending MPEG-1 Audio bitstream syntax has been used repeatedly in the past to "hide" various types of enhancement data in a backward compatible way, e.g. for embedding multi-channel audio extension [22] or bandwidth extension [30] data.

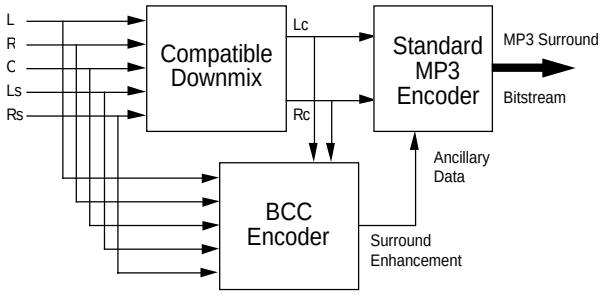

Figure 3: General structure of an MP3 Surround encoder (principle).

Figure 4 shows the decoder side of the transmission chain. The MP3 Surround bitstream is decoded into a compatible stereo downmix signal that is ready for presentation over a conventional 2-channel reproduction setup (speakers or headphones). Since this step is based on a fully compliant MPEG-1 Audio bitstream, any existing MP3 decoding device can perform this step and thus produce stereo output. MP3 Surround enabled decoders will furthermore detect the presence of the embedded surround enhancement information and, if available, expand the compatible stereo signal into a full multi-channel audio signal using a BCC-type decoder.

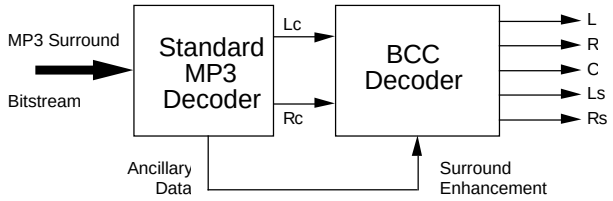

Figure 4: General structure of an MP3 Surround decoder (principle).

While the preceding example discussed the encoding/decoding of a 5 channel audio signal, other multi-channel configurations can be supported in the same way with this approach. This also includes the use of a subwoofer (LFE = "Low Frequency Enhancement") channel, as it is used frequently for the representation of movie sound (5.1 configuration).

4.2. Multi-channel vs. Stereo Representation

As can be seen from the previous section, the MP3 Surround process involves information from both a multi-channel version of the signal (for extraction of spatial parameters) and a stereo version (for actual compression and transmission). Therefore, there is a need to provide both versions of the audio item simultaneously for which a number of options are explored subsequently. The most common approach to obtaining a stereo version of a multi-channel signal is called *downmixing* and involves a linear combination of certain multi-channel signals to obtain the desired stereo signals. When downmixing multi-track/multi-channel sound material into a stereophonic representation, a number of considerations come into play which are motivated by both psychoacoustics and production practices. On one hand, it is desired to present all parts of the multi-channel sound image also to the listener of a stereo reproduction setup. On the other hand, it is known that – by collapsing front and back channels into the front-only stereo reproduction – the listener's ability to separate the sound components diminishes due to the lack of spatial separation between front and back sound sources. Consequently, sound sources from back channels are usually attenuated within a stereo mixdown in order to guarantee good audibility of the important front sound sources. In practice, there are different ways to produce corresponding stereo material for a multi-channel audio item:

Manual mix: In many cases, the sound engineer produces a "manual" downmix of the multi-channel sound sources into stereo using hand-optimized mixing parameters, thus preserving a maximum amount of artistic freedom. If further flexibility is desired, different recording methods may be used for the production of the stereo version (e.g. different microphone configurations).

Simple automatic downmixing: The most basic approach to create an automatic downmix from a given multi-channel recording is to use a fixed downmixing equation, such as the set of standard downmixing equations recommended by ITU-R for compatible 2-channel stereo reproduction of multi-channel signals [6] [31]. As an example, the well-known equations for the case of 3/2 multi-channel audio are:

$$L_c = t * (L + a*L_s + b*C)$$
$$R_c = t * (R + a*R_s + b*C)$$

where L_c and R_c denote the compatible stereo signals and t, a and b are constants. Usually, $a=b=1/\sqrt{2}$ are standard choices and t is chosen such that an appropriate output signal level is achieved while avoiding clipping. For material with a high level of ambient rear components, choosing a=0.5 is a common alternative.
Even though a fixed downmixing approach is clearly suboptimal compared to a dedicated manual stereo mix, it may in practice be sufficient for most applications.

Dynamic/advanced automatic downmixing: Over time, more advanced methods for automated multi-channel downmix have become available which take into consideration factors such as absolute source positioning, panning laws, the way sound sources were mixed into multi-channel signals and inter-channel phase relationships. Such advanced algorithms adapt their downmixing behavior to the processed material and may achieve a sonic quality that is comparable to that of a "manual downmix" [14]. Dynamic downmixing has also been applied to BCC for the case of one transmission channel [13] and can be used for generating the MP3 Surround compatible stereo downmix signals likewise.

The basic approach of MP3 Surround does not impose any restriction on what option for stereo downmixing has to be used. In fact, the downmixing process can be considered as a system component that is not necessarily part of the general coding scheme. This is illustrated in Figure 5 by showing a system that accepts both a 5-channel audio signal and a corresponding stereo version.

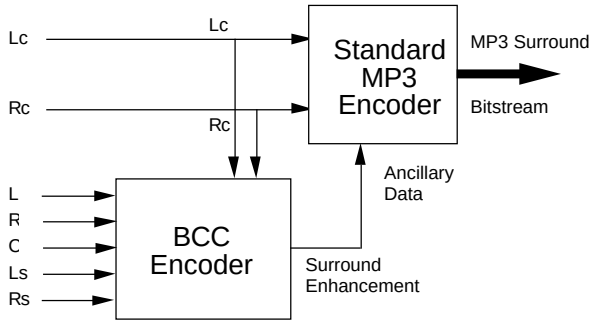

Figure 5: MP3 Surround encoding using an external downmix process.

Looking at the decoding side (Figure 4), it becomes clear that the transmitted stereo signals form the basis of the recreated multi-channel sound image. Henceforth, the sound components that are desired to be present in the multi-channel sound also need to be present in the stereo signal in an appropriate way to ensure satisfactory reproduction results. As a simple thought experiment, a solo instrument will not be reproduced properly in the multi-channel signal if omitted from the underlying stereo signal. Clearly, the use of well-behaved automatic downmixing processes guarantees that such pathological conditions do not occur. More work is needed to establish a set of minimum requirements for consistency between multi-channel and stereo signals that guarantee a proper decoded multi-channel sound image also for manual stereo downmixing.

In summary, different options exist for the production process of the stereo signal version that allow a trade-off between the delivered quality on the stereo and on the multi-channel side. These range from "use automatic stereo downmix optimized for best possible multi-channel reproduction" to "transmit best possible stereo quality, possibly at the expense of multi-channel sound quality".

5. QUALITY EVALUATION

In order to assess the subjective sound quality of MP3 Surround, a listening test was carried out to compare the performance of the proposed scheme with that of established multi-channel audio codecs. The general performance expectation of the MP3 Surround scheme can be derived from the following considerations:

- In MP3 Surround, the multi-channel sound image is multiplexed (downmixed) into a stereo signal and "unpacked" (expanded) again at the decoder side. This is analogous to the well-known formats for *matrixed surround*, such as Prologic, Logic 7 etc.

- Contrary to such matrixed surround formats, MP3 Surround makes use of some side information in order to reconstruct the multi-channel sound image at the decoder side. Ideally, the improvement due to the transmission of side information might deliver a system performance approaching that of a fully independent transmission of multi-channel material, i.e. a *discrete surround* format.

Consequently, both a common format for matrixed surround and a state-of-the-art discrete multi-channel codec were used for comparison with the MP3 Surround codec. For the matrixed surround format, Dolby ProLogic II [9] was chosen, a coder that is primarily intended for the transmission of multi channel audio over an analog stereo transmission line with the highest possible stereo compatibility. This format is widely used in current consumer electronic equipment. As a discrete multi-channel codec, MPEG-2/4 AAC was used at a bitrate of 320 kbit/s to produce near transparent quality. This coder was found to deliver EBU broadcast quality for 5-channel material at a bitrate of 256-320 kbit/s in formal verification tests [5]. The MP3 Surround coder was run at a total bitrate of 192 kbit/s (including side information) which is common for high quality stereo MP3.

In order to obtain both an absolute grade for each of the codecs as well as a consistent relative rating among them, the listening test method chosen closely resembles the ITU recommendation BS.1534 (MUSHRA) [7]. Several time-aligned audio signals were presented to the listener who performed on-the-fly switching between these signals using a keyboard and a screen. The signals included the original signal, which was labeled as "Reference", and several anonymized items, arranged at random. Using a simple graphic user interface software, the test subjects had to grade the basic audio quality of the anonymized items on a graphical scale with five equally sized regions labeled "Excellent", "Good", "Fair", "Poor" and "Bad". To check the listener's reliability and enable relating the ratings to results of other tests, a hidden reference (original) and an anchor were included in addition to the three coded/decoded items. The anchor consisted of a 3.5 kHz bandwidth reduced version of the reference, as is compulsory for

the MUSHRA test methodology. There was no limit of the number of repetitions the test subjects could listen to before they would deliver their ratings and proceed to the next test item. Due to its interactive nature, the test was taken by one subject at a time.

Eleven items were selected for the listening test, ten of them being commercial music of different styles (3 pop music, 3 jazz music, 4 classical music). The remaining item was created artificially and features a strong perceived dissimilarity between different channel groups (item "fountain": the sound of a fountain on the center channel, a piano on the front side channels and singing birds on the surround channels). The items were presented in an acoustically isolated listening lab equipped with high quality loudspeakers. Details on the equipment used and the listening test software can be found in the annex of this paper.

First experiments showed that it is very difficult to remember the surround image of the previous test signal by the time the next item is played. It is, therefore, very helpful for the listener to have the possibility of setting start and stop markers to allow looping at arbitrary positions, as is suggested in the BS.1116-1 [32] test specification and can be used for MUSHRA tests likewise. Furthermore, the possibility of instantaneous switching between different signals while they are playing contributes greatly to enhancing the sensitivity of the listening test. The switching process was implemented by a 100ms fade-out of one signal followed by a 100ms fade-in on the next signal, both using raised cosine windows. In this way, all transitions between signals sound identical. A software tool enabled these features and provided the possibility of rating and randomization of presentation order according to the MUSHRA specification [7]. Listeners were allowed to set the playback level according to their individual preference.

Eight of the ten subjects were expert listeners with years of experience in audio coding, while the other two were less experienced. Prior to the test phase the listeners were instructed to take into account both the faithful reproduction of the signal's spatial image and distortion by perceptual coding artifacts for their ratings.

Figure 6 shows the results of the listening test as mean rating and 95% confidence interval for individual test items together with their overall mean. As can be seen, the ratings of MPEG-2/4 AAC encoded items overlap with those of the hidden reference in their confidence interval for all items. This shows that the listeners were not able to distinguish between the coded/decoded items and the reference in a statistical sense. This outcome is consistent with previous characterizations of the codec at this bitrate.

The quality of the ProLogic II encoded/decoded signals was rated mostly in the "good" region of the grading scale. Although there is some degree of variability in the subjective ratings for this format, it is clearly visible that listeners are able to distinguish these signals from the original. Listeners frequently reported a change in both the general perception of the sound stage as well as the positioning of certain sound components. As a note to this result, it should be pointed out that in commercial applications of matrixed surround formats it is common to try to alleviate the limitations of the underlying format by optimizing the multi-channel mix at the time of its production for good reproduction after Prologic encoding/decoding [10].

The quality of the MP3 Surround encoded/decoded signals was mostly rated within the "excellent" range of the grading scale, and their confidence intervals overlap with those of the original signal for two out of the 11 test items. For the other nine cases, the listeners were able to distinguish statistically between the MP3 Surround coded version and the original. Test listeners reported no significant alterations of the spatial sound image.

Considering that the basic coding efficiency of the MP3 coder kernel is significantly lower than of MPEG-2/4 AAC and that MP3 Surround uses a far lower bitrate than the MPEG-2/4 AAC coder (192 kbit/s instead of 320 kbit/s), this test outcome can be seen as an excellent result. In summary, it appears that the overall subjective sound quality provided by the MP3 Surround system is much closer to that of a fully discrete multi-channel system than to a matrixed surround format even though MP3 Surround spends only a small fraction of its overall bitrate on encoding of the spatial information.

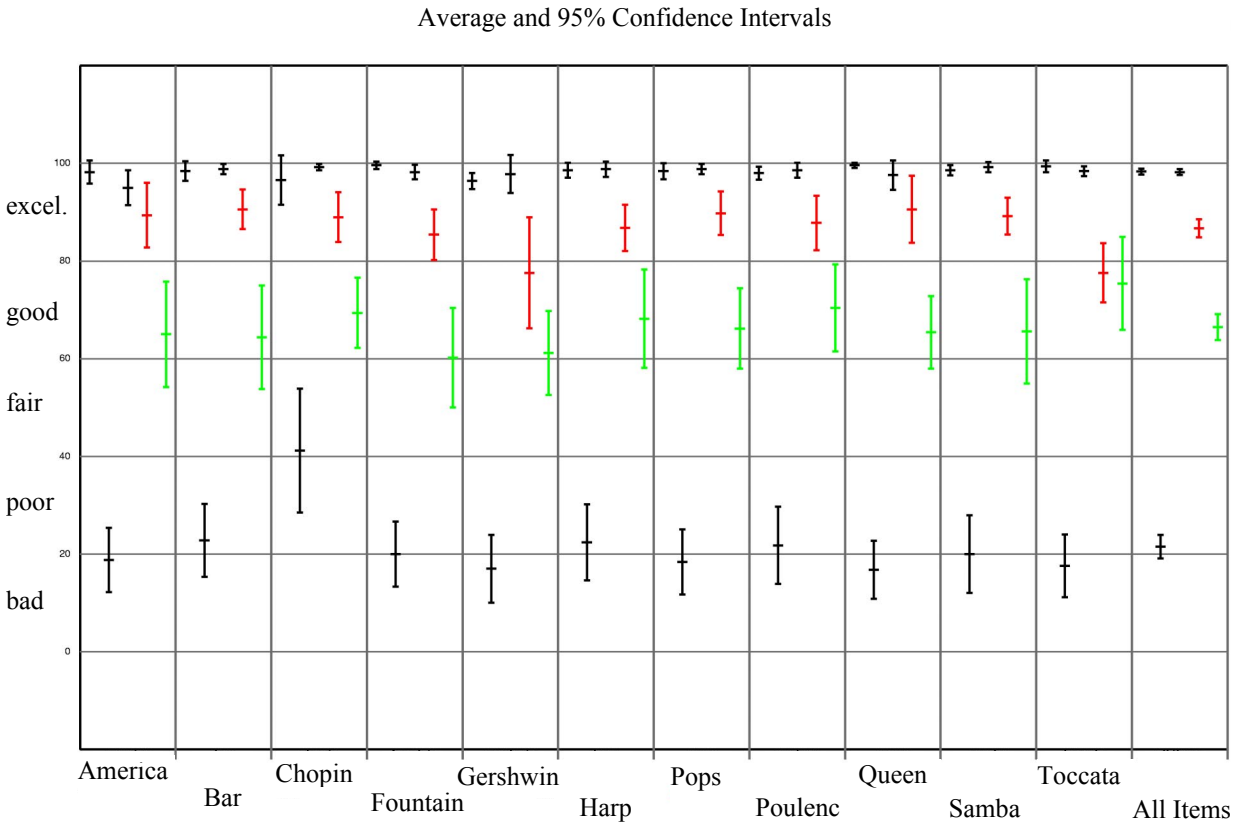

Figure 6: Results of the subjective listening test.

6. INTERNATIONAL STANDARDIZATION

As it may have become clear from the preceding discussions, the general idea of applying BCC-type processing to expand a compatible mono or stereo signal into a multi-channel sound image does not rely on the use of a particular type of audio coder. Due to the attractive properties of this approach, the MP3 Surround scheme can be considered merely the first commercial application of a new coding paradigm for multi-channel audio, and other schemes are likely to follow. Demonstrations of such combinations have been provided recently, including MPEG-2/4 AAC + BCC (67th MPEG meeting, Hawaii 2003, 5.1 multi-channel at ca. 140 kbit/s) and MPEG-4 High-Efficiency AAC + BCC (EBU workshop, Geneva 2/2004, 5.1 multi-channel at about 64 kbit/s).

In the area of international standardization, the ISO/MPEG Audio group has noted these recent advances and their market potential [19] and started a new work item with the working title *Spatial Audio Coding*. This process aims at complementing the existing MPEG-4 AAC-based general audio coding schemes with a tool for efficient and compatible representation of multi-channel audio. It addresses both technology that expands stereo signals into multi-channel sound (called "2-to-n" scheme) and the more traditional mono variant (called "1-to-n" scheme). The key requirements are [25]:

- Best possible approximation of original perceived multi-channel sound image

- Minimal bitrate overhead compared to conventional transmission of 1 or 2 audio channels

- Bitstream backward compatibility by using an MPEG-4 AAC codec for the audio transmission

- Backward compatibility of transmitted audio signal with existing mono or stereo reproduction systems, i.e. the transmitted audio channels shall represent a compatible (mono or stereo) audio signal representing all parts of the multi-channel sound image

As of the time of writing of this paper, the MPEG group has issued a "Call for Information" [24] to seek additional input on the requirements for this new work item and anticipates issuing a "Call for Proposals" (CfP) at its 68th meeting in March 2004 [25].

7. APPLICATIONS

Considering the general trend towards surround sound in consumer and professional audio, the following section illustrates some examples of applications that are enabled through spatial audio coding in general and MP3 Surround technology specifically, focusing on compatible multi-channel enhancements of existing services.

The vast majority of current audio-visual distribution infrastructure is tailored to delivering stereo rather than multi-channel audio, both in terms of available transmission bandwidth as well as the underlying technical system structure. Thus, in order to enable upgrading such distribution media to multi-channel audio, it is essential to deliver multi-channel audio at bitrates comparable to what is usually needed for the transmission of stereo which is one of the key features of the MP3 Surround/spatial audio coding approach.

- *Music download service*: Currently, a number of commercial music download services are available and working with considerable commercial success. Such services could be seamlessly extended to provide multi-channel enabled services while staying compatible for stereo users. On computers with 5.1 channel setups the MP3 Surround files are decoded in surround sound while on portable MP3 players the same files are played back as stereo music.

- *Streaming music service / Internet radio*: Many Internet radios are operating currently under severely constrained bandwidth conditions and, therefore, can offer only mono or stereo content. MP3 Surround could extend this to a full multi-channel service within the permissible range of bit-rates. Since efficiency is of paramount importance in this application, the compression aspect of MP3 Surround comes into play. As an example, representation of 5 presentation channels from two transmitted basis channels corresponds to a bit-rate saving of (5-2)/5 = 60% compared to full multi-channel coding, neglecting the small amount of surround enhancement side information.

- *Digital Audio Broadcasting*: Due to available channel capacity, the majority of existing or planned systems for digital broadcasting of audio content cannot provide multi-channel sound to the users. Adding this feature would, however, be a strong motivation for users to make the transition from their traditional FM receivers to the new digital systems. Again, compatibility with existing stereo reproduction setups is mandatory and inherent to the MP3 Surround / spatial audio coding concept.

- *Teleconferencing*: Although teleconferencing is becoming increasingly popular and important in today's global business world, most systems are still working with rather low bandwidth. A spatial audio coding approach can help expand the sound image to multi-channel sound and, thus, allow a better subjective separation and resolution of the audio contributions of each speaker in the teleconference.

- *Audio for Games*: Many personal computers have become "personal gaming engines" and are equipped with a 5.1 computer speaker setup. Synthesizing 5.1 sound from a backward compatible stereo sound basis allows for an efficient storage of multi-channel background music.

If storage efficiency is of chief importance for specific applications, a spatial audio coding approach based on a mono channel ("1-to-n scheme") can be used. For 5 channels, a bit-rate saving of 80% compared to a discrete multi-channel transmission is then achieved. For stereo-only applications, a "1-to-2" scheme may be desirable which corresponds to the case of *parametric stereo* as is addressed in the upcoming ISO/MPEG-4 Extension specification for use with a parametric audio coder [23].

8. CONCLUSIONS

This paper introduces an extension of the popular MP3 audio coding format towards the bitrate-efficient representation of multi-channel audio signals, most prominently the 5.0 and 5.1 channel configurations. This is achieved by extending a stereo MP3 scheme by means of Binaural Cue Coding (BCC) that serves as a spatial pre/post processor to the MP3 encoder/decoder chain. An important feature of the resulting format is full backward compatibility to existing MP3 decoders, which reproduce a complete stereo downmix of the multi-channel sound material. In order to guide the spatial decoding process in an MP3 Surround decoder, a small amount of spatial side information is hidden inside the MP3 bitstream in a compatible way within the ancillary data field.

Contrary to earlier approaches to parametric representation of stereo and multi-channel audio, the proposed algorithm transmits two (stereo compatible) signal channels rather than only a single channel. This is a significant and important extension of the general technical paradigm because it enables backward compatibility with the huge number of stereo reproduction setups and media currently in existence. The MP3 Surround scheme and the underlying general ideas pave the way for using multi-channel sound for a number of attractive applications which were inconceivable in the past, such as multi-channel Internet radio and music download services.

Results of first informal listening tests indicate that the proposed scheme provides an excellent combination of low bitrate and sound quality which is significantly better than that of matrixed surround formats. The idea of expanding backward compatible monophonic and stereophonic sound signals has been embraced by the MPEG standardization group. The new work item "Spatial Audio Coding" was initiated to undertake further development of this concept in the context of MPEG-4 Audio. Over time we will see further technical development of such algorithms and combinations with more powerful audio coders which will bring the vision of multi-channel audio at very low bitrates (<64 kbit/s) into reality.

9. ANNEX

The subjective listening test was conducted in an acoustically isolated listening lab that is designed to permit high-quality listening tests conforming to the BS.1116 [32] test methodology for high-quality audio listening tests. The room dimension was 5.5m x 7.1m x 2.4m. All test signals were presented on 5 Geithain RL 901 active studio loudspeakers and a Geithain TT 920 active subwoofer driven by Bryston pre-amplifiers. Playback was controlled from a 2 GHz Linux computer with an RME Hammerfall digital sound output interface connected to Lake People DAC F20 D/A converters.

For the generation of the AAC bitstreams the Fraunhofer IIS professional MPEG-4 AAC Low Complexity Profile encoder was used. The encoder was initialized with default psychoacoustic parameters at a bitrate of 320 kbps.

The Dolby ProLogic II signals were generated with the *SurCode for Dolby ProLogic II Version 2.0.3* software by Minnetonka Audio Software using the default settings. In order to guarantee perceptually equal loudness of the Dolby ProLogic II signals, a loudness alignment procedure was necessary. The amplification factor for each ProLogic II decoded signal was determined by adjusting its amplification in steps of 1 dB until there was no perceptual difference in loudness compared to the original signal when switching between both signals. The procedure was repeated for several listeners. This resulted in an equal amplification of 3 dB for all ProLogic II decoded test signals. The signals had sufficient headroom to enable this gain change without introducing clipping.

The original items were obtained partly from special audio test compilations and partly recorded from the analog output of a Panasonic DVD A7 DVD-Audio player using an RME ADI 8 DS A/D converter.

Item	Format	Title	Artist	Album	Style
America	5.1	Head and Heart	America	Homecoming DVD Audio	Pop
Bar	5.1	My Love Is Here	M. Matthews	DVD Audio (Panasonic/Technics)	Jazz
Chopin	5.0	Piano Sonata Nr. 3 Finale	n/a	Multi-Channel Demo Material (Denon Electronic GmbH)	Classic
Fountain	5.0	Fountain Music	-	MPEG-2 Multi-Channel Verification Test Material	n/a
Gershwin	5.0	I Got Rhythm	n/a	Multi-Channel Demo Material (Denon Electronic GmbH)	Jazz
Harp	5.0	G.F. Händel: Concerto For Harp And Strings	New York Symphonic Ensemble	DVD Audio (Panasonic/Technics)	Classic
Pops	5.0	n/a	n/a	n/a	Pop
Poulenc	5.0	Concerto In G Minor For Organ, String And Timpani	n/a	Multi-Channel Demo Material (Denon Electronic GmbH)	Classic
Queen	5.1	You´re My Best Friend	Queen	A Night At The Opera DVD Audio	Pop
Samba	5.0	One Note Samba	Hamamura Quintett	Audionet STANDARDS No. 1: RETOLD, DVD Audio	Jazz
Toccata	5.0	J.S. Bach: Toccata And Fugue In D Minor	Mika Oi	DVD Audio (Panasonic/Technics)	Classic

Table 1: Description of the listening test items.

10. REFERENCES

[1] F. Baumgarte and C. Faller: “Why Binaural Cue Coding is better than Intensity Stereo,” 112th AES Convention, Munich 2002. Preprint 5575

[2] F. Baumgarte and C. Faller: “Binaural Cue Coding - Part I: Psychoacoustic fundamentals and design principles,” IEEE Trans. on Speech and Audio Proc., vol. 11, no. 6, November 2003

[3] F. Baumgarte, C. Faller, P. Kroon: “Audio Coder Enhancement using Scalable Binaural Cue Coding with Equalized Mixing,” 116th AES Convention, Berlin 2004

[4] J. Blauert, “Spatial Hearing: The Psychophysics of Human Sound Localization”, revised edition, MIT Press, 1997

[5] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, Oikawa: “ISO/IEC MPEG-2 Advanced Audio Coding”, Journal of the AES, Vol. 45, No. 10, October 1997, pp. 789-814

[6] ITU-R Recommendation BS.775-1, “Multi-channel Stereophonic Sound System with or without Accompanying Picture”, International Telecommunications Union, Geneva, Switzerland, 1992-1994

[7] ITU-R Recommendation BS.1534-1, “Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)”, International

Telecommunications Union, Geneva, Switzerland, 2001

[8] M. Davis, "The AC-3 Multichannel Coder", 95th AES Convention, New York, 1993, Preprint 3774

[9] Dolby Publication, Roger Dressler: "Dolby Surround Prologic Decoder – Principles of Operation", http://www.dolby.com/tech/whtppr.html

[10] Dolby Publication, Jim Hilson: "Mixing with Dolby Pro Logic II Technology", http://www.dolby.com/tech/PLII.Mixing.JimHilson.html

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

[12] C. Faller and F. Baumgarte, "Binaural Cue Coding applied to stereo and multi-channel audio compression," 112th AES Convention, Munich 2002. Preprint 5574

[13] C. Faller and F. Baumgarte, "Binaural Cue Coding - Part II: Schemes and applications," IEEE Trans. on Speech and Audio Proc., vol. 11, no. 6, Nov. 2003

[14] D. Griesinger: "Surround from stereo", Workshop #12, 115th AES Convention, New York, 2003

[15] J. Herre, K. Brandenburg, D. Lederer: "Intensity Stereo Coding", 96th AES Convention, Amsterdam 1994, Preprint 3799

[16] J. D. Johnston, A. J. Ferreira, "Sum-Difference Stereo Transform Coding", IEEE ICASSP 1992, pp. 569-571

[17] J. D. Johnston, J. Herre, M. Davis, U. Gbur: "MPEG-2 NBC Audio - Stereo and Multichannel Coding Methods", 101st AES Convention, Los Angeles 1996, Preprint 4383

[18] M. Lutzky, M. Weishart, J. Hilpert, B. Grill, H. Gernhardt, M. Haertl: "MP3 in MPEG-4", 115th AES Convention, New York 2003, Preprint 5870

[19] ISO/IEC JTC1/SC29/WG11 (MPEG), Document M10378, "Spatial Audio Coding: Market Context and Requirements", Hawaii 2003

[20] ISO/IEC JTC1/SC29/WG11 MPEG, International Standard ISO/IEC 13818-7 "Generic Coding of Moving Pictures and Associated Audio: Advanced Audio Coding", 1997

[21] ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC 11172-3 "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s", 1992

[22] ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC 13818-3 "Generic Coding of Moving Pictures and Associated Audio: Audio", 1994

[23] ISO/IEC JTC1/SC29/WG11 (MPEG), Document N6130, "Text of ISO/IEC 14496-3:2001/FDAM2", Hawaii 2003

[24] ISO/IEC JTC1/SC29/WG11 (MPEG), Document M10378, "Call for Information on Spatial Audio Coding", Hawaii 2003

[25] ISO/IEC JTC1/SC29/WG11 (MPEG), Document M10378, "Draft Call for Proposals on Spatial Audio Coding", Hawaii 2003

[26] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," 114th AES Convention, Amsterdam 2003, Preprint 5852

[27] G. Stoll, G. Theile, S. Nielsen, A. Silzle, M. Link, R. Sedlmayer, A. Breford, "Extension of ISO/MPEG-Audio Layer II to Multi-Channel Coding: The Future Standard for Broadcasting, Telecommunication, and Multimedia Applications", 94th AES Convention, Berlin 1994, Preprint 3550

[28] AES Technical Committee of Coding of Audio Signals: "Perceptual Audio Coders: What to listen for", CD-ROM with tutorial information and audio examples, AES publications

[29] R.G. v.d.Waal, R.N.J. Veldhuis, "Subband Coding of Stereophonic Digital Audio Signals", IEEE ICASSP 1991, pp 3601-3604.

[30] T. Ziegler, A. Ehret, P. Ekstrand, M. Lutzky, "Enhancing MP3 with SBR: Features and Capabilities of the New MP3PRO Algorithm", 112th AES Convention, Munich 2002, Preprint 5560

[31] S. K. Zielinski, F. Rumsey: "Effects of Down-Mix Algorithms on Quality of Surround Sound", Journal of the AES, pp. 790, September 2003

[32] ITU-R Recommendation BS.1116-1 "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems", International Telecommunications Union, Geneva Swizerland, 1994-1997