Fraunhofer Institut
Integrierte Schaltungen

An Introduction To MP3 Surround

Jürgen Herre, Johannes Hilpert, Christian Ertel, Andreas Hoelzer, Claus Spenger, Sascha Disch, and Karsten Linzmeier

Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany

Christof Faller

Agere Systems, Allentown, PA 18109, USA

ABSTRACT

This paper presents a novel extension of the popular MP3 compression format which extends current MP3 capabilities towards the efficient and compatible representation of multi-channel audio, including the widely used 5.1 surround sound. The new format features complete backward compatibility with existing stereo MP3 decoders at bit-rates comparable to those currently used to encode stereo material. The paper discusses the underlying advanced technology and presents results of subjective listening tests. Finally several applications in multi-channel sound enabled by the MP3 Surround format are introduced.

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.

Formally speaking, the name MP3 refers to the Layer 3 coding scheme of the ISO/MPEG-1 and ISO/MPEG-2 Audio specifications [1] [2] that were finalized in 1992 and 1994, respectively[1]. Compliant to these standards, MP3 capability usually supports the 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, the name MP3 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. Although market penetration in Europe is lagging behind the US, the trend towards owning a "home theater" setup is clearly detectable.

One 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.

This raises the question about a multi-channel MP3 audio format that could serve the MP3 user community for 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. To accommodate a smooth transition towards multi-channel transmission and reproduction technology, an ideal advanced MP3 coding format needs to support equally

[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.

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

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.

The subsequent section gives 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, the paper 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. This will be rounded up with a discussion of several applications in multi-channel sound enabled by the MP3 Surround format.

2. BINAURAL CUE CODING (BCC)

Binaural Cue Coding (BCC) [3] [4] [5] is a general concept for parametric representation of spatial audio, delivering multi-channel output (with an arbitrary number of channels) from a single audio channel plus some side information. Figure 1 illustrates this concept. Several input audio channels are combined into a single output ("sum") signal by a downmix process. In parallel, the most salient inter-channel cues describing the multi-channel sound image are extracted from the input channels and coded compactly as BCC side information. Both sum signal and side information are then transmitted to the receiver side, possibly using an appropriate low bitrate audio coding scheme for coding the sum signal. Finally, the BCC decoder generates a multi-channel output signal from the transmitted sum signal and the spatial cue information by re-synthesizing channel output signals which carry the relevant inter-channel cues, such as Inter-channel Time Difference (ICTD, Inter-channel Level Difference (ICLD) and Inter-channel Coherence (ICC).

Figure 2 shows the general structure of a BCC synthesis scheme. The transmitted ("sum") signal is mapped to a spectral representation by a filterbank. For each output channel to be generated, individual time delays and level differences are imposed on the spectral coefficients, followed by a coherence synthesis process which re-introduces the most relevant aspects of coherence / (de)correlation between the synthesized audio channels. Finally, all synthesized output channels are converted back into a time domain representation by inverse filterbanks.

For a more detailed description of the BCC approach, the reader should refer to [6].

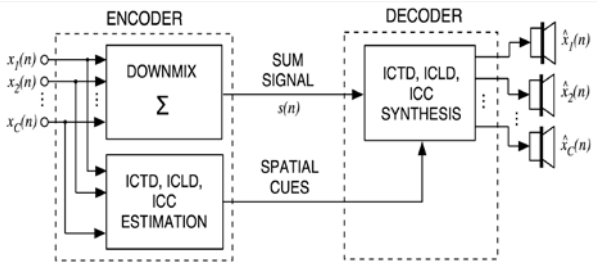

Figure 1: *Principle of Binaural Cue Coding.*

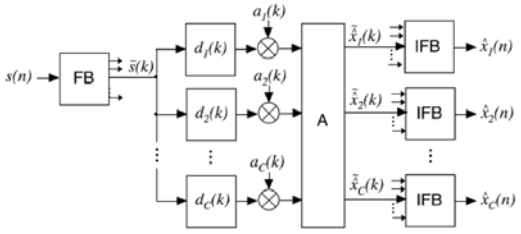

Figure 2: *Binaural Cue Coding Synthesis (Principle).*

Binaural Cue Coding exhibits a number of marked advantages of simple intensity stereo coding by being able to recreate output signals with time differences and a wide sound stage consisting of uncorrelated components. Consequently, BCC can be applied to the full audio frequency range without unacceptable signal distortion. Conversely, the traditional intensity stereo processing can be interpreted as a BCC type processing which is limited to ILD synthesis only and is subject to imperfect reconstruction due to the use of critically subsampled coder filterbanks.

An alternative type of BCC has also been used to enable bitrate-efficient transmission and flexible rendering of multiple audio sources which are represented by a single transmitted audio channel plus some cue side information [3] [4] [5].

3. 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 [7].

3.1. 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).

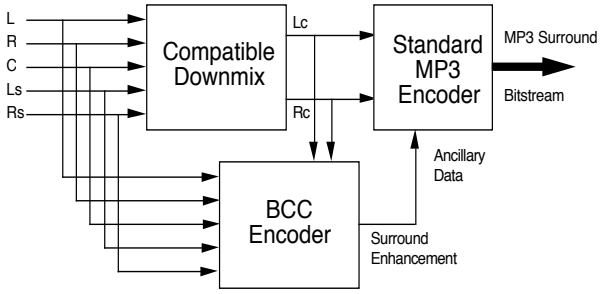

Figure 3: General structure of an MP3 Surround encoder (principle).

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.

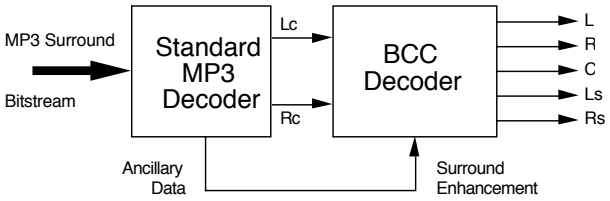

Figure 4: General structure of an MP3 Surround decoder (principle).

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).

3.2. 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 [8] [9]. 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" [10]. Dynamic downmixing has also been applied to BCC for the case of one transmission channel [11] 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.

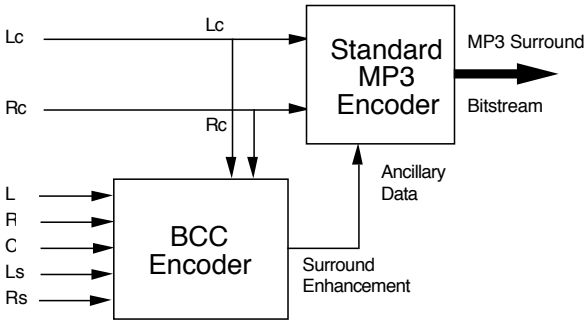

Figure 5: MP3 Surround encoding using an external downmix process.

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. On the other hand, a number of initial experiments indicate the viability of using manual downmix signals in the context of MP3 Surround. More in-depth investigations into this topic are carried out currently to establish a set of minimum requirements for consistency between multi-channel and stereo signals that guarantee a proper decoded multi-channel sound image for this approach.

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".

4. 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.

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 [12] was chosen, a coder that is primarily intended for the transmission of multi channel audio over an analog stereo transmission line while maintaining stereo compatibility. As a discrete multi-channel codec, MPEG-2/4 AAC was used at a bitrate of 320 kbit/s to produce near transparent quality. 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) [13]. 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 graphic user interface based software, the listeners 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.

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 [14] 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. 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.

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. The test participants 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.

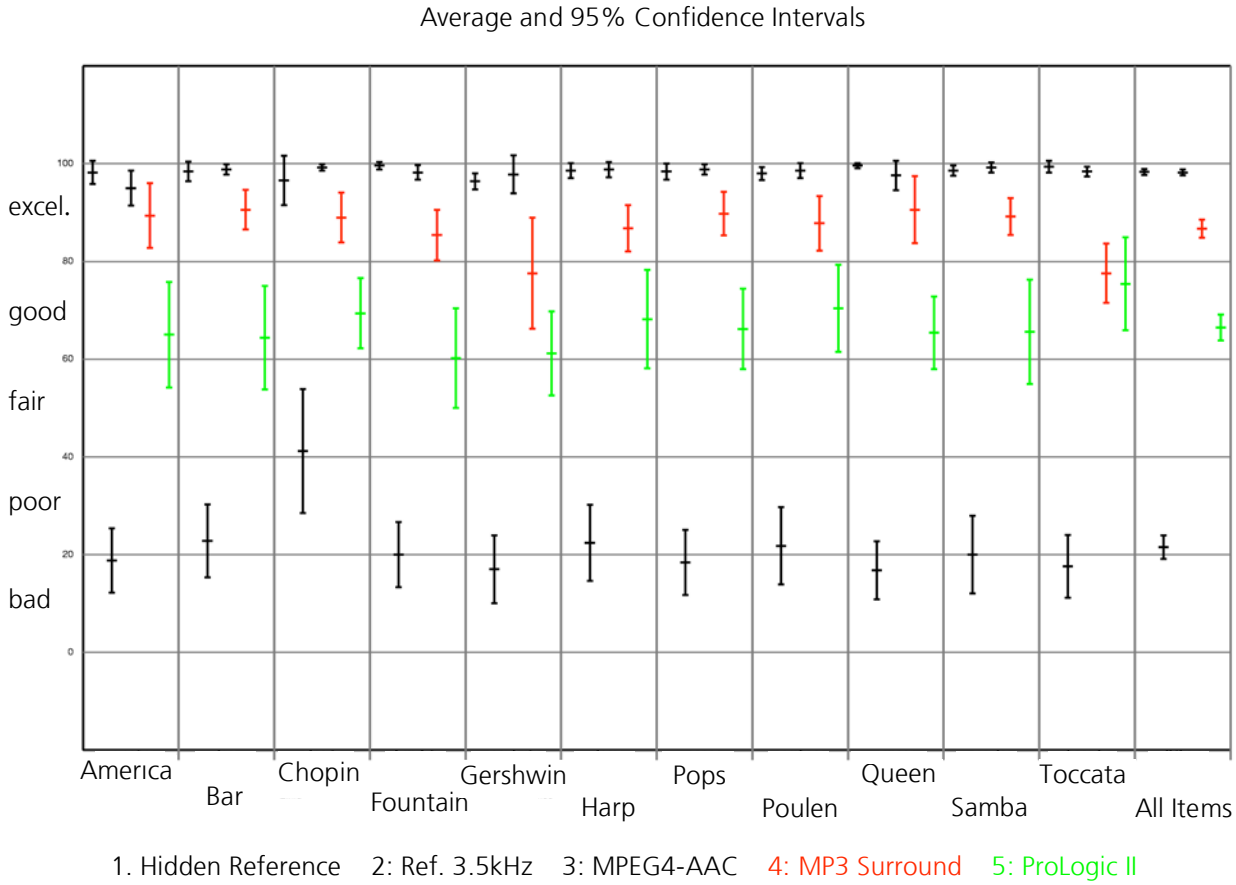

Figure 6: Results of the subjective listening test

5. 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 MP3 Surround technology, focusing on compatible multi-channel enhancements of existing services. The key feature of MP3 Surround in this context is its ability to deliver multi-channel audio at bitrates comparable to what is usually needed for the transmission of stereo.

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 radio services are currently operating under severely constrained bandwidth conditions and, therefore, can offer only mono or stereo content. MP3 Surround could extend such mono or stereo services 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.

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.

6. 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.

7. REFERENCES

[1] 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

[2] ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC 13818-3

"Generic Coding of Moving Pictures and Associated Audio: Audio", 1994

[3] C. Faller, F. Baumgarte: "Efficient Representation of Spatial Audio Using Perceptual Parametrization", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York 2001

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

[5] 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

[6] C. Faller: "Parametric Coding of Spatial Audio", 7th International Conference on Audio Effects (DAFX-04), Naples, Italy, October 2004

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

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

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

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

[11] 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

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

[13] ITU-R Recommendation BS.1534-1, "Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)", International Telecommunications Union, Geneva, Switzerland, 2001

[14] 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