A Backward Compatible MultiChannel Audio Compression Method

Xuefei Gao, Guo Yang, Jing Wang, Xiang Xie, Jingming Kuang
School of Information and Electronics, Beijing Institute of Technology
Beijing, 100081, China
e-mail:gaoxuefeibit@163.com, mryangguo@163.com, wangjing@bit.edu.cn, xiexiang@bit.edu.cn

***Abstract*—This paper proposes a backward-compatible multichannel audio codec based on downmix and upmix operation. The codec represents a multichannel audio input signal with downmixed mono signal and spatial parametric data. The encoding method consists of three parts: spatial temporal analysis of audio signal, compressing multi-channel audio into mono audio and encoding mono signals. The proposed codec combines high audio quality and low parameter coding rate and the method is simpler and more effective than the conventional methods. With this method, it's possible to transmit or store multi-channel audio signals as mono audio signals.**

Keywords-multichannel; downmix; audio; compression

I. INTRODUCTION

Digital audio has been popular for decades due to its convenience in acquisition, manipulation and distribution. With the rising popularity of the Internet, the advance of digital audio compression techniques such as MP3, AAC increases the use of digital audio files. As the need for high quality of digital audio increases, so does the need for more efficient audio compression technology. Multi-channel audio provides more realistic experiences which stereo and mono audio signals may fail to provide. But the data size of multi-channel audio signals is much higher compared to that of stereo audio signals. Furthermore, other applications such as 3D game, 3D movie and virtual reality using multi-channel audio signals also need efficient compression methods. As digital audio broadcasting services become popular, it is also necessary to efficiently compress multi-channel audio data using limited bandwidth.

A number of multi-channel audio compression techniques such as SDDS, Dolby digital [1], and DTS have been proposed. However, they have been developed mainly for maximum compression efficiency instead of backward compatibility. Although they are widely used in movie industries and high-end home applications, recent effort [3] (MPEG Surround) has been made for satisfying both high compression ratio and backward compatibility.

We propose in this paper a new approach of multichannel audio coding method with encoding sound source information. This paper is organized as follows: Section 2 describes the proposed method. Section 3 describes the experimental results. Finally, conclusions are presented in section 4.

II. THE PROPOSED METHOD

The proposed method consists of three parts. The first part analyzes the sound source location to reduce signal redundancy in the space and the time. The second part generates mono channel audio signals from multi-channel audio signals using the sound source location information. And the third part encodes the mono channel audio signals along with the sound source location information. The decoding process is the reversal process of the encoder. The proposed method is illustrated in Fig. 1.

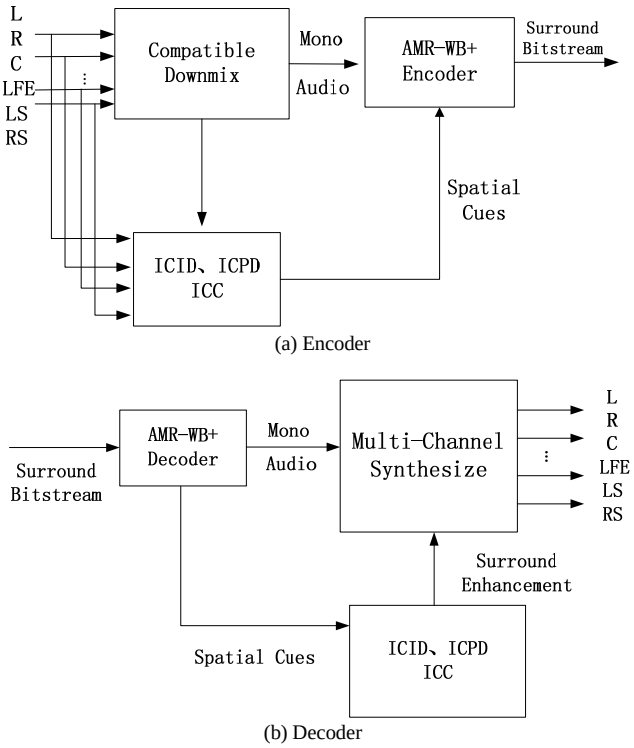


Figure 1. Overview of the proposed coding method.

The mono audio obtained from the original multi-channel audio may be combined with any type of low bitrates 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 an efficient representation of multi-channel audio signals.

Published by Atlantis Press, Paris, France.

Compared to a transmission of six 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 bit rate savings [6-7]. 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 spatial 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.

A. Basic Scheme

Figure 1(a) illustrates the general structure of the proposed surround encoder for the case of encoding a 3/2 multichannel signal (L, R, C, Ls, Rs). As a first step, a mono downmix signal is generated from the multi-channel material by a downmixing processor or other suitable means. The resulting mono signal is encoded by a conventional AMR-WB+ encoder. Meanwhile, a set of spatial parameters (ICID, ICPD, ICC) are extracted from the multi-channel signal. These spatial parameters are encoded and embedded as surround enhancement data into the ancillary data field of the AMR-WB+ bitstream within a suitable data container that unambiguously identifies the presence of such data for decoders with corresponding extended capabilities.

Figure 1(b) shows the decoder side of the transmission chain. The encoded surround bitstream is decoded into a compatible mono downmix signal that is ready for presentation over a conventional mono reproduction setup. Since this step is based on a fully compliant AMR-WB+ audio bitstream, any existing AMR-WB+ decoding device can perform this step and thus produce mono output. This proposed method enabled decoders will furthermore detect the presence of the embedded surround enhancement information and, if available, expand the compatible mono signal into a full multi-channel audio signal using the spatial cues.

While the preceding example discussed the encoding and 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 (Low Frequency Enhancement) channel, as it is used frequently for the representation of movie sound.

B. Coder Implementation

Since the spatial parameters are estimated (at the encoder side) and applied (at the decoder side) as a function of time and frequency, both the encoder and decoder require a transform or filter bank that generates individual time/frequency tiles. The frequency resolution of this stage should be nonuniform according to the frequency resolution of the human auditory system. Furthermore, the temporal resolution should generally be fairly low reflecting the concept of binaural sluggishness, except in the case of transients, where the precedence effect dictates a time resolution of only a few milliseconds.

Furthermore, the transform or filter bank should be oversampled, since time and frequency dependent changes will be made to the signals which would lead to audible aliasing distortion in a critically-sampled system. Finally, a complex valued transform or filter bank is preferred to enable easy estimation and modification of (cross-channel) phase difference information. A process that meets these requirements is a variable segmentation process with temporally overlapping segments, followed by forward and inverse FFTs. Complex-modulated filter banks can be employed as a low-complexity alternative [8].

Each segment is transformed to the frequency domain using an FFT with length of N (N = 1920 for a sampling rate frequency of 48 kHz). The frequency domain signals X1[k], X2[k]…X6[K] (k = [0, 1… N/2]) are divided into non-overlapping sub-bands. The frequency bands are formed in such a way that each band has a bandwidth, BW (in Hz), which is approximately equal to the equivalent rectangular bandwidth (ERB) [8], following

$$BW = 24.7(0.00437f + 1) \tag{1}$$

with f the (center) frequency given in Hz.

For each frequency band b, three spatial parameters are computed. The first parameter is the inter-channel intensity difference (ICID[b]), defined as the logarithm of the power ratio of corresponding subbands from the input signals:

$$ICID[b] = 10\log_{10} \frac{\sum_{K=K_b}^{K_{b+1}-1} X_i[k]X_i^*[k]}{\sum_{K=K_b}^{K_{b+1}-1} X[k]X^*[k]} \tag{2}$$

where $*$ denotes complex conjugation. The second parameter is the relative phase rotation. The phase rotation aims at optimal (in terms of correlation) phase alignment between the two signals [5]. This parameter is denoted by the inter-channel phase difference (ICPD[b]) and is obtained as follows:

$$ICPD[b] = \angle\left(\sum_{K=K_b}^{K_{b+1}-1} X_i[k]X^*[k] \right) \tag{3}$$

Using the ICPD as specified in (3), delays between the input signals which are represented as a constant phase difference in each analysis frequency band, hence result in a fractional delay. Thus, within each analysis band, the constant slope of phase with frequency is modeled by a constant phase difference per band, which is a somewhat limited model for the delay. On the other hand, constant phase differences across the input signals are described accurately, which is in turn not possible if an ICTD parameter (i.e., a parameterized slope of phase with frequency) would have been used.

The third parameter is the inter-channel coherence (ICC[b]), which is, in our context, defined as the normalized cross-correlation coefficient after phase alignment according

to the ICPD. The coherence is derived from the cross-spectrum in the following way:

$$ICC[b] = \frac{\left|\sum_{K=K_b}^{K_{b+1}-1} X_i[k]X^*[k]\right|}{\sqrt{\left(\sum_{K=K_b}^{K_{b+1}-1} X_i[k]X_i^*[k]\right)\left(\sum_{K=K_b}^{K_{b+1}-1} X[k]X^*[k]\right)}} \quad (4)$$

A suitable mono signal S[k] is obtained by a linear combination of the input signals X1[k] , X2[k]…X6[k]:

$$S[k] = w_1 X_1[k] + w_2 X_2[k] + \cdots + w_6 X_6[k] \quad (5)$$

where w1, w2,…,w6 are weights that determine the relative amount of X1, X2,…, X6 in the mono output signal. A downmix signal that is created using fixed weights however bears the risk that the power of the downmix signal strongly depends on the cross correlation of the input signals. To circumvent signal loss and signal coloration due to time and frequency dependent cross correlations, the weights w1,w2…w6 are complex-valued, to prevent phase cancellation, and varying in magnitude, to ensure overall power preservation. Specific details of the downmix procedure are however not included in the scope of this paper.

The ICID, ICPD, and ICC parameters are quantized according to perceptual criteria. The quantization process aims at introducing quantization errors which are just inaudible. For the ICID, this constraint requires a nonlinear quantizer, or nonlinearly spaced ICID values given the fact that the sensitivity for changes in ICID depends on the reference ICID. The vector ICIDs contains the possible discrete ICID values that are available for the quantizer. Each element in ICIDs represents a single quantization level for the ICID parameter and is indicated by

$$ICID_q[i] \ (i = 0, \cdots, 15) \quad (6)$$

$$ICIDs = \left[ICID_q[0], ICID_q[1], \cdots, ICID_q[15]\right] = \begin{bmatrix} -20, -15, -11, -8, -5, -3, -1, 0, 1, \\ 3, 5, 8, 11, 15, 20 \end{bmatrix} \quad (7)$$

For the ICPD parameter, the vector ICPDs represents the available quantized ICPD values:

$$ICPDs = \left[ICPD_q[0], ICPD_q[1], \cdots, ICPD_q[7]\right] = \left[0, \frac{\pi}{4}, \frac{2\pi}{4}, \frac{3\pi}{4}, \frac{4\pi}{4}, \frac{5\pi}{4}, \frac{6\pi}{4}, \frac{7\pi}{4}\right] \quad (8)$$

Finally, the repertoire for ICC, represented in the vector ICCs, is given by:

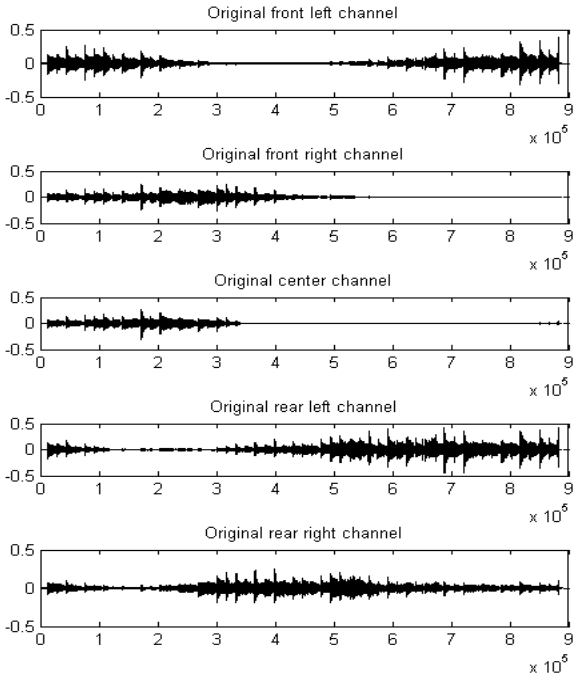
$$ICCs = \left[ICC_q[0], ICC_q[1], \cdots, ICC_q[7]\right] = \begin{bmatrix} 1, 0.937, 0.84118, 0.60092, \\ 0.36764, 0, -0.589, -1 \end{bmatrix} \quad (9)$$

C. Coding of Low frequency Effects (LFE) Audio Channels

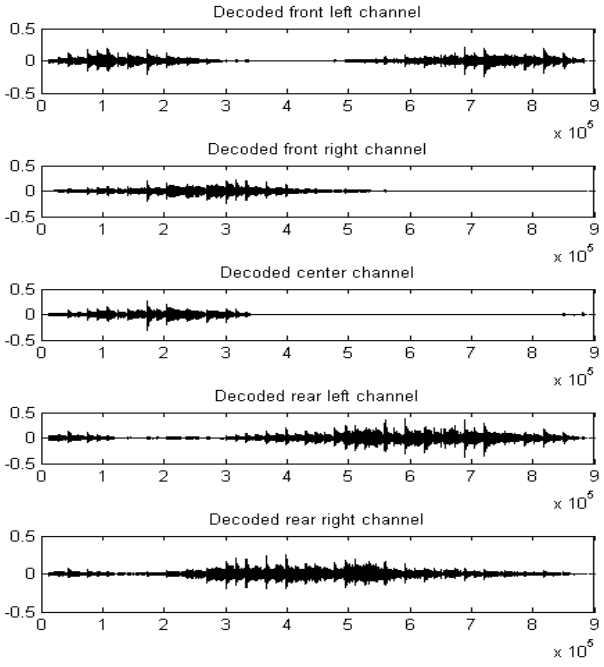
An LFE channel, as defined for the 5.1 standard, contains only frequencies up to 120 Hz. The same principles are applicable to other surround formats. At frequencies below 120 Hz, six-channel signal is applied, i.e. all six channels including the LFE channel are coded. At frequencies above 120 Hz, five-channel signal is applied, i.e. all channels except the LFE channel. The LFE channel is not considered at higher frequencies since it does not contain any signal energy there. This is implemented specifically by using a filterbank with a lowest sub-band covering 0–120 Hz. For this lowest sub-band the LFE channel is considered and for all other sub-bands the LFE channel is ignored.

III. Experiment Results

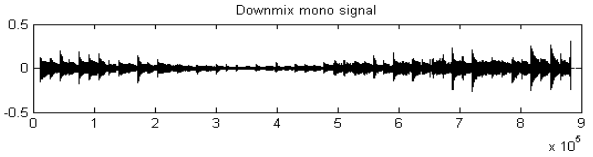
Test audio signals are extracted from a variety of audio sources and the number of test audio signals is 20. In the experiments, we compressed 5 channel audio signals into mono signals and reconstructed 5 channel audio signals from the compressed mono signals. Fig. 2 shows the waveforms of the original audio, the decoded audio and the encoded mono audio signals. It can be seen that the reconstructed multi-channel audio signals are very similar to the original multi-channel audio signals. The average correlation between the original audio and the decoded audio signals of the 20 test audio signals are shown in Table I. The center and front right channels are well preserved while the front left and surround left channels show some degradation. The center channel shows the best performance.



(a) original 5-channel audio signals



(b) decoded 5-channel audio signals



(c) downmix mono audio signals

Figure 2. Sample waveforms . (a) original 5-channel audio signals, (b) decoded 5-channel audio signals, (c) downmix mono audio signals.

TABLE I. AVERAGE CORRELATION COEFFICIENTS FOREEACH CHANNEL.

Front Left	**Front Right**	**Center**	**Surround Left**	**Surround Right**
0.8386	0.8829	0.8952	0.8292	0.7576

IV. CONCLUSIONS

This paper proposes a multi-channel audio coding method which uses spatial cues information to compress multichannel audio data as mono data. The proposed method would make it possible to substantially reduce the data size of multi-channel audio signals. Consequently, it will be possible to transmit or store multi-channel audio signals consuming the bandwidth of mono audio signals.

ACKNOWLEDGMENT

The work was supported by the Nation Natural Science Foundation: The Ultra-realistic Acoustic Interactive Communication of The Next Generation Internet, under the project number 11161140319. Also we appreciate the Special Issue of National Science and Technology: A New Generation Broadband Wireless Mobile Communication Network, under the project number 2010ZX03004-003-01.

REFERENCES

[1] Todd, C. C., Davidson, G. A., Davis, M. F., Fielder, L. D.,Link, B. D., and S. Vernon, AC-3: Flexible Perceptual Coding for Audio Transmission and Storage, 96th AES Convention, preprint 3796, 1994.

[2] K. Akagiri, M. Katakura, H. Yamauchi, E. Saito, M. Kohut, M. Nishiguchi, and K. Tsutsui, Sony systems, in Vijay K. Madisetti. ,and Douglas B Williams, The digital signal processing handbook. CRC Press, 1997.

[3] J. Herre, C. Faller, C. Ertel, J. Hilpert, A Hoelzer, C. Spenger: "MP3 Surround: Efficient and Compatible Coding of MUulti-Channel Audio", 116th AES Convention, Berlin 2004, Preprint 6049.

[4] J.D. Johnston: "Perceptual Coding of Wideband Stereo Signals", Proc. Of the ICASSP 1990.

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

[6] C. Faller: "Parametric Coding of Spatial Audio",Swiss Federal Institute of Technology Lausanne(EPFL), Ph.D. Thesis No.3062,2004.

[7] C. Faller: "Coding of Spatial Audio Compatible with Different Playback Formats", 117th AES Convention, San Francisco 2004.

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

[9] ITU-R, Multichannel stereophonic sound system with and without accompanying picture, Recommendation ITU-R BS.775-1, 1994.