A STEREO TO MONO DOWMIXING SCHEME FOR MPEG-4 PARAMETRIC STEREO ENCODER

Samsudin[1], *Evelyn Kurniawati*[2], *Ng Boon Poh*[1], *Farook Sattar*[1], *Sapna George*[2]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]STMicroelectronics Asia Pacific Pte. Ltd.

ABSTRACT

In this paper, a signal-adaptive, stereo-to-mono downmixing scheme associated with MPEG-4 Parametric Stereo (PS) encoding is presented. The proposed scheme minimizes signal cancellation and coloration due to inter-channel phase misalignment. By using the inter-channel phase difference information which is calculated as a PS spatial parameter, the phase of the stereo signals are aligned. Subsequently, a simple averaging is carried out to mix the signals. The need to perform power equalization to preserve the overall power of the stereo signals in the downmix signal is eliminated. This leads to a significant saving of computational power. The scheme allows overall phase difference (OPD), one of the spatial parameter sent in PS bitstream, to be coded with minimum bit consumption. As a result, the entropy is reduced from 1.31 bits/symbol to 1 bit/symbol. This scheme is useful especially for stereo audio with a significant amount of side signal component.

1. INTRODUCTION

Recently, Parametric Stereo (PS) has been standardized in MPEG-4 Audio standard as an audio coding tool to increase coding efficiency [1]. The idea behind PS is to represent stereo audio by a downmixed mono audio and a small amount of spatial parameters which describe the stereo image of the original audio. PS is one of the key enabling technologies in *enhanced aacPlus* audio codec, which is considered as the state-of-the-art high quality, very low bit-rate audio coding at a range of 14 – 48 kbps. The codec has been chosen as one of the optional audio codecs for 3rd Generation mobile packet-switch streaming service [2].

As the content of the original stereo signal is solely represented by the monaural signal, the downmixing process ideally has to preserve all of the signal components. The standard suggests a simple time-domain averaging of the stereo audio samples. However, simple summation often results in amplification or attenuation of signal components.

In *enhanced aacPlus* reference implementation from 3rd Generation Partnership Project (3GPP) [3], the downmixing process is carried out in complex Quadrature Mirror Filter (QMF) subband domain according to:

$$m(k,n) = \frac{l(k,n) + r(k,n)}{2} \cdot \gamma(k,n) \tag{1}$$

where $m(k,n)$, $l(k,n)$ and $r(k,n)$ are the mono, left, and right subband samples respectively, k is the frequency channel index, n is the subband sample index, and $\gamma(k,n)$ is the stereo scale factor to ensure overall power preservation, defined as:

$$\gamma(k,n) = \min\left(2, \sqrt{\frac{|l(k,n)|^2 + |r(k,n)|^2}{0.5 \cdot |l(k,n) + r(k,n)|^2}}\right). \tag{2}$$

To comply with the PS decoding process, the stereo scale factor is defined such that the power of the mono signal is half the total power of the stereo signals. It is limited to 6 dB ($\gamma(k,n) = 2$) to prevent artifacts resulting from a large gain when the attenuation of the power of the sum signal is significant [4]. As PS decoding algorithm is beyond the scope of this paper, interested readers may refer to [1][5].

This downmixing scheme may suffer from signal cancellation and coloration in the case of stereo signals which are not in phase. An example is stereo recording with significant amount of side signal components, which are usually out of phase. When the stereo signals at a particular subband are exactly anti-phase and of the same amplitude, the corresponding signal component is lost after downmixing. The proposed scheme solves this problem by performing phase alignment to the stereo signal components prior to the signal mixing.

This paper is organised as follows. First an overview of MPEG-4 PS encoding is presented. The proposed downmixing scheme and its advantages are elaborated in Secion 3. Performance comparison with a simple downmixing scheme as described by (1) is presented in Section 4, followed by conclusion in Section 5.

2. MPEG-4 PARAMETRIC STEREO ENCODING

The structure of the PS encoder is shown in Figure 1. The time domain audio samples at both channels are segmented

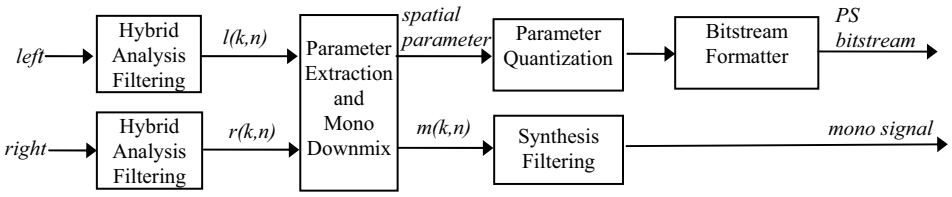

Figure 1. General structure of a parametric stereo encoder.

into non-overlapping frames of 2048 samples. Each frame is filtered by a 64-band complex-modulated quadrature mirror filter (QMF) followed by down-sampling by a factor of 64. The filtering process results in 32 complex-subband samples for each of the 64 frequency channels. The first few lower frequency channels are further filtered to increase the frequency resolution, resulting in a total of 91 subbands for a high quality PS configuration. These subbands are referred to as *hybrid subbands*, and the overall analysis filtering scheme is called *hybrid analysis filtering*. *l(k,n)* and *r(k,n)* are the left and right channel hybrid subband representation, where $k = 0,1,\ldots,90$ is the frequency channel index and $n = 0,1,\ldots,31$ is the subband sample index.

The hybrid subbands are grouped non-linearly into 34 *stereo bands* indexed by b ($b = 0,1,\ldots,33$). The grouping follows the equivalent rectangular bandwidth (ERB) with finer bandwidth at lower frequency and coarser bandwidth at higher frequency [5]. One set of spatial parameters is calculated at each stereo band and subsequently differentially coded over time or frequency (whichever requires less bits) into the bitstream. In addition to spatial parameter calculation, the stereo signal are downmixed into monaural signal *m(k,n)* and can be further processed or synthesized back into time domain and passed to a generic mono audio coder.

The spatial image of the original stereo audio is described by a set of time and frequency varying parameters: Inter-channel Intensity Difference (IID), Inter-channel Phase Difference (IPD) and Inter-channel Coherence (ICC). IID and IPD are self-explanatory, while ICC is meant to describe the spatial diffuseness of the original stereo audio. An additional parameter, Overall Phase Difference (OPD) is required to describe the relative phase distribution between the mono and the left channel. The decoder applies a phase shift equal to the OPD to reconstruct the phase of the left channel from the decoded mono signal and a phase shift equal to the OPD minus the IPD to reconstruct the phase of the right channel from the decoded mono signal.

3. A NEW SIGNAL ADAPTIVE DOWNMIXING SCHEME FOR MPEG-4 PARAMETRIC STEREO

3.1. Inter-channel phase difference calculation

IPD is one of the spatial parameter that is of particular interest. As it describes the phase difference between the two audio signals at each stereo band, its values reflect the amount of phase misalignment between the two stereo subband samples. Omitting the subband sample index n for clarity, IPD is formulated as [5]:

$$IPD[b] = \angle\left(\sum_{k=k_b}^{k_{b+1}-1} l(k) r^*(k)\right) \tag{3}$$

where k_b is the frequency boundary of the corresponding stereo band partition, and * denotes complex conjugation. The standard defines that only IPD information with stereo band index of up to $b = 16$ is sent in the bitstream. However, as it is desirable to perform phase alignment to the full range of the audio spectrum, the IPD parameter is calculated at all stereo bands.

3.2. Phase shifting

In the new scheme, the phase of the left subband signal is taken as the reference phase. At each subband, the phase of the right subband signal is shifted as much as the corresponding IPD according to:

$$r'(k,n) = \exp(jIPD(b_k)) \cdot r(k,n) \tag{4}$$

where *r'(k,n)* is the phase-shifted subband signal and b_k is the stereo band index b which corresponds to the particular hybrid subband indexed by k.

3.3. Signal averaging

Finally, simple averaging is performed on the aligned signals to obtain the mono signal according to:

$$m(k,n) = \frac{l(k,n) + r'(k,n)}{2}. \tag{5}$$

3.4. Advantages of the proposed scheme

3.4.1. Automatic power preservation

Comparing (5) with (1), the stereo scaling factor has been eliminated in the proposed scheme. In other words, the power equalization step is omitted due to the automatic power preservation of the proposed scheme. This amounts to a saving of the computation of (32×91) scaling factor per

frame (each frame consists of 32 samples for each of the 91 hybrid subbands). The power preservation can be verified by analyzing the instantaneous power of the downmix sample. The left and right complex subband samples can be represented in the polar form as:

$$l(k,n) = A_L(k,n) \cdot e^{j\theta_L(k,n)} \quad (6)$$

$$r(k,n) = A_R(k,n) \cdot e^{j\theta_R(k,n)} \quad (7)$$

where $A_L(k,n)$, $\theta_L(k,n)$, $A_R(k,n)$ and $\theta_R(k,n)$ are the amplitude and instantaneous phase of the left and right subband samples respectively.

Omitting the frequency index, k, and time index, n, for notational simplicity, the instantaneous power of the mono signal can be derived as follows:

$$\begin{aligned} |m|^2 &= \left|\frac{l+r}{2}\right|^2 \\ &= \frac{1}{4}\left|A_L \cdot e^{j\theta_L} + A_R \cdot e^{j\theta_R}\right|^2 \\ &= \frac{{A_L}^2 + {A_R}^2}{4} + \frac{A_L A_R \cos(\theta_L - \theta_R)}{2}. \end{aligned} \quad (8)$$

From (8) it can be seen that the power of the mono signal consists of the scaled total power of the stereo channels and a varying component which is dependant on the term $\cos(\theta_L - \theta_R)$. The term $(\theta_L - \theta_R)$ itself is the inter-channel phase difference, which is approximated by the IPD parameter. It can be seen that the power of the mono signal is maximized when the phase difference is zero and attenuated as the phase difference increases.

In the proposed scheme, a phase shift of the amount equal to the IPD is introduced to the right subband samples according to (4). In simplified notation, the phase of $r'(k,n)$, $\theta_{R'}$, is

$$\theta_{R'} = \theta_R + (\theta_L - \theta_R) = \theta_L. \quad (9)$$

Now the θ_R term in $\cos(\theta_L - \theta_R)$ in (8) is replaced by $\theta_{R'}$, and it can be seen that the cosine term is always maximum. Hence, the power of the mono signal is maximized without the need to perform power equalization.

Figure 2(a) shows the effect of inter-channel phase difference on the power of the mono signal of a 1 kHz stereo sinusoidal signal with equal amplitude on both channels. The plot labeled '*Stereo*' is the normalised sum of the power of both stereo channels ($|l|^2+|r|^2$), while the plot labeled '*Averaging*' is the normalised power of the downmixed mono signal obtained using simple averaging without power equalization. As the phase difference increases, the power of the mono signal is attenuated in a cosine function manner.

Figure 2(b) shows the comparison of the power of the mono signals obtained using the downmix scheme from 3GPP's *enhanced aacPlus* reference implementation [3]

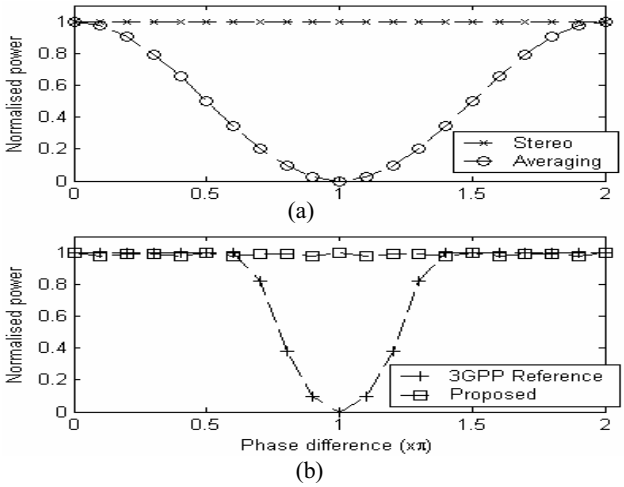

Figure 2. Power of downmixed signal as a function of inter-channel phase difference.

with that of the proposed scheme. It is evident that with the proposed scheme, the total power in the mono signal is preserved for the whole range of phase difference.

3.4.2. Omission of OPD parameter

As the phase of the right subband signal has been aligned to the phase of the left subband signal, the mono audio signal has the same phase as the left channel. This implies that OPD is always zero. As a result, the OPD parameter can be omitted in the bitstream, although this is not possible due to the bitstream definition in the standard. However, as the OPD values are always zero, the differential values are also always zero. With the specified Huffman coding table for OPD [1] whereby a differential value of 0 is coded using only 1 bit, the number of bits required to transmit OPD is always minimum. According to the entropy values obtained by entropy analysis of PS parameter symbol [5], the entropy for the OPD parameter is reduced from 1.31 bits/symbol to 1 bits/symbol. In addition, the decoder does not need to perform phase modification in order to obtain the reconstructed left channel. Rather, only a phase modification of (–IPD) is required to produce the reconstructed right channel from the monaural signal. This results in simplification on the decoding side.

4. RESULTS

The performance of the proposed downmixing scheme is compared to the downmixing scheme implemented in 3GPP *enhanced aacPlus* reference encoder as defined by (1), which will be subsequently refered to as *simple downmixing scheme*. The comparison is performed using an audio clip from Dolby Digital trailer, *canyon*. It is a downmix from a 5.1 multi-channel audio, hence it contains significant side signal components. The clip is encoded using both the simple and proposed downmixing schemes. The resulting

mono audio from each scheme are synthesized back into stereo audio by the PS decoder.

Informal listening test revealed that there is a significant loss of surround signal components of the mono audio generated using the simple downmixing scheme. The loss is clearly audible. The problem does not appear in the mono audio generated using the proposed downmixing scheme. The original, downmixed, and the decoded audio clips presented in this paper can be downloaded from *http://www.ntu.edu.sg/home5/sams0006/audio.zip*.

Figure 3 shows the spectrogram of the original and downmixed audio of a short extract from *canyon* using both schemes. It can be seen that there is significant loss of signal components due to phase cancellation from the simple downmixing scheme, as shown in Figure 3(c). Using the proposed downmixing scheme, it is evident that the scheme preserves most of the signal components, as shown in Figure 3(d).

Figure 4 shows the average Objective Difference Grade (ODG) versus time plot of the same audio extract from *canyon*, with the original extract as the reference signal and the PS decoded extracts as the test signals. The plots are obtained using OPERA, a commercial software implementation of PEAQ (perceptual evaluation of audio quality) as defined in ITU-R BS.1387 [6]. With the proposed scheme, an improvement of more than 1.5 ODG score is observed. The final ODG scores are -1.67 and -3.60 for the decoded stereo audio generated from the proposed and simple downmixing scheme respectively.

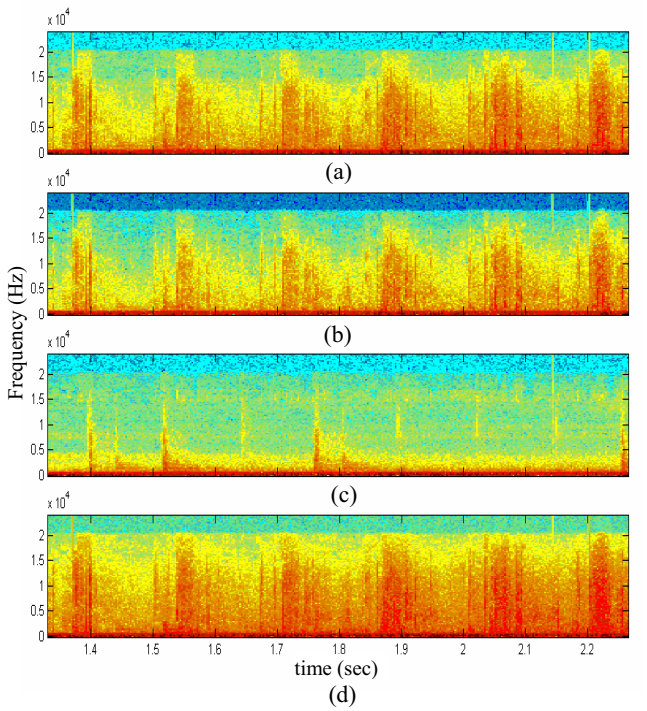

Figure 3. Spectrogram of audio extract from *canyon*: (a) original left channel, (b) original right channel, (c) mono audio with simple downmixing scheme, (d) mono audio with proposed downmixing scheme.

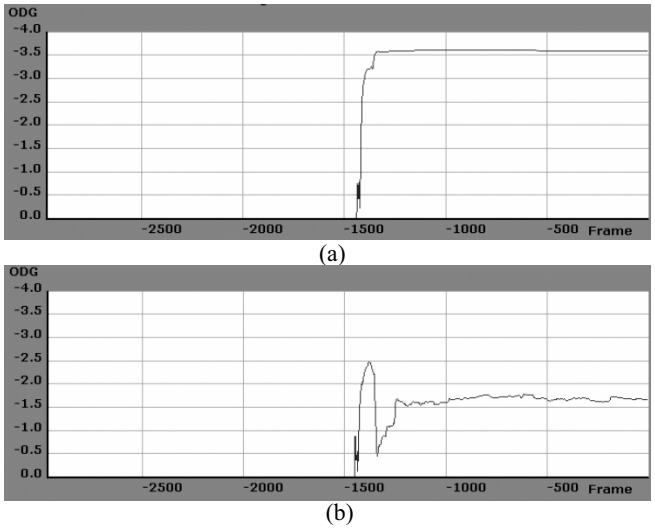

Figure 4. Average ODG versus time plot of the decoded audio extract from *canyon*: (a) PS encoding with simple downmixing scheme, (b) PS encoding with proposed downmixing scheme.

5. CONCLUSION

A signal-dependent downmixing scheme for PS audio coding tool has been described. It does not involve many additional computations due to the usage of inter-channel phase difference information, which has been readily calculated as one of the spatial parameters. Phase shifting to avoid signal cancellation is carried out easily due to complex representation of the subband samples.

It has been verified that the scheme is power preserving and it enables the OPD parameter to be coded with minimum bits for better coding efficiency. Finally, it has been shown that the performance of the proposed scheme is superior over a simple averaging scheme for real life audio signal.

6. REFERENCES

[1] *Information technology – Coding of audio-visual objects – Part 3: Audio, Amendment 2: Parametric coding for high-quality audio*, ISO/IEC 14496-3/Amd. 2: 2004.

[2] *Transparent end-to-end Packet-switched Streaming Service (PSS); Protocols and codecs*, 3GPP TS 26.234 Rel. 6, 2005.

[3] *Enhanced aacPlus general audio codec*, 3GPP TS 26 series Rel. 6, 2004.

[4] C. Faller, *Parametric Coding of Spatial Audio*, Ph.D. thesis École Polytechnique Fédérale de Lausanne, France, 2004.

[5] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "High-quality parametric spatial audio coding at low bitrates", *Proc. 116th AES Convention*, Berlin, Germany, Preprint 6072, May 2004.

[6] *Method for Objective Measurements of Perceived Audio Quality*, ITU-R Recommendation BS.1387, Dec. 1998