Low Complexity Decoding in Parametric Stereo Audio Coding Scheme

Run-Yu Tong and Pao-Chi Chang
Communication Engineering Department, National Central University, Jhongli, Taiwan

Abstract —**Parametric Stereo (PS) is an audio coding object of MPEG-4 HE-AAC v2 which utilized the Spatial Audio Coding (SAC) technique to enhance the compressing efficiency. However, the complexity at decoder is higher than that at encoder in PS. In this paper, we proposed a low complexity decoding scheme in PS. To take advantage of SAC, the encoder additionally extracts and transmits the parameters of residual and mono signals. The decoder detects the transient signal of received mono signal, and compensates the energy aliasing of reconstructed residual signal. The experiment result also shows a better average objective quality.**

I. Introduction

MPEG Layer-3 (MP3) and MPEG-4 Advance Audio Coding (AAC) exhibit high coding efficiency by utilizing the psychoacoustic model to remove the masked frequency components. However, the psychoacoustic model aims at the analysis of single channel audio signals without considering the correlation between audio channels. Parametric Stereo (PS) is a spatial audio coding structure for stereo audio, which is standardized by ISO/IEC in 2003 [1]. The encoder and decoder of PS could be regarded as a pre-processing and a post-processing in the existed infrastructure. According to that, PS combines with core codec that could be any traditional audio codec such as MP3 or AAC, and decreases the encoding bit-rates appropriate for network transmitting or storage. As shown in Fig. 1, the overall coding scheme is that PS combines with core codec.

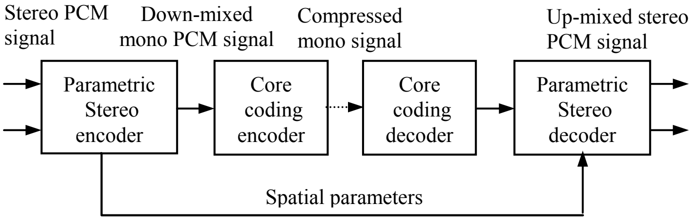

Stereo PCM signal
Down-mixed mono PCM signal
Compressed mono signal
Up-mixed stereo PCM signal
Parametric Stereo encoder
Core coding encoder
Core coding decoder
Parametric Stereo decoder
Spatial parameters

Fig.1 Parametric Stereo combined with core codec

When the stereo signal enters the PS encoder, the block would extract the spatial parameters and down-mix the stereo to mono signal, then sent the mono signal to the core coding encoder. The overall bit-rates of core coding could be utilized by mono signal. Each channel of stereo signals needs half bit-rates of core coding if we do not put PS to use.

However, the down-mixed procedure are the operations of multiplications and additions to generate a mono signal at PS encoder. Comparing to encoder, the up-mixed procedure needs a high order filter to reconstruct other single channel signal frequency-independently. The high order filter would bring much higher complexity at PS decoder. Because of the high complexity at decoder, implementing PS on portable device would cost more power. Thus we propose a method which utilizes the residual coding and spatial residual parameters to reduce the complexity at decoder. In addition, we also design an algorithm to deal with energy aliasing which is caused by transient signal. Also the Objective Difference Grade (ODG) has got improved. The new PS coding scheme combines with MPEG-4 AAC at low bit-rates which still maintain acceptable sound quality.

II. Binaural Cues of Spatial Audio Coding

The ability of human auditory system (HAS) can localize sound sources which is based on the physical distance of two ears. This reason causes sound to arrive at ears slightly different. The differences are generated by the filtering effect of the head, torso and pinna which could be described completely by the Head-Related Transfer Functions (HRTFs) [1]. By exploiting the HRTFs, we can derive the sounds that were perceived by two ears from source. Thus we can transmit single sound source with HRTFs instead of the stereo audio.

However, it is very expensive to represent HRTFs in terms of bit-rate, therefore the most important localization cues are segregated from the HRTFs:

- Inter-Channel Intensity Differences (ICID),
- Inter-Channel Phase Differences (ICPD),
- Inter-Channel Coherence (ICC).

The ICID is the main localization cue at the frequencies above approximately 1.5 kHz, since the low frequency sound travels through the head and does not substantially attenuated. The ICPD defines the difference in the arrival time of sound at two ears, also it is the main localization cue at the frequencies below 1.5 kHz, since human auditory system is not sensitive for the phase difference at high frequency.

For example, if the signals of left and right channel are coherent, that is ICID=0 and ICPD=0, auditory event appears in the center between the left and right channels of listener as Region 1 in Fig. 2 (a). By increasing the level on right channel, the auditory event moves to right side as Region 2 in Fig. 2 (a). In the extreme case, when only the signal on the left is active, the auditory event appears at the left source position as Region 3 in Fig. 2 (a).

ICC is defined as the maximum absolute value of the normalized cross-correlation function which is a measure for "signal similarity" between channels. The width of the auditory event increases (Region 1-3 in Fig. 2 (b)) as the ICC between the left and right signals decreases, until two distinct auditory events appear at the sides.

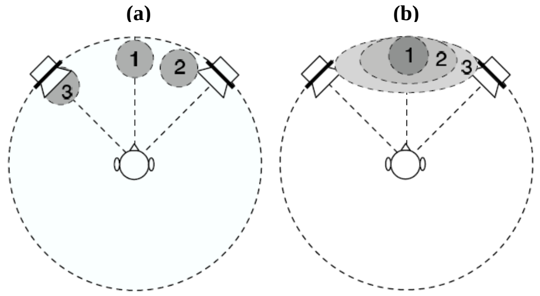


Fig. 2 (a) The example of sound source localization [2]. (b) The example of auditory event width [2].

Conclusively, SAC extracts ICID, ICPD and ICC, which is the localization and width information of auditory events, as spatial characteristic parameters for reconstructing sound field at decoder.

III. The Parametric Stereo Coding Scheme

A. *Parametric Stereo Encoder*

Parametric Stereo (PS) is a spatial audio coding structure for stereo audio, which is standardized by ISO/IEC in 2003 [3]. The encoder is shown as Fig. 3.

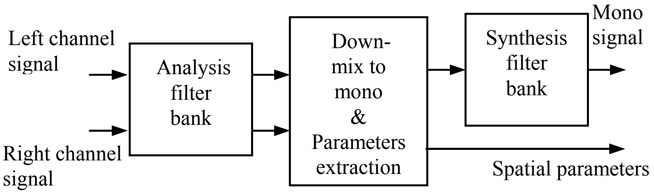


Fig. 3 The block diagram of Parametric Stereo encoder.

The PS algorithm firstly uses 64-band complex-exponential modulated Quadrature Mirror Filter (QMF) banks as time-to-frequency and frequency-to-time transforms. For the typical audio sampling rate of 44.1 kHz, the 64-band analysis filter banks result in an approximately 344 Hz effective bandwidth per band. Nevertheless, the spectral resolution of HAS closely follows the Equivalent Rectangular Band-width (ERB) scale [4]. There are finer resolution at low-frequency and coarse resolution at high-frequency in frequency domain. Thus hybrid filtering is applied to the lowest 3 QMF bands and finally forms 71-band hybrid QMF banks (refer to [5] for more details). After hybrid QMF analysis, the time/frequency (t/f) representation of input signal is shown as Fig. 4.

Consequently, we extract spatial parameters (ICID, ICPD, and ICC) from each t/f tile between left and right channel. More details of the formula could refer to [6]. However, it is impossible to extract parameters from each t/f tile since the bit-rate will be too high to transmit. Therefore the number of parameters is limited by some partitions along time and frequency axes which are called parameter sets and parameter bands, as illustrated in Fig. 4.

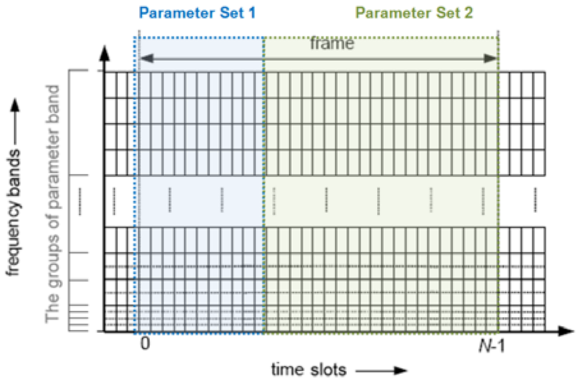


Fig. 4 Time/Frequency representation with the group scale of parameter set and band (time slots = N in a frame) (modified from [5]).

The grouping of parameter bands is also matching the human auditory characteristics. After the t/f grouping, we capture only one set of spatial parameters in each group in order to reduce the bit-rate of parameters.

B. *Parametric Stereo Decoder*

The Parametric Stereo decoder is shown as Fig. 5.

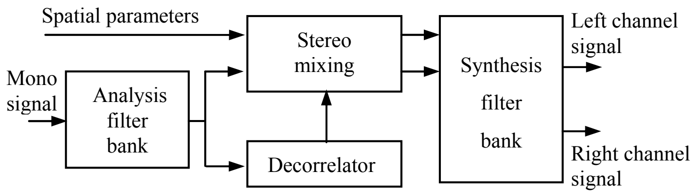


Fig. 5 The block diagram of Parametric Stereo decoder.

The mono signal and spatial parameters were received at the decoder. The mono signal also passes through the analysis filter banks as same as that in the encoder. However, we could not represent stereo sound field by mono signal. For this reason, PS would utilize a decorrelator to generate other single channel signal which is called "decorrelated signal." The decorrelator is an all-pass filter which would change the phase and the amplitude of input signal frequency-independently. For the 71 hybrid bands, different band would select the different filter delay and attenuated gain. According to the HAS, the signal at each band of 30 low frequency bands is filtered by an independent Infinite Impulse Response (IIR) filter. The signals at other 41 bands were filtered by the filters which have the same delay but distinct and attenuated gains [7]. From the previous work, we know the algorithm of decorrelator may cost high complexity, and it does occupy almost 50 percent of the decoding time. That is why decoding time is twice bigger than encoding time. In next section, we would propose a new coding scheme to solve this problem.

IV. The proposed Low Complexity Decoding Scheme

Fig. 6 and Fig. 7 show the encoder and decoder of proposed coding scheme.

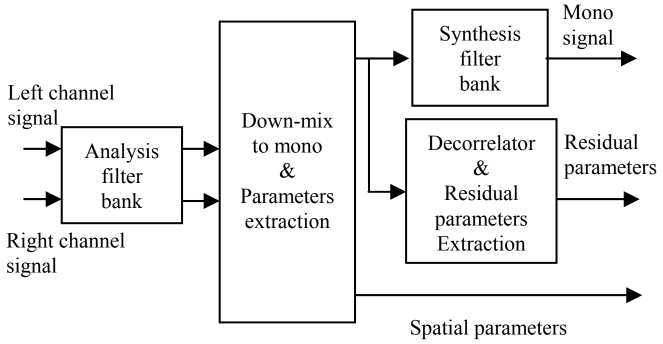


Fig. 6 The block diagram of proposed Parametric Stereo encoder.

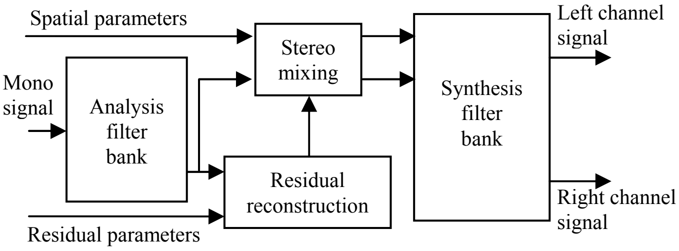


Fig. 7 The block diagram of proposed Parametric Stereo decoder.

Our works mainly consist of three parts: the residual parameters extraction, the reconstruction of residual signal, and the energy compensated algorithm. The details are described as follows.

A. The residual parameters extraction

The residual coding usually is used to improve the coding efficiency at speech coding. Because of the fundamental periodic characteristic, speech signals could get very high coding efficiency. Nevertheless, audio signals do not have the property as same as speech signals. If residual coding is used for single channel audio signal, that would cost very high bitrate about 200k bits per second. Now we utilize the high quality of SAC to extract the residual parameters of the mono signal and residual signal as shown in Fig. 6. The residual signal $Res(n,k)$ is the difference of mono signal $M(n,k)$ and decorrelated signal $D(n,k)$ as in (1).

$$Res(n,k) = M(n,k) - D(n,k) \tag{1}$$

Where the indices n and k mean the n^th^ time-slot and the k^th^ hybrid-band. According to this method, we could reduce the bitrates and just cost about 10~15k bits per second.

B. The reconstruction of residual signal

As shown in Fig. 7, the residual signal could be reconstructed by mono signal and the residual parameters as in (2).

$$Res^{'}(n,k) = R_{res}(n,k) * M(n,k) \tag{2}$$

The reconstructed coefficient $R_{res}(n,k)$ refers to the residual parameters. It is derived with physical meaning from the mixing matrix of stereo mixing in Fig. 7 [8][9]. Then we could get the reconstructed decorrelated signal $D^{'}(n,k)$ and reduce the complexity by means of (3).

$$D^{'}(n,k) = M(n,k) - Res^{'}(n,k) \tag{3}$$

The decorrelator was replaced by the reconstruction of residual signal which could reduce about half decoding time.

C. The energy compensated algorithm

When we use the residual coding in the encoder, the aliasing effect would get worse in the decoder which is causing by transient signals. The decorrelated signal was got by passing the mono signal through the all-pass filter. The all-pass filter just changes the phase and the amplitude of the input signal. If we implement (1), extract and quantize the residual parameters. It would cause the Inter-Channel Intensity Difference of the residual parameters $ICID_{res}$ centralizing to specific quantization level as Fig. 8 shown.

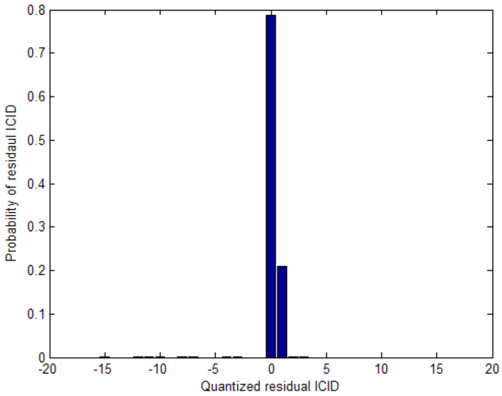
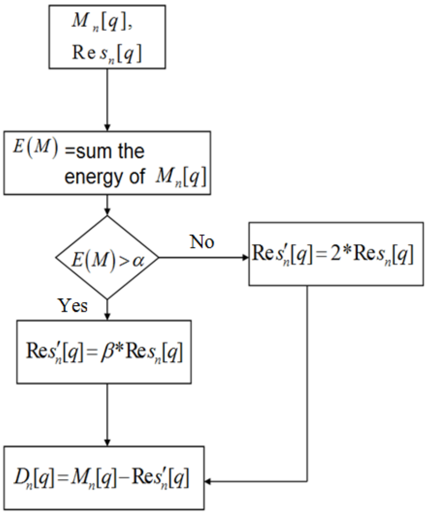


Fig. 8 The probability statistics of quantized $ICID_{res}$ (Moon.wav)

Therefore the reconstructed residual signal would be more sensitive to transient signal in the decoder. To solve the energy aliasing which is causing by transient signal, we propose an energy compensated algorithm to fix the problem. The energy compensated algorithm is described in Fig. 9.



$M_n[q]$, $\mathrm{Re}s_n[q]$
$E(M)$ =sum the energy of $M_n[q]$
$E(M)>\alpha$
No
$\mathrm{Res}'_n[q]=2*\mathrm{Res}_n[q]$
Yes
$\mathrm{Res}'_n[q]=\beta*\mathrm{Res}_n[q]$
$D_n[q]=M_n[q]-\mathrm{Res}'_n[q]$

Fig. 9 The energy compensated algorithm

For the lowest 30 hybrid-bands, we sum all the hybrid-band energy of mono signal first. If the sum of energy exceeds the threshold α, let residual signal multiply by β. If not, the residual signal would multiply by 2. Threshold α is an experimental value which is set for 2.5 and β is a scalar which multiplies the energy ratio of mono signal to residual signal by 1.5 times.

V. Experimental Results

The Table I and Table II show the complexity analysis of encoder and decoder of PS and Low Complexity Decoding Parametric Stereo (LCDPS). The complexity of encoder in LCDPS would increase because of the decorrrelator. Most of the computing machines are powerful at the encoder. So we do not discuss the complexity at the encoder especially.

TABLE I

COMPLEXITY ANALYSIS OF ENCODER (DOMINGO.WAV)

Block	Time(Second)	
	Complexity of PS Encoder	**Complexity of LCDPS Encoder**
QMF_Analysis	12.28	12.75
QMF_Hyb_Analysis	0.83	0.92
Down-Mix/Para_ Extra	1.04	1.05
Decorrelator/Residual Para_Extra	-	19.9
QMF_Hyb_Synthesis	0.25	0.27
QMF_Synthesis	7.13	7.08
Encoding Time	22.16	40.44

According to experiment results, the complexity of decoder in LCDPS actually saves about half time. This is the focus in our work. For most of portable device at the decoder, the complexity is the main consideration because of the finite power.

TABLE II

COMPLEXITY ANALYSIS OF DECODER (DOMINGO.WAV)

Block	Time(Second)	
	Complexity of PS Decoder	**Complexity of LCDPS Decoder**
QMF_Analysis	6.02	5.39
QMF_Hyb_Analysis	0.48	0.42
Decorrelator	18.53	-
Residual reconstruction	-	1.14
Stereo Reconstruction	1.11	0.98
QMF_Hyb_Synthesis	0.53	0.48
QMF_Synthesis	12.39	10.91
Decoding time	39.86	19.28

A. EAQUAL - Evaluation of Audio Quality

EAQUAL is an objective measurement technique based on the ITU-R recommendation BS.1387. The results of EAQUAL are interpreted as an ODG (Objective Difference Grade) score. The ODG scores range from -4 to 0.The -4 means annoying impairment, and 0 means imperceptible impairment. However, EAQUAL is developed for the quality measurement of lossy audio coding based on utilizing the masking effect (e.g. MP3、AAC), thus it is not suitable for spatial audio coding.

The algorithm of EAQUAL regards the waveform differences as quantization noise due to compression and calculates its effect on HAS. In the case of spatial audio coding, waveform difference is not simply resulting from quantization noise. It generally includes quantization noise from core coding and waveforms mixed from other channels. If there exists high correlation on 2 channels, waveforms mixed from other channels will not lead to obvious quality degradation. However, EAQUAL algorithm considers it as quantization noise and results in the under estimation of audio quality.

B. System performance of proposed low complexity decoding parametric stereo scheme

The length of test sequences are 15~18 seconds. The varieties of sequence include musical instruments, symphony, and tenor …etc. Some of them are got by Sound Quality Assessment Material (SQAM) which came from European Broadcast Union (EBU). The experimental environment was set up by OS: Windows XP, Platform: Matlab R2008a, CPU: Intel Core i7 2.67GHz, Ram: DDRII-800 2GB x2.

Although the measurement does not perfectly reflect the actual presentation of all test sequences. Consulting the average score, LCDPS still performs well as the same as PS, even makes better than PS. Table III shows the comparison of ODG scores between PS and LCDPS.

TABLE III

THE ODG SCORES OF LCDPS AND PS

	LCDPS ODG	PS ODG
(1)Moon	-2.85	-3.34
(2)Domingo	-2.52	-3.09
(3)Fanfare	-3.44	-3.48
(4)EBU13_Flute	-3.78	-3.32
(5)EBU43_Organ	-3.54	-2.91
(6)EBU58_Guitar	-1.98	-2.64
(7)EBU64_Choir	-1.72	-1.85
Average score	-2.83	-2.94

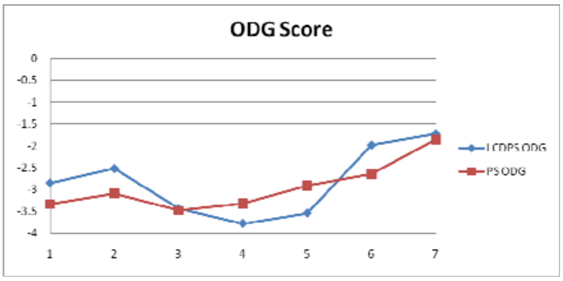


Fig. 10 The broken line graph of ODG scores of LCDPS and PS

Figure. 10 is the related graph of the ODG scores. The ODG score of LCDPS seems to be slightly different form PS. To consider the complexity, we prefer using the LCDPS to reduce the complexity at decoder.

VI. CONCLUSION

The residual coding whose bitrates is reduced greatly by utilizing the residual parameters at the encoder. Furthermore, the proposed PS coding scheme actually reduces the complexity of decoder about half time but not overall coding scheme, and the sound quality had been improved by energy compensated algorithm. From this point of view, it would be a nice choice for the customers.

REFERENCES

[1] J. Jakka, “Binaural to Multichannel Audio Upmix,” Department of Electrical and Communications Engineering, HELSINKI UNIVERSITY OF TECHNOLOGY, Espoo, June 6, 2005.

[2] C. Faller, “Parametric coding of spatial audio,” in *Proc. of the 7th International Conference on Digital Audio Effects (DAFx’04)*, Naples, Italy, October 5-8, 2004.

[3] ISO/IEC JTC1/SC29/WG11, “Text of ISO/IEC 14496-3:2001 /FPDAM2 (parametric coding for high quality audio),” ISO/IEC JTC1/SC29/WG11 N5713, July 2003.

[4] J. Hall and M. Fernandes, “The role of monaural frequency selectivity in binaural analysis,” in *J. Acoust. Soc. Amer.*, 1984, vol. 76, pp. 435 – 439.

[5] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdeg˙ard, “Low complexity parametric stereo coding,” in *Proc. 116th AES Convention*, Berlin, Germany, May 2004.

[6] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers “Parametric coding of stereo audio”, *EURASIP Journal*, Applied Signal Processing 9:1305-1322, 2005.

[7] J. Engdegard, H. Purnhagen, J. Roden, and L. Liljeryd, “Sythetic ambience in parametric stereo coding,” in *Proc. 116th AES Convention*, Berlin, Germany, May 2004, pp.1-12.

[8] T.C. Li, and P.C. Chang, “Spatial characteristic based scalable audio coding structure, “ in *Proc. of National Symposium on Telecommunications (NST)*, Taipei, Taiwan, pp. 452-456, Dec. 2009.

[9] R. Irwan and R. M. Aarts, “Two-to-five channel sound processing,” *Journal of the Audio Engineering Society*, vol. 50, no. 11, pp. 914–926, 2002.