PARAMETRIC STEREO CODING SCHEME WITH A NEW DOWNMIX METHOD AND WHOLE BAND INTER CHANNEL TIME/PHASE DIFFERENCES

Wenhai Wu[1], Lei Miao[1], Yue Lang[2], David Virette[2]

[1]Huawei Technologies, China, [2]Huawei European Research Center, Germany

ABSTRACT

This paper presents a novel low bit rate parametric stereo coding scheme which uses whole band inter-channel time difference (WITD) and whole band inter-channel phase difference (WIPD) together with a new effective downmixing method. The inter-channel level differences and inter-channel phase differences are also employed in the proposed stereo coding to further improve the quality in an embedded structure. This scheme is applied to the stereo extension of ITU-T G.722 at 56+8 kbits/s with a frame length of 5 ms. Listening test results are provided to assess the quality of the proposed downmixing method and WITD/WIPD coding scheme separately.

Index Terms—Stereo coding, whole band inter-channel time difference, whole band inter-channel phase difference, downmixing

1. INTRODUCTION

Nowadays applications such as spatial audio conference or immersive audio communication are getting more and more popular. For stereo communications, several possible stereo recording configurations can be used with various constraints and quality, such as A-B microphone, M/S (Mid/Side) microphone, and binaural microphone. There is a need for robust stereo audio coding to address those scenarios and to improve the user experience. The use of one mono codec per channel does not offer a suitable solution as the required bit rate increases almost linearly with the number of channels and the direct backward compatibility is not completely ensured. For example, using only one of the channels for a legacy mono decoder could definitely affect the perceived quality, if the content of the two channels is different. There is a strong need to deliver stereo with reasonable bit rate and offering the mono compatibility for legacy terminals.

Stereo coding using a parametric representation based on spatial cues has been investigated in a number of papers. Some techniques are already widely used, such as Binaural Cue Coding (BCC) [1][2] and Parametric Stereo (PS) [3][4]. Parametric stereo coding techniques generate a mono downmix. The mono signal is encoded by a legacy mono coder together with side information representing the spatial cues [5]. The stereo decoder uses this side information to reconstruct the original number of audio channels based on the decoded waveform provided by the legacy mono decoder. The side information usually consists of several parameters per frequency sub-band [3][4]:

- Inter-channel Level Difference (ILD) measuring the level difference (or balance) between channels.
- Inter-channel Time Difference (ITD) or Inter-channel Phase Difference (IPD) describing the time or phase difference between channels, respectively,
- Inter-channel Coherence (IC) which represents the correlation or coherence between channels.

This work presents an approach derived from the BCC and addresses the low bit rate stereo coding for conversational applications with low delay and small frame size. It relies on two parameters which represent the whole band time and phase differences between the stereo channels and a new downmixing method aligning the phase of the downmix signal to the main perceived direction.

This paper is organized as follows. Firstly, the theory of the low bit rate stereo analysis and synthesis using WITD and WIPD is described in Section 2. Section 3 describes the novel downmixing method with phase alignment in the frequency domain. In Section 4, the integration of the proposed stereo coding in ITU-T G.722 Annex D is discussed. In Section 5, some experimental results are presented to demonstrate the benefit of the proposed method before concluding and reviewing the relation to prior work.

2. STEREO ANALYSIS/SYNTHESIS WITH WITD/WIPD

For MS microphone recorded signals, it has turned out that by merely considering ILD and IC, a good quality stereo coding can be achieved [6][7]. For AB microphone and binaural microphone recorded signals, ITD and IPD play an important role for the localization of sound sources, as it provides spatial cues to identify the direction of the sources. ITD represents the time delay between the amplitude envelop of the two channels. It is related to the interaural time difference, i.e. the difference in arrival time of a sound between two ears.

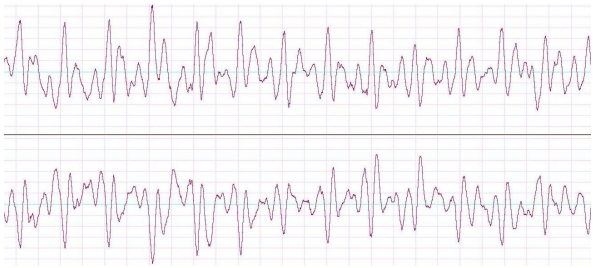

Fig. 1. Stereo signal with out of phase channels.

When considering sub-band IPD, the phase difference can be directly related to a sub-band delay and then consistently represents the same information. Consequently, both methods can be used to express the spatial cues for sub-band analysis of stereo signals.

In recently proposed parametric stereo coding schemes [1-4], the sound localization cues have been represented by either sub-band ITDs or sub-band IPDs. However, in this work, we consider low bit rate stereo side information, and the transmission of all sub-band ITD/IPD is not possible. We propose to estimate and transmit only a whole band time difference and a whole band phase difference to represent the stereo image. As opposed to other stereo coding scheme [1-4], both cues must be considered at the same time. The direct relation between WITD and WIPD is not verified anymore. An example of stereo signal, where the channels are out of phase, is shown on Fig.1. This phase difference cannot be observed in most of the natural recording, but artificial mixing may introduce such phase difference. The inversion of one channel with regard to the other cannot be simply represented by a whole band delay, while a delay between two channels cannot be represented by a whole band phase difference. Hence, the proposed parametric stereo coding algorithm uses both WITD and WIPD to faithfully represent the stereo image at low bit rate.

In the following sections, the estimation, synthesis and downmixing algorithms are based on *N*-point short-time Fourier transform. As a convention, it is defined that ITD is positive if a waveform comes first on the left channel.

2.1. Whole band ITD and IPD estimation

Sub-band IPDs are generally used to estimate the sub-band ITDs [2][3]. For the low frequencies, no phase wrapping occurs and the sub-band ITDs are computed as the slope of the sub-band IPDs [8]. Considering a delay range of $[-\tau_{max}, \tau_{max}]$ in seconds, it can be noted that no phase wrapping occurs for frequencies below $1/2\tau_{max}$ Hz. Consequently, the $IPD(k)$ per frequency bin k is computed as follows:

$$IPD(k) = \angle(L(k)R^*(k)). \quad (1)$$

From paper [2, 3, 8], we can see that an estimation of the IPD can be computed as a linear regression function of the WITD and the WIPD.

$$\hat{IPD}(k) = \frac{-2\pi WITD}{N}k + WIPD. \quad (2)$$

Assuming $\hat{IPD}(k) = IPD(k)$, WITD and WIPD can be estimated as follows:

$$WITD = -N\frac{\sum_{k=0}^{k_{max}-1}(IPD(k+1) - IPD(k))/k_{max}}{2\pi}, \quad (3)$$

and

$$WIPD = \sum_{k=0}^{k_{max}-1}(IPD(k) + k*\frac{2\pi WITD}{N})/k_{max}, \quad (4)$$

where k_{max} is the frequency bin index corresponding to the frequency $1/2\tau_{max}$ Hz.

2.2. Stereo synthesis with WITD/WIPD

At the decoder, IPD must be properly distributed to each channel. Without a priori information on phase distribution, IPD is equally distributed to each channel [8]. Using the proposed *WITD* and *WIPD*, the estimated $\hat{IPD}(k)$ is obtained using (2). Given the spectrum of the decoded mono signal $\hat{M}(k)$, the synthesized spectrums of the left and right channels are computed as follows:

$$\hat{L}(k) = \hat{M}(k)e^{j(\hat{IPD}(k)/2)}, \quad (5)$$

and

$$\hat{R}(k) = \hat{M}(k)e^{j(-\hat{IPD}(k)/2)}, \quad (6)$$

3. DOWNMIX WITH PHASE ALIGNMENT

Downmxing in the time domain does not finely take into account sub-band phase differences between channels. Hence, it cannot preserve the energy per frequency band due to potential cancellation of some frequency bands. In frequency domain, magnitude equalization and phase alignment control are commonly used [3], since magnitude equalization alone cannot sufficiently correct the undesired effect of signal cancellation for out-of-phase signals, such as the example shown in Fig. 1.

A special case of phase alignment downmix algorithm was proposed in [6]. The downmixed mono signal amplitude is calculated as the average of the left and right channel amplitudes and the phase is the sum of the left and right channel phases. For stereo signals as illustrated on Fig.1, the phase of the sum signal according to [6] will be close to the zero for all frequency bins. This wrong phase introduces audible distortions in the reconstructed downmix signal.

We proposed a new downmixing method in frequency domain to solve the problem of phase instability. The proposed downmixing signal is calculated by

$$M(k) = |M(k)|e^{j\angle M(k)}, \quad (7)$$

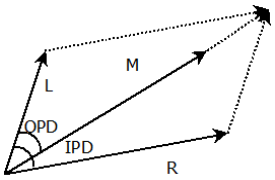

Fig. 2. Proposed downmix with phase alignment.

with the amplitude computed according to [6]:

$$|M(k)| = (|L(k)| + |R(k)|)/2. \quad (8)$$

The phase of the mono downmix is given by the difference between the phase of the left channel and the overall phase difference (OPD) [9] as shown in Fig. 2,

$$\angle M(k) = \angle L(k) - OPD(k). \quad (9)$$

The idea here is to make the phase of downmixed signal close to the phase of higher energy channel, in order to favor the reconstruction of the main sound sources in the downmixed signal. The estimation of OPD from IPD and ILD was presented based on the geometrical relationship in [10]. For simplicity, we consider that IPD is distributed between the channels as a function of the channel energy. Thus the OPD is defined as:

$$OPD(k) = \frac{1}{1+f(b)} IPD(k), \quad (10)$$

where b is the sub-band index, k is the frequency bin and $f(b) = 10^{ILD(b)/20}$. ILD parameters are estimated as in [1][2]. Thus equation (9) can be rewritten as

$$\angle M(k) = \angle L(k) - \frac{1}{1+f(b)} IPD(k). \quad (11)$$

For the communication applications, normally the frame size of the codec is short, which leads to poor frequency resolution and low accuracy in the sub-band IPD estimation. If the signal is out of phase (Fig. 1), the extracted sub-band IPD may rapidly change from π to –π. When applied in (11), it results in a large difference of the downmix phase between consecutive frames. This problem affects the reconstruction and introduces audible artifacts. In order to overcome this problem, the WIPD is subtracted from the *IPD*(*k*), which has the effect of setting the downmix phase to the left channel phase for such critical signal, while focusing on the most energetic channel phase for other cases. The phase of the mono downmix is expressed by:

$$\angle M(k) = \angle L(k) - \frac{1}{1+f(b)} (IPD(k) - WIPD). \quad (12)$$

4. PROPOSED STEREO CODING

4.1. Proposed stereo coding scheme

The proposed stereo coding scheme is shown in Fig. 3. The "Downmix" is performed using the algorithm detailed in Section 3.

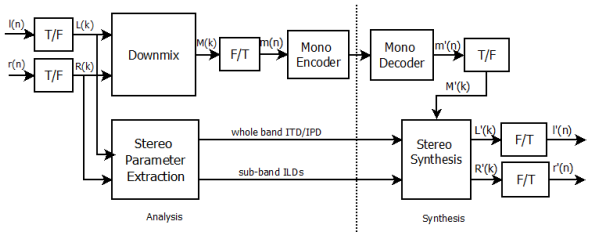

Fig. 3. Parametric stereo coding scheme

ILD are estimated and transmitted in the frequency domain following perceptual sub-band decomposition [1-4]. WITD and WIPD are extracted as described in Section 2. The decoded amplitudes of the left and right channel signals, $|\hat{L}(k)|$ and $|\hat{R}(k)|$, are obtained from the mono downmix using:

$$|\hat{L}(k)| = |\hat{M}(k)| \frac{f(b)}{1+f(b)}, \quad (13)$$

and

$$|\hat{R}(k)| = |\hat{M}(k)| \frac{1}{1+f(b)}. \quad (14)$$

Based on the decoded WITD and WIPD, the sub-band phase differences between the left and right channels are calculated using (2). The phases of left and right channels can be recovered from the phase of the mono signal and the decoded WITD/WIPD:

$$\angle\hat{L}(k) = \angle\hat{M}(k) - \frac{1}{1+f(b)} \frac{2\pi k \cdot W\hat{I}TD}{N}, \quad (15)$$

and

$$\angle\hat{R}(k) = \angle\hat{M}(k) - \frac{f(b)}{1+f(b)} I\hat{P}D(k) - \frac{W\hat{I}PD}{1+f(b)}. \quad (16)$$

For low frequency bands, where the ITD can be perceived, the synthesized spectrums are obtained by combining (13) with (15), and (14) with (16) for left and right channels respectively:

$$\hat{L}(k) = \frac{|\hat{M}(k)| f(b)}{1+f(b)} e^{j\left(\angle\hat{M}(k) - \frac{1}{1+f(b)} \frac{2\pi k \cdot W\hat{I}TD}{N}\right)}, \quad (17)$$

and

$$\hat{R}(k) = \frac{|\hat{M}(k)|}{1+f(b)} e^{j\left(\angle\hat{M}(k) - \frac{f(b)}{1+f(b)} I\hat{P}D(k) - \frac{W\hat{I}PD}{1+f(b)}\right)}. \quad (18)$$

For higher frequency bands, the phase of the reconstructed channels is simply set to the phase of decoded mono signal. At the end, the left and right channels are converted back into the time domain with the frequency-time transform.

4.2. Application to stereo extension of ITU-T G.722

The proposed stereo coding scheme has been developed in the course of the recent ITU-T standardization on stereo

extension of G.722 [11]. This novel parametric stereo codec has been tested with G.722 core bitstream at 56 kbit/s and with an 8kbit/s stereo extension layer. This codec operates with a frame length of 5ms. The G.722 stereo extension encoder operates in the FFT domain. The proposed WITD/WIPD and the downmixing method are applied together with the ILD-based stereo coding scheme described in [6]. At the decoder, G.722 compatible bitstream is first fed into the G.722 mono decoder to produce a time domain mono decoded signal. At the same time, stereo parameters which are sub-band ILDs and WITD/WIPD included in the stereo bitstream are decoded. The mono decoded signal is converted to the frequency domain using the FFT. The stereo synthesis operates in this frequency domain using (17) and (18). Finally the two channels are converted back to the time domain using inverse FFT.

5. EXPERIMENTAL RESULTS

Two subject listening tests were conducted to assess the performance of the proposed algorithms:

- Test 1: A/B test to compare the quality of the proposed downmixing scheme with the method described in [6].
- Test 2: Ref/A/B test to evaluate the performance of the WITD/WIPD synthesis.

For each listening test, 17 critical items were selected including clean speech and noisy speech recorded with various microphone setup (Binaural, MS), as well as music items. The speech signals are in Chinese language. All the test items were pre-filtered to 50-7000 Hz using the P.341 filter prototype and normalized in level at -26dBov. There were seven expert listeners for each test. Both tests were conducted using high-quality headphones Sennheiser HD280Pro. For the first test, listeners were asked to give their preference as no absolute reference can be used for mono downmix. For the second test, the listeners were asked to listen to three conditions: original signal, reference system and proposed coding scheme. The comparison mean opinion scores (CMOS) are illustrated per item as well as in average in Fig. 4 and 5.

The first listening test reveals that the overall quality of the proposed downmixing technique is statistically better than the reference method [6] as shown in Fig. 4. For some specific items (item8, item13, item17) the quality improvement is significant.

The second listening test was conducted to evaluate the performance of WITD/WIPD. The reference codec uses only sub-band ILD information with the same total bit rate. The comparison results for the proposed codec and the reference codec are shown in Fig. 5. The overall quality of the proposed stereo coding scheme is slightly better than the reference codec in the mean score, but Fig. 5 shows a significant improvement for some items (number 2, 8, 9 and 15) which are binaural speech recording or artificial music.

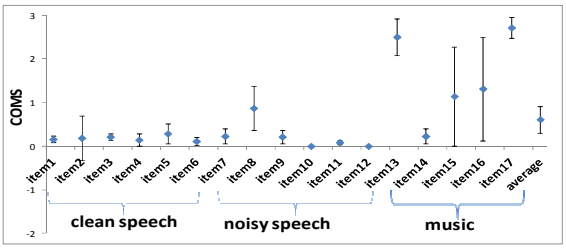

Fig. 4. CMOS results for downmixng methods.

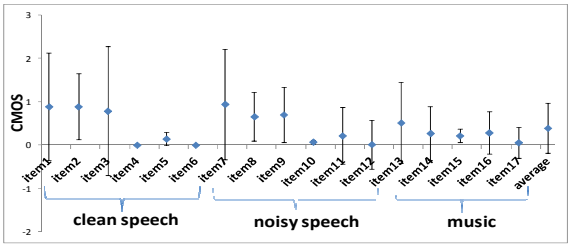

Fig. 5. CMOS results for whole band ITD/IPD.

6. CONCLUSIONS

In this paper, a novel low bitrate parameter stereo coding scheme has been described. It relies on the uses of WITD and WIPD which represent important spatial cues for low bit rate stereo coding. As part of this coding scheme, a new dowmixing method is provided. The low bitrate stereo coding scheme is successfully applied to ITU-T G.722 stereo extension as illustrated in the subjective quality assessment test of the new coding modules.

7. RELATION TO PRIOR WORK

The work presented here has focused on low bitrate stereo coding with the use of WITD and WIPD to enhance the spatial images. Moreover, a novel stereo to mono downmixing algorithm is also described using sub-band phase alignment. The work in [1,2,3,4] describes parametric stereo coding scheme relying on sub-band ILD and ITD/IPD which is not applicable for low bit rate and low delay applications. The presented study simplifies the relationship between OPD and IPD compared to [9,10]. T.M.N. Hoang and S. Ragot [6] presented a low bit rate stereo extension framework which uses a downmix method aligning the phase of the downmix to a sum of the left and right channel phases, and only considers ILD as spatial cue. This coding scheme causes distortion in the downmix for out of phase signal and does not offer sufficient quality for stereo signals with delay or phase differences.

ACKNOWLEDGEMENTS

The authors would like to thank Herve Taddei and Xu Jianfeng for their contributions to this work.

8. REFERENCES

[1] F. Baumgarte and C. Faller, “Binaural Cue Coding - Part I: Psychoa-coustic fundamentals and design principles,” IEEE Trans. on Speech and Audio Proc., vol. 11, no. 6, pp. 509–519, Nov. 2003.

[2] C. Faller and F. Baumgarte, “Binaural Cue Coding - Part II: Schemes and applications,” IEEE Trans. on Speech and Audio Proc., vol. 11,no. 6, pp. 520–531, Nov. 2003.

[3] E. Schuijers, W. Oomen, B. Brinker, and J. Breebaart, “Advances in Parametric Coding for High-Quality Audio,” in Preprint 5852, 114th AES convention, Amsterdam, Mar. 2003.

[4] Jeroen Breebaart, Steven van de Par, Armin Kohlrausch, Erik Schuijers, “Parametric Coding of Stereo Audio” EURASIP Journal on Applied Signal Processing 2005:9, 1305–1322.

[5] J. Blauert, “Spatial Hearing: The Psychoacoustics of Human Sound Localization”, MIT Press, Cambridge, USA, 1997.

[6] T.M.N. Hoang, S. Ragot,, B. Kövesi, P. Scalart, “Parametric stereo extension of ITU-T G.722 based on a new downmixing scheme, ” in *Proc. IEEE MMSP*, St Malo, France, Oct. 2010.

[7] Y. Lang, D. Virette, C. Faller, “Novel low complexity coherence estimation and synthesis algorithms for parametric stereo coding,” in *Proc. EUSIPCO*, pp. 2427 – 2431, 2012.

[8] C. Tournery and C. Faller, “Improved time delay analysis/synthesis for parametric stereo audio coding,” in Preprint 120th Conv. Aud. Eng. Soc., May 2006.

[9] J. Lapierre and R. Lefebvre, "On Improving Parametric Stereo Audio Coding," Proc. 120th AES Convention, Paris, France, May 2006.

[10] JungHoe Kim, Eunmi Oh, and Julien Robilliard, “Enhanced Stereo coding with phase parameters for MPEG Unified Speech and Audio Coding”, in Proc. 127th AES Convention, New York, USA, 2009, preprint 7875.

[11] ITU-T Rec. G.722 Annex D (pre-published), Sep. 2012.