G.722 ANNEX D AND G.711.1 ANNEX F - NEW ITU-T STEREO CODECS

David Virette[1], Yue Lang[1], Lei Miao[1], Wenhai Wu[1], Balàzs Kövesi[2], Claude Lamblin[2], Stéphane Ragot[2]

[1]Huawei Technologies, China, [2]France Telecom Orange, France

ABSTRACT

This paper presents the two new ITU-T Recommendations G.722 Annex D and G.711.1 Annex F, which are stereo extensions of the wideband codecs ITU-T G.722 and G.711.1 and their superwideband extensions (G.722 Annex B and G.711.1 Annex D). An embedded scalable structure is used to add stereo extension layers on top of the wideband or superwideband core coding. Wideband stereo modes are supported at the bit rates of 64/80 and 96/128 kbit/s for G.722 and G.711.1 (respectively), while superwideband stereo modes are supported at 80/96/112/128 and 112/128/144/160 kbit/s. The parametric stereo coding model is based on a frequency domain downmix, wideband inter-channel differences estimation, quantization and synthesis, low complexity coherence analysis and synthesis, stereo transient detection and stereo post-processing. An overview of formal ITU-T characterization listening tests illustrates the performance of these codecs.

Index Terms— speech coding, audio coding, parametric stereo coding, G.722 Annex D, G.711.1 Annex F

1. INTRODUCTION

Low bit rate stereo coding is mainly applied to broadcast and streaming applications [1,2] with several MPEG and 3GPP standards relying on parametric stereo coding [3,4]. In conversational applications such as Voice over IP services, wideband (WB, 50-7000 Hz) and superwideband (SWB, 50-14000 Hz) are now widely used. In particular, the ITU-T WB codecs G.722 [5] and G.711.1 [6] have been deployed in fixed VoIP services. Their superwideband extensions, G.722 Annex B (G.722B) and G.711.1 Annex D (G.711.1D), were standardized in 2010 to improve quality with a larger audio bandwidth while coding all types of audio signals (speech, music, etc.) [7]. The next evolution in enriching audio communication is the support of stereo to give a more natural and immersive experience to end users. Hence, the G.722/G.711.1 stereo standardization effort was initiated to further extend those communication codecs (G.722, G.711.1, G.722B, G.711.1D) with embedded scalable stereo extensions. Approved in Summer 2012, the new ITU-T annexes to G.722 (G.722D) [8] and to G.711.1 (G.711.1F) [9] are respectively backward compatible with G.722 and G.711.1 and their SWB mono annexes.

The paper presents the new ITU-T codecs G.722D and G.711.1F, and it is structured as follows. Sections 2 and 3 provide a general description of the stereo extensions outlining their novelties. In Sections 4 and 5, the encoder and decoder are described. Finally, the performance is discussed in Section 6.

2. MAIN FEATURES OF G.722D AND G.722F

Both stereo extensions, G.722D and G.722F, are based on the same stereo coding model, and use common algorithms. The encoder input and decoder output are sampled at 16 kHz and 32 kHz for WB and SWB operating modes respectively. The input signal is processed with 5 ms frames. The operating bit rate range depends on the selected core codec and core bit rate, and on the number of enhancement layers.

G.722D comprises three stereo extension layers with bit rates of 8, 8 and 16 kbit/s generating two bit rates at 64 and 80 kbit/s for WB stereo and four bit rates at 80, 96, 112 and 128 kbit/s for SWB stereo. G.711.1F has only two 16 kbit/s layers corresponding to two bit rates at 96 and 128 kbit/s for WB stereo and four bit rates at 112, 128, 144 and 160kbit/s for SWB stereo. The main difference between the two stereo extensions lies in the first 16 kbit/s G.711.1F layer which is split in two 8 kbit/s layers for G.722D.

3. REVIEW OF PRIOR WORK AND NOVELTY

The work presented in this paper describes an embedded low delay parametric stereo codec standardized in ITU-T.
Principles of embedded speech and audio coding are presented in [15]. The coding model used in this work is based on parametric stereo coding, where the stereo signal is downmixed to a mono signal and inter channel cues are represented to synthesize a stereo image close to the original stereo signals [1,2]. Typically, inter channel cues comprise the inter channel coherence (IC) and the inter channel time/phase/level differences (ITD/IPD/ILD) between the two channels (left and right). ITD and IPD represent the global time/phase differences between the two channels and IC is defined as the normalized inter channel coherence after

phase alignment according to the estimated IPD/ITD. IC represents the width of the stereo image.

Unlike [1,2], this work addresses conversational applications with strong constraints on delay and frame size. It is based on recent approaches in low delay parametric stereo coding [10,11] targeting new coding schemes to support short frames and high robustness against packet loss.

The encoder key modules are stereo transient detection driving stereo parameter quantization, estimation of whole WB inter-channel cues and their selective transmission, and frequency domain downmix.

4. ENCODER OVERVIEW

The SWB stereo encoder block diagram is shown in Fig. 1. The WB stereo encoder is a subset of the SWB encoder as shown in the dash line box. A pre-processing high-pass filter (HPF) is applied to the left and right input signals to remove 0-50 Hz components. For SWB stereo, the pre-processed signals are divided into two 16-kHz sampled WB and super higher-band (SHB, 8-16 kHz) signals, with a 32-tap quadrature mirror filterbank (QMF). The WB signals are transformed in frequency domain by short-term Fast Fourier Transform (FFT) and downmixed to generate a frequency domain mono signal. The downmixed mono signal is converted back to time domain and encoded by the G.722 or G.711.1 WB core encoder.

The stereo parameters are estimated from frequency domain left and right channel signals. SHB stereo parameters are extracted from left and right channel signals in the modified discrete cosine transform (MDCT) domain. First, the SHB left and right signals are downmixed in time domain to get the sum and difference signals, which are then transformed into the MDCT domain. From these sum and difference signal MDCT coefficients, the MDCT coefficients of the left and right channel signals are generated. After gain correction, the SHB sum signal, representing the SHB downmix, is encoded by the SHB encoder part of G.722B/G.711.1D.

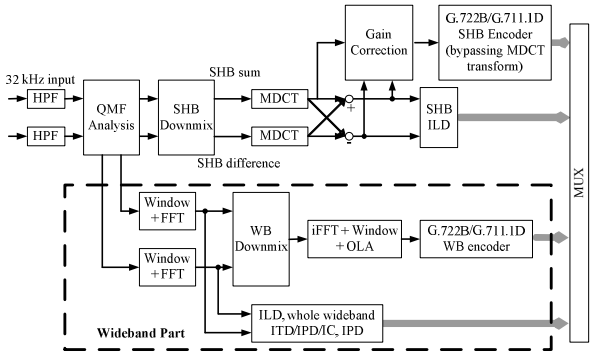

Fig. 1. G.722D/G.711.1F encoder block diagram

4.1. Inter channel cues estimation

Normally, ITD, IPD and IC are estimated and transmitted for each subband in the frequency domain [1, 2], but due to bit budget limitation in this work, there are not enough bits to transmit all these subband inter channel cues in every frame. Only the whole WB ITD, IPD and IC are then transmitted; these WB parameters represent a global inter channel difference for the whole WB spectrum. These three WB parameters are estimated in frequency domain based on the cross correlation.

To have a more stable ITD estimation while tracking fast changes, two smoothed versions of subband cross correlations are computed with strong and weak smoothing respectively. The estimated ITD is obtained by averaging the corresponding subband ITD. Normally, the ITD calculated from strongly smoothed cross correlation is chosen as the final WB ITD. Yet, the weakly smoothed ITD is selected if its standard deviation across the subbands is low, showing a more regular ITD across the spectrum. To ensure stability in the subsequent parameter estimation, the weakly smoothed cross correlation is also used to update the strongly smoothed cross correlation memory. ITD is set positive, if a sound comes first on the left channel. Yet, even for stationary sound sources, the sign of the ITD varies among the subbands due to the low frequency resolution. As only one WB ITD is estimated per frame, its sign depends on the number of positive and negative subband ITDs, and of their standard deviations across the respective subbands.

The WB IPD is computed by summing the perceptually weighted unwrapped IPDs from each subband. A direct sound is usually more energetic than the ambience or reverberated part in a stereo signal. This background ambience is usually less relevant to represent the general spatial image. Therefore the energy weighting emphasize the most perceptually important subbands in the WB IPD estimation.

The WB IC is estimated from the subband IPD stability [10]. The average of subband IPDs varies slowly over consecutive frames for stable stereo image in case of directional sources with no diffuse sources; if the channels greatly differ, this average changes quickly.

Due to bit rate constraints, the WB ITD, IPD and IC cannot all be transmitted every frame. A selection procedure based on the stereo parameter characteristics determines which parameter is perceptually more important and must be transmitted for each frame.

4.2. Stereo transient detection and ILD quantization

Stereo transient is detected when the energy difference between the left and right channels varies greatly and quickly. This difference is based on ILD parameters which are the logarithmic energy ratio per subband in the frequency domain [11]. First the sum of the ILDs in the

current frame is calculated. Then the frame is classified as transient (resp. normal) stereo frame if the difference between this sum and the average of the sum over past frames is higher (resp. lower) than a predetermined threshold. One bit is used to transmit the stereo class information.

The ILD quantization splits the 20 WB subbands in groups of interleaved subbands and only one group is quantized and transmitted per frame. The splitting and the quantization depend on the stereo class. As for transient stereo frame, the ILDs change very fast between consecutive frames, all the subband ILDs must be transmitted in a short time period. Thus, a 2-frame quantization mode is used: the 20 subbands are split in 2 groups, with even and odd indexes. For normal stereo frame, the stereo image is more stable and a 4-frame mode is used: the subbands are split in 4 groups and only the odd or even subbands of 2 of the 4 groups are transmitted for each frame.

4.3. WB downmixing

Since downmixing in time domain cannot take into account phase differences between channels, nor preserve the energy per frequency subband [11], downmixing is performed in the frequency domain (FFT), bin by bin. The downmixed mono signal amplitude is set as the half sum of the left and right channel amplitudes. Its phase is derived from the IPD of each frequency bin and the energy ratio between the two channels. The goal is to set the phase of the mono signal close to the one of the higher energy channel which is perceptually dominant in the downmix.

5. DECODER OVERVIEW

A general description of the decoder is now given in Fig. 2 which shows a block diagram of the SWB stereo decoder. For each 5-ms frame, the decoder can receive any supported bit rates, from 64 kbit/s up to 128 kbit/s for G.722D and from 96 kbit/s to 160 kbit/s for G.711.1F. The bitstream is de-multiplexed into three parts: G.722 / G.711.1 compatible core bitstream, G.722B / G.711.1D SHB bitstream and stereo parameters bitstream. In the WB part, the core compatible bitstream is decoded into time domain WB mono signal, and converted to the frequency domain by a FTT in order to perform the inter channel level adjustment, the phase and coherence synthesis based on the decoded stereo parameters.

In the SHB part, the stereo signal is synthesized in MDCT domain using the SHB decoded mono signal and the received SHB ILD. After converting the SHB signal back to the time domain by an inverse MDCT, a stereo post-processing is applied. For the SWB stereo modes, the final output signal is reconstructed by the QMF synthesis filter.

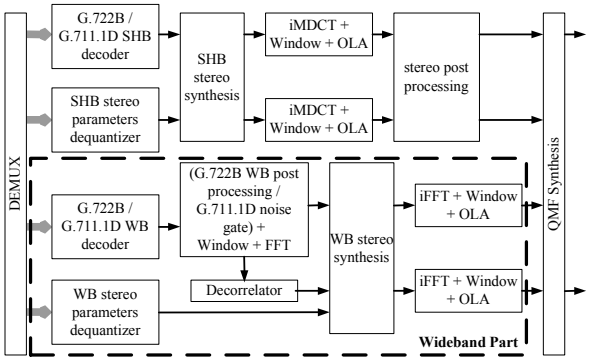

Fig. 2. G.722D/G.711.1F decoder block diagram

5.1. ILD synthesis

Depending on the frame classification in WB and SHB, only a limited number of subband ILDs is received. In a given frame, the non-transmitted subband ILDs are set to their last received values. The transmission of parameters for interleaved subbands ensures a better robustness to packet loss, as the lost subband ILDs can be approximated from the last received neighbor subband.

5.2. Phase and coherence synthesis

The FFT subband phases of the output channels are reconstructed from the phases of the decoded mono signal combined with the energy weighted subband IPDs. As the phase of the downmixed signal is closer to the phase of the channel with higher energy, the synthesized channel phase is obtained from the downmix phase and the inversely weighted IPDs.

When IC is not equal to one, coherence synthesis is performed based on a low complexity time domain decorrelator [10]. The diffuse sounds are obtained by delaying and energy adjustment of the decoded mono signal. The energy of the decorrelated sound depends on the WB IC. The lower IC is, the higher the energy of the diffuse sound is set. The delays for the two decorrelated channels are different and set to a multiple of the frame size in order to perform coherence synthesis directly in the frequency domain and therefore limit the associated complexity. The frequency domain diffuse sounds are added to the left and right channel signals before converting the two channels back to the time domain.

5.3. SHB stereo post-processing

When the SHB downmix mono signal is classified as transient, a stereo post-processing based on the decoded SHB mono envelope is applied to reduce pre-echoes. This processing depends on the stereo classification. If the stereo signal is classified as transient, only one channel contains a

transient, and the time envelope weighted by the channel energy is applied to the higher energy channel. Otherwise, both channels contain the transient signal, and the post-processing is applied to both channels. In that case, the weighted mono envelope is applied directly to the channel where the transient occurs first, and delayed by the absolute value of WB ITD samples before being applied to the other channel.

6. ITU-T PERFORMANCE EVALUATION

6.1. Subjective Quality

The G.722D and G.711.1F WB and SWB stereo codecs were evaluated in ITU-T characterization tests in February 2012. The test methodology was compliant with ITU-R BS.1116-1, with a triple stimulus/hidden reference/double blind ('Ref', 'A', 'B') and a five grade impairment scale [12]. Both codecs were evaluated in 3 sets of experiments (clean speech recorded with binaural and MS microphones, noisy speech, music). The G.722.1 and G.722.1 Annex C codecs operating in dual-mono mode were used as reference in all experiments. Each experiment was run by two listening laboratories in different languages (Lab A: Chinese, Lab B: French; Lab C: Japanese). Laboratories A and B performed the clean speech (Fig. 3) and noisy speech (Fig. 4) experiments, and laboratories A and C performed the music (Fig. 5) experiment.

Selected test results from the official G.722D and G.711.1F test report [13] are summarized in Figs. 3 to 5 in terms of difference scores comparing the mean difference of the coded outputs to the originals. The x axis denotes the codec bit rates in kbit/s and it is grouped per codec. G.722.1 at 48 (2×24) kbit/s, G.722D at 64 kbit/s and G.711.1F at 96 kbit/s operate in WB stereo, while all other bit rates operate in SWB stereo. In Figs. 3 and 4, WB and SWB test results are presented together, but they were obtained from separate experiments.

In clean speech and in both labs, all requirements but two and all objectives but one were passed. In noisy speech, for each noise, all requirements but two were passed in both labs. In music, though G.722D and G.711.1F did not reach the quality level of dual-mono codecs, their quality was similar to parametric stereo codec for broadcast applications.

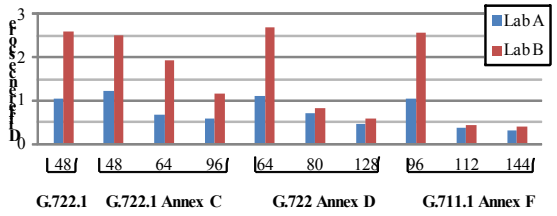

Fig. 3. Clean speech quality (-26dBov).

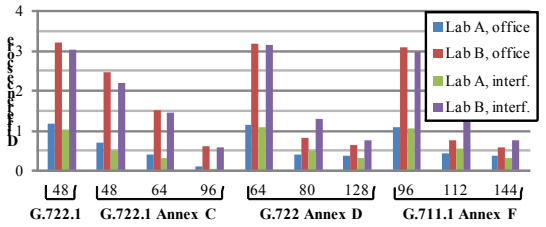

Fig. 4. Noisy speech quality.

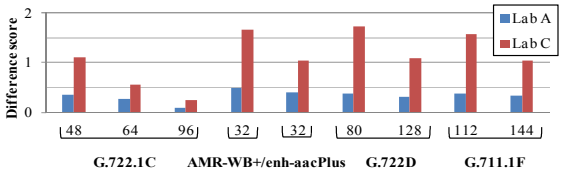

Fig. 5. Music quality.

6.2. Complexity and delay

The observed worst-case complexity and storage requirements for SWB modes in 16-bit words (encoder plus decoder) based on the ITU-T Software Tool Library STL2009 [14] are given in Table 1, together with the algorithmic delays.

Table 1. Complexity, memory and delay of G.722D/G.711.1F

Codec		**G.722D**	**G.711.1F**
Complexity (WMOPS)	Encoder	23.17	25.43
	Decoder	19.15	16.84
	Overall	42.32	42.27
Dynamic/Static RAM [kwords]		7906	7780
Data ROM [kwords]		6336	7576
Program ROM [instructions]		9174	10132
WB stereo delay (in ms)		13.625	18.125
SWB stereo delay (in ms)		15.9375	19.0625

7. CONCLUSION

This paper presented an overview of the recently standardized codecs ITU-T G.711.1F and G.722D. These new codecs provide low-delay stereo coding in wideband and superwideband while being backward compatible with G.711.1/G.711 and G.722 thanks to their embedded structure. The codecs demonstrated high quality for various audio contents, with good performance compared to dual-mono conversational codecs.

ACKNOWLEDGEMENTS

The authors would like to thank Simão Campos Neto, Thi Minh Nguyet Hoang, Xu Jianfeng, Yusuke Hiwasaki, for their contributions to the G.722D and G.711.1F work.

REFERENCES

[1] 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.

[2] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," in *Preprint 114th Convention AES*, Mar. 2003.

[3] 3GPP TS 26.401, e-AAC+, General audio codec audio processing functions; Enhanced aacPlus general audio codec; General description.

[4] ISO/IEC 23003-1:2007 Information technology – MPEG audio technologies – Part 1: MPEG Surround.

[5] X. Maitre, "7 kHz audio coding within 64 kbit/s," *IEEE Select. Areas. Com.*, vol. 6, no. 2, pp. 283-298, Feb. 1988.

[6] Y. Hiwasaki and H. Ohmuro, "ITU-T G.711.1: Extending G.711 to Higher-Quality Wideband Speech," *IEEE Commun. Mag.*, vol. 47, no. 10, pp. 110-116, Oct. 2009.

[7] L. Miao; Z. Liu; C. Hu; V. Eksler, S. Ragot, C. Lamblin, B. Kovesi, J. Sung; M. Fukui, S. Sasaki, Y. Hiwasaki, "G.711.1 Annex D and G.722 Annex B - New ITU-T superwideband codecs," in *Proc. ICASSP,* pp. 5232 – 5235, 2011.

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

[9] ITU-T Rec. G.711.1 Annex F (pre-published), Sep. 2012.

[10] 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.

[11] 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.

[12] ITU-R Rec. BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.

[13] ITU-T TD534-GEN/16, "Summary of test results of ITU-T G.711.1 Annex D and ITU-T G.722 Annex B layered stereo", May 2012

[14] ITU-T Rec. G.191, "Software tools for speech and audio coding standardization," March 2010.

[15] B. Geiser, S. Ragot, and H. Taddei, "Embedded Speech Coding: From G.711 to G.729.1," Chapter 8 in *Advances in Digital Speech Transmission* (R. Martin, U. Heute, C. Antweiler, eds.), Wiley, Jan. 2008, pp. 201-248